Міністерство освіти і науки України
Харківський національний університет імені В.Н. Каразіна

Кафедра статистики, обліку та аудиту

«ЗАТВЕРДЖУЮ»

Завідувач кафедри

_____Володимир СОБОЛЄВ

Протокол № 10 від “22” червня 2020 р.

НАВЧАЛЬНО-МЕТОДИЧНИЙ КОМПЛЕКС

дисципліни «Методи класифікації даних в пакеті Statistica»

для студентів денної (заочної) форми навчання

рівень вищої освіти другий (магістерський)

галузь знань 05 «Соціальні та поведінкові науки»

спеціальність 051 «Економіка»

освітня програма «Бізнес-аналітика та міжнародна статистика»
«Економічна аналітика та статистика»

Розроблено:
д.е.н., доцент, доцент кафедри статистики, обліку та аудиту
Корепанов Олексій Сергійович

2020/2021 навчальний рік

ЗМІСТ

1. Робоча програма навчальної дисципліни
2. Навчальний контент (розширений план лекцій)
3. Плани практичних (семінарських) занять, самостійної роботи
4. Питання, задачі, завдання або кейси для поточного та підсумкового контролю знань і вмінь здобувачів вищої освіти, для контрольних робіт, передбачених навчальним планом, післяатестаційного моніторингу набутих знань і вмінь з навчальної дисципліни
    4.1. Питання, задачі, завдання або кейси для поточного та підсумкового контролю знань і вмінь здобувачів вищої освіти
    4.2. Контрольні роботи, передбачені навчальним планом
5. Завдання семестрових екзаменів (письмових залікових робіт)
6. Критерії оцінювання знань студентів та розподіл балів

1. РОБОЧА ПРОГРАМА НАВЧАЛЬНОЇ ДИСЦИПЛІНИ

Міністерство освіти і науки України

Харківський національний університет імені В. Н. Каразіна

Кафедра статистики, обліку та аудиту

"**ЗАТВЕРДЖУЮ**"

Проректор з науково-педагогічної роботи

_____ Антон ПАНТЕЛЕЙМОНОВ

"_____"_____2020 р.

Робоча програма навчальної дисципліни

"Методи класифікації даних в пакеті Statistica"

(шифр і назва навчальної дисципліни)

рівень вищої освіти другий (магістерський)

галузь знань 05 «Соціальні та поведінкові науки»

спеціальність 051 «Економіка»

освітня програма «Бізнес-аналітика та міжнародна статистика»

спеціалізація _____

вид дисципліни за вибором

факультет економічний

2020 / 2021 навчальний рік

Міністерство освіти і науки України

Харківський національний університет імені В. Н. Каразіна

Кафедра статистики, обліку та аудиту

"**ЗАТВЕРДЖУЮ**"

Проректор з науково-педагогічної роботи

_____ Антон ПАНТЕЛЕЙМОНОВ

"_____"_____2020 р.

Робоча програма навчальної дисципліни

"Методи класифікації даних в пакеті Statistica"

(шифр і назва навчальної дисципліни)

рівень вищої освіти другий (магістерський)

галузь знань 05 «Соціальні та поведінкові науки»

спеціальність 051 «Економіка»

освітня програма «Економічна аналітика та статистика»

спеціалізація _____

вид дисципліни за вибором

факультет економічний

2020 / 2021 навчальний рік

2. НАВЧАЛЬНИЙ КОНТЕНТ

ПЛАН ЛЕКЦІЙ

Тема 1. Основні поняття методів багатомірної класифікації
1. Сутність задач класифікації об'єктів
2. Систематизація задач класифікації об'єктів
3. Типологізація математичних постановок задач класифікації

Тема 2. Кластерний аналіз. Ієрархічні методи класифікації
1. Загальна характеристика методів кластерного аналізу
2. Відстань між об’єктами й міра близькості
3. Відстань між кластерами
4. Алгоритми ієрархічного й дивізимного методів кластерного аналізу

Тема 3. Кластерний аналіз. Ітеративні методи класифікації
1. Метод k - середніх
2. Метод пошуку згущень
3. Критерії якості класифікації

Тема 4. Дискримінантний аналіз
1. Основні положення дискримінантного аналізу
2. Дискримінантні функції і їхня геометрична інтерпретація.
3. Класифікація при наявності навчальних вибірок.
4. Взаємозв'язок між дискримінантними змінними й дискримінантними функціями

Тема 5. Аналіз даних методами нечіткої кластерізації
1. Постановка задачі нечіткої кластерізації
2. Алгоритм розв’язування задачі нечіткої кластеризації
3. Виконання алгоритму FCM в системі MATLAB
4. Приклад реалізації алгоритму FCM

КОНСПЕКТ ЛЕКЦІЙ

Тема 1. Основні поняття методів багатомірної класифікації

1. Сутність задач класифікації об'єктів
2. Систематизація задач класифікації об'єктів
3. Типологізація математичних постановок задач класифікації

Після вивчення лекції студент повинен:
Знати:

- що розуміється під терміном «класифікація»;
- загальну задачу класифікації в термінах статичного варіанта різних форм завдання вихідних статистичних даних;
- типи задач класифікації та варіанти (приклади) кінцевих прикладних цілей дослідження для даного типу задачі класифікації;
- здійснити вибір математичної постановки задачі класифікації;
- доцільність і ефективність застосування тих або інших методів класифікації.

Вміти:

- визначити підхід до вирішення задачі *класифікації при наявності навчальних вибірок* «*класифікація з навчанням*» та задачі без *наявності навчальних вибірок* «*класифікації без навчання*»;
- виконати систематизацію задач класифікації відповідно з кінцевими прикладними цілями дослідження;
- здійснити вибір математичної постановки задачі класифікації;

1. Сутність задач класифікації об'єктів

До розробки апарата багатомірного статистичного аналізу й, головне, до появи й розвитку досить потужної електронно-обчислювальної бази проблеми теорії й практики класифікації відносилися не до розробки методів і алгоритмів, а до повноти й старанності відбору й теоретичного аналізу досліджуваних об'єктів, ознак, що їх характеризують, змісту й числа градацій по кожній із цих ознак.

Всі методи класифікації зводилися власне кажучи до методу так називаного *комбінаційного групування*, коли всі ознаки, що характеризують об'єкт, носять дискретний характер або зводяться до таких (стать або мотив міграції індивідуума, рівень житлових умов або число дітей у родині й т.п.), а *два об'єкти належать до однієї групи тільки при точному збігу зареєстрованих на них градацій одночасно по всіх ознаках, що їх характеризують* (однакова стать, мотив міграції й т.д.).

Однак у звязку із поступовим збільшенням обсягів інформації, що перероблюється, і, зокрема, числа об'єктів, що класифікуються, і ознак, що їх характеризують, можливість ефективної реалізації подібної логіки дослідження ставала все менш реальною (так, наприклад, число k груп або класів, підраховуване при комбінаційному угрупованні за формулою $k = m_1 \cdot m_2 ... m_p$, де m_j - число градацій по $x^{(j)}$ ознаці, а p - загальне число аналізованих ознак, вже при $m_j = 3$ й $p = 5$ виявляється рівним 243). Саме електронно-обчислювальна техніка стала тим головним інструментом, що дозволив по-новому підійти до рішення цієї важливої проблеми й, зокрема, конструктивно скористатися розробленим до цього часу потужним апаратом багатомірного статистичного аналізу: методами розпізнавання образів «із учителем» (дискримінантний аналіз) і «без учителя» (автоматична класифікація або кластер-аналіз).

Розвиток електронно-обчислювальної техніки як засобу обробки великих масивів даних стимулював проведення в останні роки широких комплексних досліджень складних со-

ціально-економічних, технічних, медичних і інших процесів і систем, таких, як образ і рівень життя населення, удосконалювання організаційних систем, регіональна диференціація соціально-економічного розвитку, планування й прогнозування галузевих систем, закономірності виникнення збоїв (у техніці) або захворювань (у медицині) і т.п. У зв'язку з багатоплановістю й складністю цих об'єктів і процесів дані про них по необхідності носять *багатомірний* і *різнотипний* характер, тому що до їхнього аналізу звичайно буває неясно, наскільки істотна та або інша властивість для конкретної мети. У цих умовах виходять на перший план проблеми побудови угруповань і класифікацій за багатомірними даними (тобто проблеми *класифікації багатомірних спостережень*), причому з'являється можливість оптимізації цієї побудови з погляду найбільшої відповідності одержуваного результату поставленій кінцевій меті класифікації.

Цілі класифікації істотно розширюються, і одночасно зміст самого процесу класифікації стає незмірно складніше. Він, зокрема, доповнюється *проблемою побудови самої процедури класифікації*, що раніше носила чисто технічний характер.

Перш ніж переходити до прикладів і типологізації (у прикладному й математичному аспектах) задач класифікації, визначимо сам термін *класифікація*.

У самому загальному формулюванні під **класифікацією** *ми будемо розуміти поділ розглянутої сукупності об'єктів або явищ на однорідні, у певному сенсі, групи або віднесення кожного із заданої множини об'єктів до одного із заздалегідь відомих класів* (при цьому «задана множина», що класифікується, може складатися з єдиного об'єкта). Помітимо, що термін «класифікація» використовується, залежно від контексту, для позначення як самого процесу «розділення-віднесення», так і його результату.

Найпоширеніші наступні дві форми подання вихідних статистичних даних (в.с.д.).

По-перше, у вигляді матриці (або таблиці) «об'єкт - властивість»:

$$\left(\text{в.с.д.}\right)_1 = \begin{pmatrix} x_1^{(1)}(t) & x_1^{(2)}(t) & \ldots & x_1^{(p)}(t) \\ x_2^{(1)}(t) & x_2^{(2)}(t) & \ldots & x_2^{(p)}(t) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_n^{(1)}(t) & x_n^{(2)}(t) & \ldots & x_n^{(p)}(t) \end{pmatrix}, \quad t = t_1, t_2, \ldots, t_N, \qquad (1.1)$$

де $x_i^{(j)}(t_k)$ - значення j-ї аналізованої ознаки, що характеризує стан i-го об'єкта в момент часу t_k.

По-друге, матрицею парних порівнянь розміру $n \times n$ (якщо розглядаються характеристики парних порівнянь об'єктів) або $p \times p$ (якщо розглядаються характеристики парних порівнянь ознак):

$$\left(\text{в.с.д.}\right)_2 = \begin{pmatrix} \gamma_{11}(t), & \gamma_{12}(t), \ldots, & \gamma_{1m}(t) \\ \gamma_{21}(t), & \gamma_{22}(t), & \ldots, \gamma_{2m}(t) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \gamma_{m1}(t), & \gamma_{m2}(t), & \ldots, \gamma_{mm}(t) \end{pmatrix} \begin{pmatrix} m = n \quad \text{или} \quad p; \\ t = t_1, \quad \ldots, \quad t_N \end{pmatrix}. \qquad (1.2)$$

У статичному варіанті, тобто при $N = 1$, дослідник розташовує лише однією матрицею парних порівнянь (γ_{ij}), що описує ситуацію в один якийсь фіксований момент часу.

Сформулюємо загальну задачу класифікації в термінах статичного варіанта різних форм завдання вихідних статистичних даних (формули (1.1), (1.2)) *за схемою «на вході - на виході задачі»*.

На «вході» задачі дослідник має:

а) *n класифікуємих об'єктів*, представлених даними виду (1.1) (тоді кожний *i-й* рядок матриці (1.1) відбиває значення *p, що* характеризують *i-й* об'єкт ознак $x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(p)}$) або

даними виду (1.2) (тоді кожний *i-й* рядок матриці (1.2) задає попарні відносини $\gamma_{i1}, \gamma_{i2}, ..., \gamma_{in_i}$ *i-го* об'єкта з усіма іншими об'єктами, що класифікуються);

б) *навчальні вибірки*

$$X_{i1}, X_{i2}, ..., X_{in_i}, \qquad i = 1,2,...,k, \qquad (1.3)$$

кожна (*j-а*) з яких визначає значення аналізованих ознак $X_{ji} = \left(x_{ji}^{(1)}, x_{ji}^{(2)}, ..., x_{ji}^{(p)}\right)^T$ на n_j об'єктах (тобто $i = 1,2,...,n$), про які апріорі відомо, що всі вони належать *j-му* класу, причому число *k* різних вибірок (1.3) дорівнює загальному числу всіх можливих класів (так що кожний клас представлений своєю порцією вибіркових даних).

На «виході» задачі результат може бути представлений в одній із двох форм:

(*1*) *якщо число класів k і їхній зміст відомі заздалегідь*, те кожне з *n* класифікуємих багатомірних спостережень повинне бути позначене «адресою» (номером) класу, до якого воно належить;

(*2*) *якщо число класів і* (*або*) *їхній зміст виявляються в процесі класифікації*, то результатом класифікації є поділ множини класифікуємих об'єктів на певне число однорідних (у певному змісті) груп, кожна з яких оголошується «класом».

Якщо дослідник розташовує на «вході» задачі не тільки даними, що класифікуються, але й навчальними вибірками, то говорять, що вирішується *задача класифікації при наявності навчальних вибірок* («*класифікація з навчанням*»); у противному випадку мова йде про задачу «*класифікації без навчання*».

Для пояснення сутності основних прикладних типів задач класифікації й кінцевих прикладних цілей, які ставить при цьому перед собою дослідник, розглянемо приклади.

2. Систематизація задач класифікації об'єктів

Аналіз розглянутих вище прикладів з обліком, звичайно, і іншого досвіду, що накопичився до теперішнього часу, рішення практичних задач класифікації в економіці, соціології, психології й інших сферах практичної й наукової діяльності людини дозволяє зробити певну систематизацію цих задач відповідно з кінцевими прикладними цілями дослідження (табл. 1.1).

Таблиця 1.1

Систематизація задач класифікації

№з/п	Тип задачі класифікації	Варіанти (приклади) кінцевих прикладних цілей дослідження для даного типу задачі класифікації
1	2	3
1	*Комбінаційні угруповання* і їхні безперервні узагальнення	1.1. Складання частотних таблиць і графіків, що характеризують розподіл статистично обстежених об'єктів по градаціях або інтервалам групування ознак, що їх характеризують
2	*Проста типологізація*: виявлення «стратифікованої структури» множини статистично обстежених об'єктів, «намацування» і опис чітко виражених скупчень («згустків», «кластерів», «образів», класів) цих об'єктів в аналізованому багатомірному просторі й побудова правила віднесення кожного нового об'єкта до одному з виявлених класів	2.1. Класифікація як необхідний попередній етап дослідження, коли до проведення основної статистичної обробки множини аналізованих даних (побудови регресійних моделей, оцінки параметрів генеральної сукупності й т.д.) домагаються розшарування цієї множини на однорідні (у сенсі проведеного потім статистичного аналізу) групи даних. 2.2. Виявлення й опис розшарованої природи аналізованої сукупності статистично обстежених об'єктів з метою формування плану вибіркових обстежень цієї сукупності

№з/п	Тип задачі класифікації	Варіанти (приклади) кінцевих прикладних цілей дослідження для даного типу задачі класифікації
1	2	3
		2.3 Перший крок у побудові зв'язкових типологій
3	*Зв'язна неупорядкована типологізація*: дослідження залежностей між класифікаціями, що не піддаються впорядкуванню тієї самої множини об'єктів у різних ознакових просторах, одне з яких побудовано на результуючих (поведінкових) ознаках, а друге — на описових	3.1. Прогноз економіко-соціологічних ситуацій або окремих соціально-економічних показників, включаючи задачу виявлення так званих типоутворюючих ознак, у тому числі латентних, тобто безпосередньо не спостережуваних. 3.2. Діагностика в промисловості, техніці, георозвідці, медицині. 3.3. Кластер-аналіз, автоматичне (машинне) розпізнавання образів - зорових, слухових.
4	*Зв'язна впорядкована типологізація*: модифікація зв'язної неупорядкованої типологізації (див. п. 3 табл.), обумовлена додатковим допущенням, що класи, одержувані в просторі результуючих (поведінкових) ознак, піддаються експертному впорядкуванню по деякій зведеній властивості: ефективності функціонування, якості, ступеню прогресивності (оптимальності) поведінки й т.п.	4.1. Побудова й інтерпретація єдиної (вільної) латентної ознаки-класифікатора у вигляді функції від вихідних описових ознак: класифікація хімічних елементів по заряду їхнього атомного ядра (періодична система Д. І. Менделєєва); побудова фактору загальної обдарованості в педагогіці й психології; побудова зведеного показника ефективності функціонування підприємства; побудова інтегральної характеристики рівня майстерності спортсменів в ігрових видах спорту й т.д.
5	*Структурна типологізація*: доповнення й розвиток простої типологізації (див. п. 2 табл.) у напрямку вивчення й опису структури взаємозв'язків отриманих класів, включаючи побудову відповідних ієрархічних систем, аналіз ролі й місця кожного елемента й класу в загальній структурній класифікаційній схемі. При цьому структурна класифікаційна схема визначається складовими її класами (підсистемами) і характеристиками (правилами) їхньої взаємодії	5.1. Класифікація задач багатоцільового комплексу (великої програми, наукового напрямку виробничого комплексу й т.п.) 5.2. Класифікація елементів і підсистем по їхньому функціональному призначенню (виробництв - у територіально-виробничому комплексі, територіальних одиниць - у народногосподарському розподілу праці й споживання, елементів організаційних структур і т.і.). 5.3. Класифікація осіб, що приймають рішення, по їхній ролі й близькості позицій у розумінні ситуації й способі рішення задачі. 5.4. Класифікація досліджуваних ознак і аналіз структури зв'язків між ними
6	Класифікація динамічних траєкторій розвитку систем: типологізація траєкторій багатомірних тимчасових рядів $X(t)=\left(x^{(1)}(t),...,x^{(p)}(t)\right)^T$ серед компонентів $x^{(j)}(t)$ яких можуть бути як кількісні, так і якісні змінні	6.1. Задачі аналізу типів динаміки сімейної структури, споживчої поведінки домашніх господарств та ін.

3. Типологізація математичних постановок задач класифікації

Доцільність і ефективність застосування тих або інших методів класифікації так само, як їхня предметна свідомість, обумовлені конкретизацією базової математичної моделі, тобто математичною постановкою задачі. Визначальним моментом у виборі математичної постановки задачі є відповідь на питання, на якій апріорній інформації будується модель. При цьому апріорна інформація складається із двох частин:

1) з апріорних відомостей про досліджувані класи;

2) з апріорної статистичної (вибіркової) інформації, тобто так званих навчальних вибірок.

Апріорні відомості про досліджувані генеральні сукупності належать звичайно до виду або деяких загальних властивостей закону розподілу досліджуваного випадкового вектора X у відповідному просторі й виходять або з теоретичних, предметно-професійних міркувань про природу досліджуваного об'єкта, або як результат попередніх досліджень. Одержання апріорної вибіркової інформації в економіці й соціології, як правило, пов'язане з організацією системи експертних оцінок або із проведенням спеціального попереднього етапу, присвяченого рішенню задачі простої типологізації аналізованих об'єктів у просторі результуючих показників (див. приклад 1.1).

Класифікація задач розбивки об'єктів на однорідні групи (залежно від наявності апріорної й попередньої вибіркової інформації) і відповідний розподіл опису апарата рішення цих задач представлені в табл. 1.2.

Таблиця 1.2

Застосовність методів класифікації

Апріорні відомості про класи (генеральні сукупності)	Попередня вибіркова інформація	
	Немає інформації	Є навчальні вибірки
Деякі самі загальні припущення про закон розподілу досліджуваного вектора: гладкість, зосередженість усередині обмеженої області й т.п.	Класифікація без навчання: кластер-аналіз, таксономія, розпізнавання образів «без навчання», ієрархічні класифікації.	Непараметричні методи дискримінантного аналізу.
Генеральні сукупності, що розрізняють, задані у вигляді параметричного сімейства законів розподілу ймовірностей (параметри невідомі).	Інтерпретація досліджуваної генеральної сукупності як суміші декількох генеральних сукупностей. «Розщеплення» цієї суміші за допомогою методів оцінювання невідомих параметрів.	Параметричні методи дискримінантного аналізу.
Генеральні сукупності, що розрізняють, задані однозначним описом відповідних законів.	Класифікація при повністю описаних класах: розрізнення статистичних гіпотез.	Навчальні вибірки не потрібні

Контрольні запитання

1. Що розуміється під класифікацією об'єктів?
2. У чому полягає проблема класифікації об'єктів за багатомірними даними?
3. У якій формі можуть представлятися вихідні дані в задачах класифікації об'єктів?
4. Що таке навчальна вибірка?
5. Які кінцеві прикладні цілі ставить перед собою дослідник при проведенні класифікації?
6. Які фактори називаються типоутворюючими?
7. Як будується комбінаційне угруповання?
8. Яка ідея покладена в основу методу відбору найбільш інформативних ознак-детермінантів?

9. Дати характеристику класифікації як необхідного попереднього етапу статистичної обробки багатомірних даних.
10. Як використовується класифікація в задачах планування вибіркових обстежень?
11. Як класифікуються задачі розбивки об'єктів на однорідні групи залежно від наявності апріорної й попередньої вибіркової інформації?

Тема 2. Кластерний аналіз. Ієрархічні методи класифікації

1. Загальна характеристика методів кластерного аналізу
2. Відстань між об'єктами й міра близькості
3. Відстань між кластерами
4. Алгоритми ієрархічного й дивізимного методів кластерного аналізу

Після вивчення лекції студент повинен:
Знати:
- визначення та мету кластерного аналізу;
- три різних підходи до проблеми кластерного аналізу;
- задачі, які вирішують методи кластерного аналізу;
- дві групи методів кластерного аналізу;
- найбільш уживані міри відстані між об'єктами;
- методи об'єднання кластерів, які є найбільш уживаними;

Вміти:
- виконати групування первинних даних;
- визначити критерії якості, цільову функцію, значення якої дозволять зіставити різні схеми класифікації;
- здійснити вибір метрики (або міри близькості) між об'єктами;
- провести нормування значеннь вихідних змінних;
- використовувати методи, які засновані на мінімізації внутрігрупових сум квадратів (відхилень)

1. Загальна характеристика методів кластерного аналізу

Кластерний аналіз — це сукупність методів, що дозволяють класифікувати багатомірні спостереження, кожне з яких описується набором вихідних змінних $X_1, X_2, ..., X_m$. Метою кластерного аналізу є утворення груп схожих між собою об'єктів, які прийнято називати кластерами. Слово *кластер* англійського походження (*cluster*), переводиться як *згусток, пучок, група*. Родинні поняття, використовувані в літературі, — *клас, таксон, згущення*.

Кластерний аналіз – це сукупність методів, підходів і процедур, які розробляються для розв’язування проблеми формування класів – сукупностей даних, однорідних за заданими ознаками.

Кластерний аналіз (автоматична класифікація сукупності даних) займає одно з центральних місць серед методів аналізу даних і являє собою сукупність підходів та алгоритмів знаходження деякого розбиття досліджуваної сукупності об’єктів на підмножини відносно схожих між собою елементів. Такі підмножини отримали назву кластерів.

Виділення кластерів серед сукупності даних має відповідати наступним вимогам:

1. кожний кластер представляє собою сукупність об’єктів, які схожі між собою значеннями деяких властивостей або ознак;
2. сукупність всіх кластерів має бути вичерпаною, тобто всі об’єкти досліджуваної сукупності мають належить до деякого кластеру;
3. кластери мають бути взаємно-виключні; тобто, жоден з об’єктів не має належить до двох різних кластерів.

Формально, під задачею кластерного аналізу розуміється задача знаходження деякого теоретико-множинного розбиття початкової множити об’єктів на підмножини, які не перетинаються, таким чином, щоб елементи, які відносяться до однієї підмножини відрізнялися між собою в значно меншій степені, ніж об’єкти з різних підмножин.

Кластерний аналіз — один з напрямків статистичного дослідження. Особливо важливе місце він займає в тих галузях науки, які пов'язані з вивченням масових явищ і процесів. Необхідність розвитку методів кластерного аналізу і їхнього використання продиктована насамперед тим, що вони допомагають побудувати науково обґрунтовані класифікації, виявити внутрішні зв'язки між одиницями спостережуваної сукупності. Крім того, методи кластерного аналізу можуть використовуватися з метою стиснення інформації, що є важливим чинником в умовах постійного збільшення й ускладнення потоків статистичних даних.

У статистичних дослідженнях групування первинних даних є основним прийомом рішення задачі класифікації, а значить і основою всієї подальшої роботи із зібраною інформацією.

Традиційно ця задача вирішується в такий спосіб. Із множини ознак, що описують об'єкт, відбирається один, найбільш інформативний з погляду дослідника, і виконується групування у відповідності зі значеннями даної ознаки. Якщо потрібно провести класифікацію по декількох ознаках, ранжируваним між собою по ступені важливості, то спочатку виконується класифікація по першій ознаці, потім кожний з отриманих класів розбивається на підкласи по другій ознаці, і т.д. Подібним чином будується більшість комбінаційних статистичних групувань .

У тих випадках, коли впорядкувати класифікаційні ознаки не представляється можливим, застосовується найбільш простий метод багатомірного групування - **створення інтегрального показника (індексу)**, що функціонально залежить від вихідних ознак, з наступною класифікацією по цьому показнику.

Розвитком цього підходу є варіант класифікації по декількох узагальнюючих показниках (головних компонентах), отриманим за допомогою **методів факторного аналізу**.

При наявності декількох ознак (вихідних або узагальнених) задача класифікації може бути вирішена **методами кластерного аналізу**, які від інших методів багатомірної класифікації відрізняються відсутністю навчальних вибірок, тобто апріорної інформації про розподіл генеральної сукупності, що являє собою вектор X.

Розходження між схемами рішення задач класифікації багато в чому визначаються тим, що розуміють під поняттями «подібність» і «ступінь подібності».

Після того, як сформульована мета роботи, необхідно спробувати визначити критерії якості, цільову функцію, значення якої дозволять зіставити різні схеми класифікації.

В економічних дослідженнях цільова функція, як правило, повинна мінімізувати деякий параметр, визначений на множині об'єктів (наприклад, метою класифікації встаткування може бути групування , мінімізуюче сукупність витрат часу й засобів на ремонтні роботи).

У випадках, коли формалізувати мету задачі не вдається, критерієм якості класифікації може служити можливість змістовної інтерпретації знайдених груп.

Розглянемо наступну задачу. Нехай досліджується сукупність n об'єктів, кожний з яких характеризується по k заміряним на ньому ознакам X. Потрібно розбити цю сукупність на однорідні в деякому сенсі групи (класи). При цьому практично відсутня апріорна інформація про характер розподілу вимірів X усередині класів.

Отримані в результаті розбивки групи звичайно називаються кластерами (від англ. *cluster* - група елементів, що характеризується якою-небудь загальною властивістю), а також таксонами (від англ. *taxon* - систематизована група будь-якої категорії) або образами. Методи знаходження кластерів називаються кластер-аналізом (відповідно чисельною таксономією або розпізнаванням образів із самонавчанням).

При цьому із самого початку необхідно чітко представити, яка із двох задач класифікації підлягає рішенню. Якщо вирішується звичайна задача типізації, то сукупність спосте-

режень розбивають на порівняно невелике число областей групування (наприклад, інтервальний варіаційний ряд у випадку одномірних спостережень) так, щоб елементи однієї такої області по можливості перебували друг від друга на невеликій відстані.

Рішення іншої задачі типізації полягає у визначенні природного розшарування вихідних спостережень на чітко виражені кластери, що лежать друг від друга на деякій відстані.

Якщо перша задача типізації завжди має рішення, то при другій постановці може виявитися, що множина вихідних спостережень не виявляє природного розшарування на кластери, тобто утворить один кластер.

Методи кластерного аналізу можна розділити на дві великі групи: агломеративні (об'єднуючі) і дивізимні (поділяючі). *Агломеративні методи* послідовно поєднують окремі об'єкти в групи (кластери), а *дивізимні методи* розчленовують групи на окремі об'єкти. У свою чергу кожний метод як об'єднуючого, так і поділяючого типу може бути реалізований за допомогою різних алгоритмів. Найбільш докладно у підручниках описаний самий доступний для розуміння ієрархічний агломеративний кластерный аналіз. Варто помітити, що як агломеративні, так і дивізимні алгоритми трудомісткі і їх складно використати для великих сукупностей. Крім того, результати роботи таких алгоритмів (їхнє графічне зображення) важко піддаються візуальному аналізу.

2. Відстань між об'єктами й міра близькості

Вибір метрики (або міри близькості) між об'єктами, кожний з яких представлений значеннями багатомірної ознаки, що його характеризує, є найважливішим моментом дослідження, від якого вирішальним образом залежить остаточний варіант розбивки об'єктів на класи при будь-якому використовуваному для цього алгоритмі розбивки. У кожній конкретній задачі цей вибір повинен виконуватися по-своєму, залежно від головних цілей дослідження, фізичної й статистичної природи аналізованої багатомірної ознаки, апріорних відомостей про його імовірнісну природу й т.д.

Подібність або розходження між класифікуемими об'єктами встановлюється залежно від метричної відстані між ними. Якщо кожний об'єкт описується *k* ознаками, то він може бути представлений як точка в *k-мірному* просторі, і подібність із іншими об'єктами буде визначатися як відповідна відстань.

Нехай $Q = \{Q_1, Q_2, ..., Q_n\}$ - множина об'єктів, кожний з яких характеризується набором спостережуваних показників або характеристик $C = (C_1, C_2, ..., C_p)^T$. Для множини об'єктів Q дослідник має у своєму розпорядженні множину векторів вимірів $X = \{X_1, X_2, ..., X_n\}$. У випадках, коли кожний об'єкт Q_i представлений вектором ознак X_i (тобто у випадку вихідних даних, представлених у формі X), часто зручніше у формулах і різних співвідношеннях замість Q_i писати відразу X_i. Наприклад, $d(X_i, X_j)$ замість $d(Q_i, Q_j)$.

Ненегативна кількісно визначена функція $d(X_i, X_j)$ називається функцією відстані (метрикою) якщо:

а) $d(X_i, X_j) \geq 0$ для всіх X_i і X_j з p - мірного евклідового простору E_p;

б) $d(X_i, X_j) = 0$ тоді й тільки тоді, коли $X_i = X_j$;

в) $d(X_i, X_j) = d(X_j, X_i)$;

г) $d(X_i, X_j) \leq d(X_i, X_k) + d(X_k, X_j)$, де X_i, X_j и X_k - будь-які три вектори з E_p.

Значення $d(X_i, X_j)$ для заданих X_i і S_l називається відстанню між X_i і X_j й еквівалентно відстані між об'єктами Q_i й Q_j відповідно до обраних характеристик $(C_1, C_2, ..., C_p)^T$.

У кластерному аналізі використовуються різні міри відстані між об'єктами, найбільш уживаними з яких є наступні.

1) **Звичайна евклідова відстань**:

$$d_{ij} = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2},$$

де x_{ik}, x_{jk} - величина *k-ї* компоненти в i-го (j -го) об'єкта ($k = 1,2,...,m;\ i, j = 1,2,...,n$).

Використання цієї відстані виправдано у випадках, якщо:

а) спостереження X беруться з генеральних сукупностей, що мають багатомірний нормальний розподіл з коваріаційною матрицею виду $\sigma^2 E_k$, тобто компоненти X взаємно незалежні й мають ту саму дисперсію, де E_k - одинична матриця;

б) компоненти вектора спостережень X однорідні по своєму фізичному змісту й однаково важливі для класифікації;

в) ознаковий простір збігається з геометричним простором.

2) **Зважена евклідова відстань**:

$$d_{ij} = \sqrt{\sum_{k=1}^{m} w_k (x_{ik} - x_{jk})^2}.$$

Вона застосовується в тих випадках, коли кожному компоненту x_k вектора спостережень X вдається приписати деяку «вага» w_k, пропорційну ступеню важливості ознаки в задачі класифікації. Звичайно приймають $0 \le w_k \le 1$, де $k = 1,2,...,m$.

Визначення «ваг», як правило, пов'язане з додатковими дослідженнями, наприклад, організацією опитування експертів і обробкою їхніх думок. Визначення ваг w_k тільки за даними вибірки може привести до помилкових висновків.

3) **Квадрат евклідової відстані:**

$$d_{ij} = \sum_{k=1}^{m} w_k (x_{ik} - x_{jk})^2.$$

Звичайну евклідову відстань підносять до квадрата, щоб додати більші ваги більше віддаленим друг від друга об'єктам.

4) **Хеммінгова відстань:**

$$d_{ij} = \sum_{k=1}^{m} \left| x_{ik} - x_{jk} \right|.$$

Вона використовується як міра розходження об'єктів, що задають діхотомічними ознаками. Хеммінгова відстань дорівнює числу розбіжностей значень відповідних ознак у розглянутих i-му і j-му об'єктах.

5) **Відстань Чебишева:**

$$d_{ij} = \max \left| x_{ik} - x_{jk} \right|.$$

Ця відстань може виявитися корисною, коли бажають визначити два об'єкти як «різні», якщо вони розрізняються по якій-небудь одній координаті (яким-небудь одним виміром).

6) **Степінна відстань:**

$$d_{ij} = \left(\sum_{k=1}^{m} \left| x_{ik} - x_{jk} \right|^{p} \right)^{1/r},$$

де r й p - параметри, обумовлені користувачем. Параметр p відповідальний за поступове зважування різниць по окремих координатах, параметр r - за прогресивне зважування великих відстаней між об'єктами. Якщо обидва параметри - r і p, рівні 2, то ця відстань збігається з евклідовою.

Ця відстань застосовується в тих випадках, коли бажають прогресивно збільшити або зменшити вагу, що відноситься до розмірності, для якої відповідні об'єкти сильно відрізняються.

7. **Відстань Махаланобіса:**

$$d_{ij} = (X_i - X_j)^T W^{-1} (X_i - X_j),$$

де W^{-1} - матриця, зворотна матриці розсіювання.

Відстань Махаланобіса часто називають узагальненою евклідовою відстанню. Вона інваріантна відносно невироджених лінійних перетворень. Іншою властивістю відстані Махаланобіса є те, що вона робить деякі критерії кластеризації еквівалентними. Наступні три критерії кластеризації при користуванні відстанню Махаланобіса еквівалентні: 1) *tr W* (слід матриці W); 2) $|T|/|W|$; 3) $tr\ W^{-1}B$, де T, B, W - відповідно матриці повної, міжгрупової й внутрішньогрупової відстані.

Оцінка близькості між об'єктами сильно залежить від абсолютного значення ознаки й від ступеня її варіації в сукупності. Щоб усунути подібний вплив на процедуру класифікації, можна значення вихідних змінних нормувати одним з наступних способів:

1) $z_{ij} = \dfrac{x_{ij} - \bar{x}_j}{\sigma_j}$, 2) $z_{ij} = \dfrac{x_{ij}}{x_{\max j}}$, 3) $z_{ij} = \dfrac{x_{ij}}{\bar{x}_j}$, 4) $z_{ij} = \dfrac{x_{ij}}{x_{\min j}}$.

Однак ця операція може привести до небажаних наслідків. Якщо кластери добре розділені по одній ознаці й не розділені по іншому, то після нормування дискримінуючі можливості першої ознаки будуть зменшені у зв'язку зі збільшенням «шумового» ефекту другого.

Вибір міри відстані й ваг для класифікуемих змінних - дуже важливий етап кластерного аналізу, тому що від цих процедур залежать склад і кількість формованих кластерів, а також ступінь подібності об'єктів усередині кластерів. Залежно від типів вихідних змінних вибирається один з видів показників, що характеризують близькість між ними.

3. Відстань між кластерами

При конструюванні різних процедур класифікації в ряді випадків виявляється необхідним введення поняття відстані між групами об'єктів (кластерами). Основні зусилля в розвитку методів кластеризації й класифікації були спрямовані на побудову методів, заснованих на мінімізації внутрігрупових сум квадратів (відхилень). Вони можуть бути виражені в термінах евклідових відстаней і називаються методами мінімальної дисперсії.

Найбільш уживаними методами об'єднання кластерів є наступні.

1) ***Метод мінімальної локальної відстані («найближчого сусіда»)***. У цьому методі відстань між двома кластерами визначається відстанню між двома найбільш близькими об'єктами (найближчими сусідами) у різних кластерах:

$$d(S_l, S_m) = \min_{\substack{x_i \in S_l \\ x_j \in S_m}} d_{ij}.$$

2) ***Метод максимальної локальної відстані («далекого сусіда»)***. У цьому методі відстані між кластерами визначаються найбільшою відстанню між будь-якими двома об'єктами в різних кластерах:

$$d(S_l, S_m) = \max_{\substack{x_i \in S_l \\ x_j \in S_m}} d_{ij}.$$

3) ***Центроїдний метод***. У цьому методі відстань між двома кластерами визначається як відстань між їхніми центрами тяжіння:

$$d(S_l, S_m) = d(\bar{x}_l, \bar{x}_m).$$

4) ***Метод групових середніх***. У цьому методі відстані між двома кластерами визначається як середнє арифметичне всіх попарних відстаней між представниками розглянутих груп:

$$d_{cp} = (S_l, S_m) = \frac{1}{n_l n_m} \sum_{x_i \in S_l} \sum_{x_j \in S_m} d(x_i, x_j).$$

5) ***Метод медіанного зв'язку***. Відстань між будь-яким кластером S і новим кластером, що вийшов у результаті об'єднання кластерів P і U, визначається як відстань від центра кластера S до середини відрізка, що з'єднує центри кластерів P і U.

6) ***Метод Уорда***. Даний метод припускає, що на першому кроці кожний кластер складається з одного об'єкта. Спочатку поєднуються два найближчих кластери. Для них визначаються середні значення кожної ознаки й розраховується сума квадратів відхилень V_k:

$$V_k = \sum_{i=1}^{n_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{jk})^2,$$

де k - номер кластера;

i - номер об'єкта;

j - номер ознаки;

p - кількість ознак, що характеризують кожний об'єкт;

n_k - кількість об'єктів в k-м кластері.

Надалі на кожному кроці роботи алгоритму поєднуються ті об'єкти або кластери, які дають найменший приріст величини V_k. Метод Уорда приводить до утворення кластерів приблизно рівних розмірів з мінімальною внутрішньокластерною варіацією. У підсумку всі об'єкти виявляються об'єднаними в один кластер.

Якщо процес об'єднання в ієрархічному кластерному аналізі проводиться вручну, те оцінка подібності може бути дана або візуально по матриці відстаней між об'єктами D, або по узагальненій формулі:

$$d_{l,(mq)} = d(S_l, S_{(mq)}) = \alpha d_{lm} + \beta d_{lq} + \gamma d_{mq} + \delta |d_{lm} - d_{lq}|,$$

де $d_{l,(mq)}$ - відстань між класом S_l і класом $S_{(mq)}$, що є об'єднанням двох інших класів S_m і S_q;

$d_{lm} = d(S_l, S_m)$; $d_{lq} = d(S_l, S_q)$; $d_{mq} = d(S_m, S_q)$ - відстані між класами S_l, S_m й S_q;

α, β, γ і δ - числові коефіцієнти, значення яких визначають специфіку процедури, її алгоритм.

Наприклад, при $\alpha = \beta = -\delta = 1/2$ й $\gamma = 0$ приходимо до відстані, побудованій за принципом «найближчого сусіда». При $\alpha = \beta = \delta = 1/2$ й $\gamma = 0$ відстань між класами визначається за принципом «далекого сусіда», тобто як відстань між двома самими далекими елементами цих класів. І нарешті, при

$$\alpha = \frac{n_m}{n_m + n_q},\ \beta = \frac{n_q}{n_m + n_q},\ \gamma = \delta = 0$$

одержуємо відстань d між класами, обчислену як середня з відстаней між всіма парами елементів, один із яких береться з одного класу, а другий – з іншого класу.

4. Алгоритми ієрархічного й дивізимного методів кластерного аналізу

Із всіх методів кластерного аналізу найпоширенішими є ієрархічні агломеративні методи. Сутність цих методів полягає в тім, що на першому кроці кожний об'єкт вибірки розглядається як окремий кластер. Процес об'єднання кластерів відбувається послідовно: на підставі матриці відстаней або матриці подібності поєднуються найбільш близькі об'єкти. Якщо матриця подібності спочатку має розмірність $m \times m$, то повністю процес кластеризації завершується за $m-1$ кроків, у підсумку всі об'єкти будуть об'єднані в один кластер. Послідовність об'єднання легко піддається геометричній інтерпретації й може бути представлена у вигляді графа-дерева (дендрограми). На дендрограмі вказуються номери поєднуваних об'єктів і відстань (або інша міра подібності), при якій відбулося об'єднання.

Розглянемо алгоритм ієрархічного кластерного аналізу на прикладі. Необхідно провести класифікацію п'яти підприємств, кожне з яких характеризується трьома змінними: X_1 - середньорічна вартість основних фондів, млн. грн.; X_2 - матеріальні витрати на 1 грн. виробленої продукції, коп.; X_3 - об'єм виробленої продукції, млн. грн. Значення змінних наведені в табл. 2.1.

Таблиця 2.1

Вихідні дані

Номер підприємства	X_1	X_2	X_3
1	120,0	94,0	164,0
2	85,0	75,2	92,0
3	145,0	81,0	120,0
4	78,0	76,8	86,0
5	70,0	75,9	104,0
Середнє значення ($\bar{x}_j$)	99,6	80,6	113,2
Середнє квадратичне відхилення (σ)	28,4	10,9	27,9

Перед тим як обчислювати матрицю відстаней, нормуємо вихідні дані по формулі:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_{j`}}.$$

Матриця значень нормованих змінних буде мати вигляд:

$$Z = \begin{pmatrix} 0{,}718 & 1{,}229 & 1{,}821 \\ -0{,}514 & -2{,}238 & -0{,}760 \\ 1{,}598 & 0{,}037 & 0{,}244 \\ -0{,}760 & -0{,}349 & -0{,}975 \\ -1{,}042 & -0{,}431 & -0{,}330 \end{pmatrix}.$$

Класифікацію проведемо за допомогою ієрархічного агломеративного методу. Для побудови матриці відстаней скористаємося евклідовою відстанню. Тоді, наприклад, відстань між першим і другим об'єктами:

$$d_{12} = \left\{[0{,}718-(-0{,}514)]^2 + [1{,}229-(-2{,}238)]^2 + [1{,}821-(-0{,}760)]^2\right\}^{1/2} = 4{,}49.$$

Первісна матриця відстаней D_0 характеризує відстані між окремими об'єктами, кожний з яких на першому кроці є окремим кластером:

$$D_0 = \begin{pmatrix} 0 & 4{,}49 & 2{,}16 & 3{,}53 & 3{,}24 \\ & 0 & 3{,}26 & 1{,}92 & 1{,}93 \\ & & 0 & 2{,}68 & 2{,}74 \\ & & & 0 & 0{,}71 \\ & & & & 0 \end{pmatrix}.$$

Як видно по елементах матриці D_0, найбільш близькими є об'єкти n_4 й n_5 ($d_{45} = 0{,}71$). Об'єднаємо їх в один кластер і привласнимо йому номер S_4. Перерахуємо відстані всіх об'єктів (кластерів), що залишилися, до кластера S_4, та одержимо нову матрицю D_1 відстаней :

$$D_1 = \begin{pmatrix} 0 & 4{,}49 & 2{,}16 & 3{,}53 \\ & 0 & 3{,}26 & 1{,}92 \\ & & 0 & 2{,}74 \\ & & & 0 \end{pmatrix}.$$

У матриці D_1 відстані між кластерами визначені по алгоритму «далекого сусіда». Тоді відстань між об'єктом n_1 і кластером S_4 дорівнює:

$$d_{S_1,S_4} = \max\{d_{14}, d_{15}\} = \max\{3{,}53; 3{,}24\} = 3{,}53 \quad \text{і т.д.}$$

У матриці D_1 знову знаходимо найближчі кластери. Це будуть S_2 і S_4, оскільки $d_{24} = 1{,}93$. Отже, на цьому кроці поєднуємо кластери S_2 й S_4; одержимо новий кластер, що містить об'єкти n_2, n_4, n_5. Привласнимо йому номер S_2. Тепер маємо три кластери $S_1\{1\}, S_2\{2,4,5\}, S_3\{3\}$. Перераховуємо відстані d_{12} й d_{23}, одержуємо матрицю D_2 :

$$d_{12} = \max\{d_{1,2}, d_{1,IV}\} = \max\{4{,}49; 3{,}53\} = 4{,}49.$$

$$d_{23} = \max\{d_{2,3}, d_{3,IV}\} = \max\{3{,}26; 2{,}74\} = 3{,}26.$$

$$D_2 = \begin{pmatrix} 0 & 4{,}49 & 2{,}16 \\ & 0 & 3{,}26 \\ & & 0 \end{pmatrix}.$$

Судячи з матриці D_2, на наступному кроці поєднуємо кластери S_1 й S_3 ($d_{13} = 2{,}16$) в один кластер і привласнимо йому номер S_1. Тепер маємо тільки два кластери:

$$\left.\begin{array}{l} S_1 \text{ кластер (ообъект } n_1, \mathrm{n}_3) \\ S_2 \text{ кластер (ообъект } n_2, \mathrm{n}_4, \mathrm{n}_5) \end{array}\right\} d_{12} = \max\{d_{1,2}, d_{3,2}\} = \max\{4{,}49; 3{,}26\} = 4{,}49.$$

$$D_3 = \begin{pmatrix} 0 & 4{,}49 \\ & 0 \end{pmatrix}.$$

І нарешті, на останньому кроці поєднуємо кластери S_1 й S_2 на відстані 4,49. Представимо результати класифікації у вигляді дендрограми (рис. 2.1). Дендрограма свідчить про те,

що кластер S_2 більше однорідний по складу вхідних об'єктів, тому що в ньому об'єднання відбувалося при менших відстанях, чим у кластері S_1.

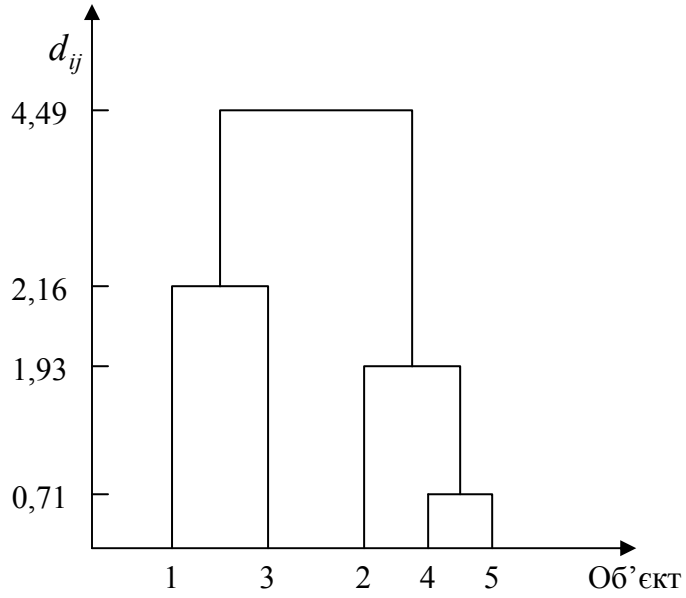

Рис. 2.1. Дендрограма кластеризації п'яти об'єктів

Розглянутий приклад проведення класифікації дозволяє зробити висновок про те, що алгоритм ієрархічного кластерного аналізу можна представити у вигляді послідовності процедур:

крок 1. Значення вихідних змінних нормуються одним з відомих способів.

крок 2. Розраховується матриця відстаней або матриця мір подібності.

крок 3. Визначається пара найближчих кластерів. По обраному алгоритму поєднуються ці два кластери. Новому кластеру привласнюється менший з номерів поєднуваних кластерів.

крок 4. Процедури 2, 3 і 4 повторюються доти, поки всі об'єкти не будуть об'єднані в один кластер або до досягнення заданого «порога» подібності.

Крім розглянутих агломеративних методів ієрархічного кластерного аналізу існують методи, протилежні їм по логічній побудові процедур класифікації. Вони називаються *ієрархічні дивізимні методи*. Основною вихідною посилкою дивізимних методів є те, що спочатку всі об'єкти належать одному кластеру (класу). У процесі класифікації за певними правилами поступово від цього кластера відокремлюються групи схожих між собою об'єктів. Таким чином, на кожному кроці кількість кластерів зростає, а міра відстані між кластерами зменшується. Дендрограма для дивізимних ієрархічних методів зображена на рис. 2.2.

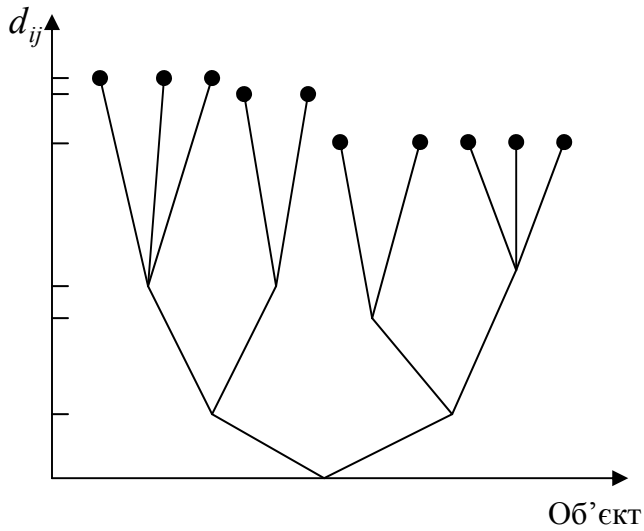

Рис. 2.2. Дендрограма ієрархічного дивізимного алгоритму

Розглянемо приклад. Нехай дана наступна матриця відстаней між об'єктами:

$$D = \begin{pmatrix} 0 & 4{,}49 & 2{,}16 & 3{,}53 & 3{,}24 \\ & 0 & 3{,}26 & 1{,}92 & 1{,}93 \\ & & 0 & 2{,}68 & 2{,}74 \\ & & & 0 & 0{,}71 \\ & & & & 0 \end{pmatrix}.$$

Проведемо класифікацію по дивізимному алгоритму. Найбільш віддаленими є об'єкти n_1 й n_2; оцінимо відстані об'єктів, що залишилися, до першого й другого об'єктів:

$d_{31} < d_{32}$ - об'єкт n_3 ближче до n_1,

$d_{41} > d_{42}$ - об'єкт n_4 ближче до n_2,

$d_{51} > d_{52}$ - об'єкт n_5 ближче к. n_2

Отже, одержали тепер два кластери: $S_1\{1,3\}$ і $S_2\{2,4,5\}$. У кожному з них аналізуємо відстані між об'єктами, і на черговому кроці відбувається поділ того кластера, де досягається максимум відстані між об'єктами.

$$d_{13} = 2{,}16, \qquad d_{25} = 1{,}93, \qquad d_{24} = 1{,}92, \qquad d_{45} = 0{,}71.$$

Найбільша відстань $d_{13} = 2{,}16$, отже, об'єкти n_1 й n_3 виділяємо в окремі кластери. У кластері $S_2\{2,4,5\}$ шукаємо максимальну відстань $\{d_{24}, d_{25}, d_{45}\} = 1{,}93$. На наступному кроці із цього кластера виділяємо об'єкт n_2, і, нарешті, на останньому кроці розділяємо кластер $S_4\{4,5\}$ на два кластери на відстані 0,71.

Дендрограма процесу класифікації по ієрархічному дивізимному методу представлена на рис. 2.3.

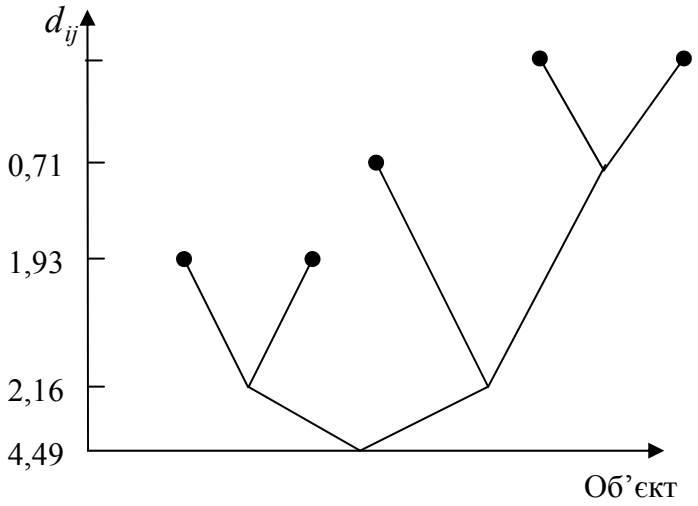

Рис. 2.3. Дендрограма кластеризації об'єктів по ієрархічному дивізимному методу

Як видно із цього приклада, дивізимний алгоритм не вимагає перерахування матриці відстаней на кожному кроці класифікації на відміну від агломеративних методів, що сприяє зниженню трудомісткості розрахунків.

Контрольні запитання.

1. Що таке кластерний аналіз?
2. Яка мета кластерного аналізу?
3. Чим обумовлена необхідність розвитку методів кластерного аналізу і їхнього використання?
4. Які задачі дозволяють вирішувати методи кластерного аналізу?
5. У чому суть агломеративних методів кластерного аналізу?
6. У чому суть дивізимних методів кластерного аналізу?
7. Чому вибір способу обчислення відстані між об'єктами є вузловим моментом дослідження?
8. Які відстані між об'єктами найбільш часто використовуються в задачах кластерного аналізу?
9. Які показники можуть бути використані в якості мір подібності?
10. У чому полягає сутність ієрархічних агломеративних методів?
11. Як будується дендрограма?
12. Які відстані між групами об'єктів є найбільш уживаними?
13. Чому якість проведення кластеризації залежить від алгоритму об'єднання в ієрархічних агломеративних методах?
14. У вигляді яких послідовних процедур можна представити алгоритм ієрархічного кластерного аналізу?

Тема 3. Кластерний аналіз. Ітеративні методи класифікації

1. Метод k - середніх
2. Метод пошуку згущень
3. Критерії якості класифікації

Після вивчення лекції студент повинен:
Знати:

- сутність *ітеративних методів* кластерного аналізу;

- кроки проведення процедури класифікації за допомогою ітеративних методів;
- основні переваги та недоліки ітеративного алгоритму типу «форель»;
- яку розбивку за обраним функціоналом варто вважати найкращою.

Вміти:
- здійснити вибір алгоритму класифікації;
- грамотоно та змістовно використовувати метод k-середніх;
- провести оцінку стійкості отриманої розбивки.

1. Метод k - середніх

Поряд з ієрархічними методами класифікації існує численна група так званих *ітеративних методів* кластерного аналізу. Сутність їх полягає в тім, що процес класифікації починається із завдання деяких початкових умов (кількість утворених кластерів, поріг завершення процесу класифікації й т.д.). Ітеративні методи більшою мірою, чим ієрархічні, жадають від користувача інтуїції при виборі типу класифікаційних процедур і завдання початкових умов розбивки, тому що більшість цих методів дуже чутливі до зміни параметрів, що задають. Наприклад, обране випадковим образом число кластерів може не тільки сильно збільшити трудомісткість процесу класифікації, але й привести до утворення «розмитих» або мало наповнюваних кластерів. Тому доцільно спочатку провести класифікацію по одному з ієрархічних методів або на підставі експертних оцінок, а потім уже підбирати початкову розбивку й статистичний критерій для роботи ітераційного алгоритму. Як і в ієрархічному кластерному аналізі, в ітераційних методах існує проблема визначення числа кластерів. У загальному випадку їхнє число може бути невідомо. Не всі ітеративні методи вимагають первісного завдання числа кластерів. Але для остаточного рішення питання про структуру досліджуваної сукупності можна випробувати кілька алгоритмів, міняючи або число утворених кластерів, або встановлений поріг близькості для об'єднання об'єктів у кластери. Тоді з'являється можливість вибрати найкращу розбивку згідно критерію якості, що задається.

Метод k-середніх належить до групи ітеративних методів еталонного типу. Сама назва методу була запропонована Дж. Мак-Куіном в 1967 р.

На відміну від ієрархічних процедур метод k-середніх не вимагає обчислення й зберігання матриці відстаней або подібностей між об'єктами. Алгоритм цього методу припускає використання тільки вихідних значень змінних. Для початку процедури класифікації повинні бути задані k випадково обраних об'єктів, які будуть служити еталонами, тобто центрами кластерів. Уважається, що алгоритми еталонного типу зручні й швидкодіючі. У цьому випадку важливу роль грає вибір початкових умов, які впливають на тривалість процесу класифікації й на його результати.

Метод k-середніх зручний для обробки великих статистичних сукупностей. У дослідженнях Хартігана, Болла, В.Н. Йолкіної, Н.Г. Загоруйко пропонуються різні модифікації цього методу.

Розглянемо математичний опис алгоритму методу k-середніх (Мак-Куіна).

Нехай є n спостережень, кожне з яких характеризується p ознаками $X_1, X_2, ... X_p$. Ці спостереження необхідно розбити на k кластерів. Для початку із n точок досліджуваної сукупності відбираються випадковим чином або задаються дослідником виходячи з яких-небудь апріорних міркувань k точок (об'єктів). Ці точки приймаються за еталони. Кожному еталону привласнюється порядковий номер, що одночасно є й номером кластера. На першому кроці із $(n-k)$ об'єктів, що залишилися, витягається точка X_i, з координатами $x_{i1}, x_{i2}, ... x_{ip}$ й перевіряється, до якому з еталонів (центрів) вона перебуває ближче всього. Для цього використовується одна з метрик, наприклад, евклідова відстань:

$$d_{il} = \sqrt{\sum_{j=1}^{p} (x_{ij} - x_{lj})^2} .$$

Об'єкт, що перевіряється, приєднується до того центра (еталону) , якому відповідає $\min d_{il} (l = 1,..., k)$. Еталон заміняється новим, переліченим з урахуванням приєднаної точки, і вага його (кількість об'єктів, що входять у даний кластер) збільшується на одиницю. Якщо зустрічаються дві або більше мінімальних відстані, то i-й об'єкт приєднують до центра з найменшим порядковим номером. На наступному кроці вибираємо точку X_{i+1} й для неї повторюються всі процедури. Таким чином, через (*n* — *k*) кроків всі точки (об'єкти) сукупності виявляться віднесеними до одному з *k* кластерів, але на цьому процес розбивки не закінчується. Для того щоб домогтися стійкості розбивки по тому ж правилу, всі точки $X_1, X_2, ... X_n$ знову приєднуються до отриманих кластерів, при цьому ваги продовжують накопичуватися. Нова розбивка рівняється з попередньою. Якщо вони збігаються, то робота алгоритму завершується. У противному випадку цикл повторюється. Остаточна розбивка має центри тяжіння, які не збігаються з еталонами, їх можна позначити $C_1, C_2, ... C_k$. При цьому кожна точка $X_i = (i = 1, 2, ... n)$ буде відноситися до такого кластера (класу) l, для якого

$$d(x_j, c_l) = \min_{1 \le j \le R} = d(x_j, C_j).$$

Можливі дві модифікації методу k-середніх. Перша припускає перерахування центра тяжіння кластера після кожної зміни його складу, а друга - лише після того, як буде завершений перегляд всіх даних. В обох випадках ітеративний алгоритм цього методу мінімізує дисперсію усередині кожного кластера, хоча в явному виді такий критерій оптимізації не використовується.

Розглянутий метод k-середніх допускає у якості вихідної розбивки використати угруповання, отримане одним з методів ієрархічного кластерного аналізу. Такий підхід можна рекомендувати для скорочення часу обробки в тому випадку, коли сукупність об'єктів досить велика і користувач утрудняється вказати кількість утворених кластерів.

Розрахункові процедури більшості ітеративних методів класифікації зводяться до виконання наступних кроків:

Крок 1. Вибір числа кластерів, на які повинна бути розбита сукупність, завдання первісної розбивки об'єктів і визначення центрів тяжіння кластерів.

Крок 2. Відповідно до обраних мір подібності визначення нового складу кожного кластера.

Крок 3. Після повного перегляду всіх об'єктів і розподілу їх по кластерах здійснюється перерахування центрів тяжіння кластерів.

Крок 4. Процедури 2 і 3 повторюються доти, поки наступна ітерація не дасть такий же склад кластерів, що й попередня.

2. Метод пошуку згущень

Одним з ітеративних методів класифікації, не потребуючих завдання числа кластерів, є *метод пошуку згущень*.

У теорії й на практиці існує кілька різних модифікацій цього методу. Кожна модифікація відрізняється початковим станом, що задається, і критеріями завершення класифікації. Зупинимося докладно на одному з алгоритмів пошуку згущень, що одержав назву «форель». Суть ітеративного алгоритму типу «форель» полягає в застосуванні гіперсфери заданого радіуса, що переміщається в просторі класифікаційних ознак з метою пошуку локальних згущень точок. Розглянемо схему даного алгоритму в загальному виді й на конкретному прикладі.

Метод пошуку згущень вимагає обчислення матриці відстаней (або матриці мер подібності) між об'єктами. Потім вибирається об'єкт, що є первісним центром першого кластера. Вибір такого об'єкта може бути довільним, а може ґрунтуватися на попередньому аналізі точок і їхніх околиць. При використанні другого підходу можна значно скоротити число ітерацій, що приводять до розподілу всіх точок по кластерах.

Обрана точка приймається за центр гіперсфери заданого радіуса *R*. Визначається сукупність точок, що потрапили усередину цієї сфери, і для них обчислюються координати центра (вектор середніх значень ознак). Далі знову розглядаємо гіперсферу такого ж радіуса, але з новим центром, і для сукупності точок, що потрапили в неї, знову розраховуємо вектор середніх значень, приймаємо його за новий центр сфери й т.д. Коли чергове перерахування координат центра сфери приводить до такого ж результату, як і на попередньому кроці, переміщення сфери припиняється, а точки, що потрапили в неї, утворять кластер і з подальшого процесу кластеризації виключаються. Для всіх точок, що залишилися, процедури повторюються, тобто знову вибирається довільний об'єкт, що є первісним центром сфери радіуса *R*, і т.д.

Таким чином, робота алгоритму завершується за кінцеве число кроків і всі точки виявляються розподіленими по кластерах. Число кластерів, що утворилися, заздалегідь невідомо й сильно залежить від вибору радіуса сфери. Деякі модифікації алгоритму дозволяють розділити сукупність на задане число кластерів шляхом послідовної зміни радіуса сфери.

Для оцінки стійкості отриманої розбивки доцільно повторити процес кластеризації кілька разів для різних значень радіуса сфери, змінюючи щораз радіус на невелику величину.

Існує кілька способів вибору радіуса сфери. Якщо d_{lk} - відстань між l-м і *k-м* об'єктами, то як нижня границя радіуса R_H обирається $R_H = \min\{d(X_l, X_k)\}$, а верхня границя радіуса R_B може бути визначена як $R_B = \max\{d(X_l, X_k)\}$.

Якщо починати роботу алгоритму з величини $R = \min d(X_l, X_k) + \delta$ й при кожному його повторенні змінювати δ на невелику величину, то можна виявити значення радіусів, які приводять до утворення того самого числа кластерів, тобто до стійкої розбивки.

3. Критерії якості класифікації

При використанні різних методів кластерного аналізу для однієї й тієї ж сукупності можуть бути отримані різні варіанти розбивки. Істотний вплив на характеристики кластерної структури роблять: по-перше, набір ознак, по яких здійснюється класифікація, по-друге, тип обраного алгоритму. Наприклад, ієрархічні й ітеративні методи приводять до утворення різного числа кластерів. При цьому самі кластери розрізняються й по складу, і по ступені близькості об'єктів. Вибір міри подібності також впливає на результат розбивки. Якщо використовуються методи з еталонними алгоритмами, наприклад, метод k-середніх, то початкові умови розбивки, що задають, в значній мірі визначають кінцевий результат розбивки.

Після завершення процедур класифікації необхідно оцінити отримані результати. Для цієї мети використовується деяка міра якості класифікації, що прийнято називати функціоналом або критерієм якості. Найкращою по обраному функціоналу варто вважати таку розбивку, при якій досягається екстремальне (мінімальне або максимальне) значення цільової функції - функціонала якості.

У більшості випадків алгоритми класифікації й критерії якості зв'язані між собою, тобто певний алгоритм забезпечує одержання екстремального значення відповідного функціонала якості. Наприклад, використання методу Уорда приводить до одержання кластерів з мінімальною внутрішньокласовою дисперсією.

Розглянемо найпоширеніші функціонали якості.

1) Сума квадратів відстаней до центрів класів:

$$F_1 = \sum_{l=1}^{k} \sum_{i \in S_1} d^2\left(X_i, \overline{X}_l\right),$$

де l - номер кластера $(l = 1, 2, ..., k)$;

$\overline{X}$ - центр l-го кластера;

X_i - вектор значень змінних для i-го об'єкта, що входить в i-й кластер;

$d\left(X_i, \overline{X}_l\right)$ — відстань між i-м об'єктом і центром l-го кластера.

При використанні цього критерію прагнуть одержати таку розбивку сукупності об'єктів на *k* кластерів, при якому значення F_1 було б мінімальним.

2) Сума внутрішньокласових відстаней між об'єктами:

$$F_2 = \sum_{l=1}^{k} \sum_{i,\, j \in S_1} d_{ij}^2 .$$

У цьому випадку найкращим варто вважати таку розбивку, при якій досягається мінімальне значення F_2, тобто отримані кластери великої «щільності». Об'єкти, що потрапили в один кластер, близькі між собою за значеннями тих змінних, які використовувалися для класифікації.

3) Сумарна внутрішньокласова дисперсія:

$$F_3 = \sum_{l=1}^{k} \sum_{j=1}^{p} \sigma_{ij}^2 ,$$

де σ_{ij}^2 - дисперсія j-ї змінної в кластері S_l.

У цьому випадку розбивку, при якій сума внутрішньокласових (внутрішньогрупових) дисперсій буде мінімальною, варто вважати оптимальною. Існує кілька алгоритмів кластерного аналізу, що забезпечують оптимальну розбивку з погляду функціонала F_3. Наприклад, ітераційний алгоритм, що включає наступні обчислювальні процедури:

а) як початкова розбивка задається розбивка сукупності об'єктів на *k* кластерів. Вона може бути отримана одним з ієрархічних методів;

б) для кожного кластера S_l визначається центр $\overline{X}_l = \left(\overline{x}_{1l}, \overline{x}_{2l}, ..., \overline{x}_{pl}\right)$. Кожна координата центра обчислюється в такий спосіб:

$$\overline{x}_{jl} = \frac{1}{n_l} \sum_{i=1}^{n_l} x_{ij} ,$$

де i - номер об'єкта;

j - номер змінної;

l — номер кластера;

n_l - кількість об'єктів у кластері S_l;

в) всі об'єкти вихідної сукупності розподіляються по кластерах залежно від їхньої відстані до центрів цих кластерів, тобто i-і об'єкт буде включений у кластер S_1 у тому випадку, якщо його відстань до центра цього кластера:

$$d_{i\overline{x}_l} = \min_{q=1...k} \{d_{i\overline{x}_q}\} .$$

Після розподілу об'єктів по *k* кластерах порівнюють первісний склад цих кластерів зі знову отриманим. Якщо виявляється розбіжність, тоді робота алгоритму триває, повторюються процедури (б) і (в). Локальний екстремум досягається в тому випадку, якщо збігаються результати наступного й попереднього угруповань. Варто помітити, що для іншої початкової розбивки оптимальне значення функціонала F_3 буде відрізнятися. На принципі мініміс-

зації внутрішньокластерної дисперсії засновані алгоритми методу k-середніх і методу Уорда.

Контрольні запитання

1. У чому полягає суть ітеративних методів кластерного аналізу?
2. Яким чином в ітеративних методах задаються початкові умови?
3. Яким чином проводиться класифікація об'єктів методом k-середніх?
4. До виконання яких кроків зводяться обчислювальні процедури більшості ітеративних методів класифікації?
5. У чому складається істотна відмінність методу пошуку згущень від інших ітеративних методів класифікації?
6. У чому полягає суть ітеративного алгоритму типу «форель»?
7. Якими способами вибирається радіус сфери для пошуку локальних згущень точок у методі пошуку згущень?
8. Для якої мети використовуються критерії якості класифікації?
9. Які найпоширеніші функціонали якості використовуються в кластерному аналізі?
10. Які найпростіші прийоми дозволяють судити про якість розбивки об'єктів на кластери?

Тема 4. Дискримінантний аналіз

1. Основні положення дискримінантного аналізу.
2. Дискримінантні функції і їхня геометрична інтерпретація.
3. Класифікація при наявності навчальних вибірок.
4. Взаємозв'язок між дискримінантними змінними й дискримінантними функціями.

Після вивчення лекції студент повинен:
Знати:
- Визначення та зміст дискримінантного аналізу;
- Алгоритм механізму визначення виду та обчислення дискримінантної функції;
- Геометричну інтерпретацію дискримінантної функції;
- У яких випадках лінійну дискримінантну функцію можна застосовувати у якості опису поділяючої поверхні між множинами.

Вміти:
- визначити коефіцієнти дискримінантної функції;
- знайти коефіцієнти дискримінантної функції;
- стандартизувати значення вихідних змінних таким чином, щоб їхні середні значення дорівнювали нулю, а дисперсії - одиниці;
- обчислити стандартизовані коефіцієнти виходячи зі значень коефіцієнтів у нестандартній формі.

1. Основні положення дискримінантного аналізу

1. ***Дискримінантний аналіз*** - це розділ математичної статистики, змістом якого є розробка методів рішення задач розрізнення (дискримінації) об'єктів спостереження по певних ознаках. Наприклад, розбивка сукупності підприємств на кілька однорідних груп за значеннями яких-небудь показників виробничо-господарської діяльності.

Методи дискримінантного аналізу знаходять застосування в різних областях: медицині, соціології, психології, економіці й т.д. При спостереженні великих статистичних сукупностей часто з'являється необхідність розділити неоднорідну сукупність на однорідні групи

(класи). Таке розчленовування надалі при проведенні статистичного аналізу дає кращі результати моделювання залежностей між окремими ознаками.

Дискримінантний аналіз виявляється дуже зручним і при обробці результатів тестування окремих осіб. Наприклад, при виборі кандидатів на певну посаду можна всіх опитуваних претендентів розділити на дві групи: «підходить» і «не підходить».

Можна привести ще один приклад застосування дискримінантного аналізу в економіці. Для оцінки фінансового стану своїх клієнтів при видачі їм кредиту банк класифікує їх на надійні й ненадійні по ряду ознак. Таким чином, у тих випадках, коли виникає необхідність віднесення того або іншого об'єкта до одного з реально існуючим або виділених певним способом класів, можна скористатися дискримінантним аналізом.

Апарат дискримінантного аналізу розроблявся багатьма вченими-фахівцями, починаючи з кінця 50-х років XX в. Дискримінантним аналізом, як і іншими методами багатомірної статистики, займалися П.Ч. Махаланобіс, Р. Фішер, Г. Хотедлінг і інші видні вчені.

Всі процедури дискримінантного аналізу можна розбити на дві групи й розглядати їх як зовсім самостійні методи. Перша група процедур дозволяє інтерпретувати розходження між існуючими класами, друга - проводити класифікацію нових об'єктів у тих випадках, коли невідомо заздалегідь, до якому з існуючих класів вони належать.

Нехай є множина одиниць спостереження - генеральна сукупність. Кожна одиниця спостереження характеризується декількома ознаками (змінними) x_{ij} — значення j-ї змінної в i-го об'єкта $i = 1,..., N; \quad j = 1,..., p.$

Припустимо, що вся множина об'єктів розбита на кілька підмножин (дві й більше). З кожної підмножини взята вибірка об'ємом n_k, де k - номер підмножини (класу), $k = 1, ..., q$.

Ознаки, які використовуються для того, щоб відрізняти один клас (підмножину) від іншого, називаються ***дискримінантними змінними***. Кожна із цих змінних повинна вимірятися або по інтервальній шкалі, або по шкалі відносин. Інтервальна шкала дозволяє кількісно описати розходження між властивостями об'єктів. Для завдання шкали встановлюються довільна точка відліку й одиниця виміру. Прикладами таких шкал є календарний час, шкали температур і т.п. Як оцінка положення центра використовуються середня величина, мода й медіана.

Шкала відносин - окремий випадок інтервальної шкали. Вона дозволяє співвіднести кількісні характеристики якої-небудь властивості в різних об'єктів, наприклад, стаж роботи, заробітна плата, величина податку.

Теоретично число дискримінантних змінних не обмежене, але на практиці їхній вибір повинен здійснюватися на підставі логічного аналізу вихідної інформації й одного із критеріїв, про яке мова йтиме трохи нижче. Число об'єктів спостереження повинне перевищувати число дискримінантних змінних, як мінімум, на два, тобто $p < N$. Дискримінантні змінні повинні бути лінійно незалежними. Ще одним припущенням при дискримінантному аналізі є нормальність закону розподілу багатомірної величини, тобто кожна з дискримінантних змінних усередині кожного з розглянутих класів повинна бути підлегла нормальному закону розподілу. У випадку, коли реальна картина у вибіркових сукупностях відрізняється від висунутих передумов, варто вирішувати питання про доцільність використання процедур дискримінантного аналізу для класифікації нових спостережень, тому що в цьому випадку утруднюються розрахунки кожного критерію класифікації.

2. Дискримінантні функції і їхня геометрична інтерпретація

Перед тим як приступити до розгляду алгоритму дискримінантного аналізу, звернемося до його геометричної інтерпретації.

На рис. 4.1 зображені об'єкти, що належать двом різним множинам M_1 і M_2. Кожний об'єкт характеризується в цьому випадку двома змінними x_1 й x_2. Якщо розглядати

проекції об'єктів (точок) на кожну вісь, то ці множини перетинаються, тобто по кожній змінній окремо деякі об'єкти обох множин мають подібні характеристики. Щоб щонайкраще розділити дві розглянутих множини, потрібно побудувати відповідну лінійну комбінацію змінних x_1 і x_2. Для двовимірного простору ця задача зводиться до визначення нової системи координат. Причому нові осі *L* і *C* повинні бути розташовані таким чином, щоб проекції об'єктів, що належать різним множинам на вісь *L*, були максимально розділені. Вісь *C* перпендикулярна осі *L* і розділяє дві «хмари» точок щонайкраще, тобто щоб множини опинилися по різні сторони від цієї прямої. При цьому ймовірність помилки класифікації повинна бути мінімальною. Сформульовані умови повинні бути враховані при визначенні коефіцієнтів a_1 і a_2 наступної функції:

$$f(x) = a_1 x_1 + a_2 x_2 . \qquad (4.1)$$

Функція *f(x)* називається *канонічною дискримінантною функцією*, а величини x_1 й x_2; - дискримінантними змінними.

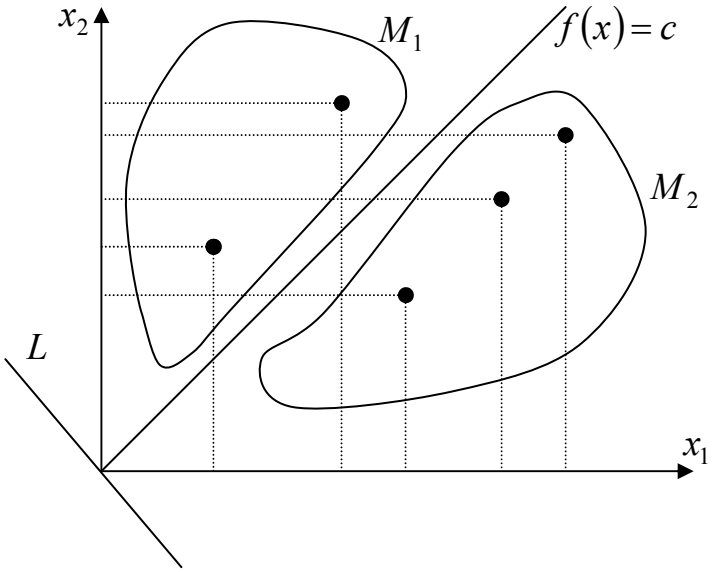

Рис. 4.1. Геометрична інтерпретація дискримінантної функції в дискримінантних змінних

Позначимо x_{ij} - середнє значення j-ї ознаки в об'єктах i-ї множини (класу). Тоді для множини M_1 середнє значення функції $f_1(x)$ буде дорівнювати:

$$\overline{f}_1(x) = a_1 \bar{x}_{11} + a_2 \bar{x}_{12};$$

для множини M_2 середнє значення функції $f_2(x)$ дорівнює:

$$\overline{f}_2(x) = a_1 \bar{x}_{21} + a_2 \bar{x}_{22} .$$

Дискримінантна функція може бути як лінійною, так і нелінійною. Вибір її виду залежить від геометричного розташування поділюваних класів у просторі дискримінантних змінних. Для спрощення викладень надалі розглядається лінійна дискримінантна функція.

Коефіцієнти дискримінантної функції α_i визначаються таким чином, щоб $\overline{f_1}(x)$ і $\overline{f_2}(x)$ якнайбільше розрізнялися між собою, тобто щоб для двох множин (класів) було максимальним вираження:

$$\overline{f_1}(x) - \overline{f_2}(x) = \sum_{i=1}^{n_1} a_1 x_{1i} - \sum_{i=1}^{n_2} a_i x_{2i} . \qquad (4.2)$$

Тоді можна записати наступне:

$$f_{kt}(x) - \overline{f}_k(x) = a_1(x_{1kt} - \bar{x}_{1k}) + a_2(x_{2kt} - \bar{x}_{2k}) + ... + a_p(x_{pkt} - \bar{x}_{pk}), \qquad (4.3)$$

де k - номер групи;

p - число змінних, що характеризують кожне спостереження.

Отримані значення коефіцієнтів підставляють у формулу (4.1) і для кожного об'єкта в обох групах (множинах) обчислюють дискримінантні функції, потім знаходять середнє значення для кожної групи. Таким чином, кожне i-е спостереження, що спочатку описувалося m змінними, буде як би поміщено в одномірний простір, тобто йому буде відповідати одне значення дискримінантної функції, отже, розмірність ознакового простору знижується.

3. Класифікація при наявності навчальних вибірок

Перед тим як приступити безпосередньо до процедури класифікації, потрібно визначити границю, що розділяє в окремому випадку дві розглянуті групи. Такою величиною може бути значення функції, рівновіддалене від $\overline{f}_1$ і $\overline{f}_2$, тобто

$$C = \frac{1}{2}(\overline{f}_1 + \overline{f}_2).$$

Величина C називається *константою дискримінації.*

На рис. 4.1 видно, що об'єкти, розташовані над прямою $f(x) = a_1x_1 + a_2x_2 + ... + a_px_p = C$, перебувають ближче до центра множини M_1 й, отже, можуть бути віднесені до першої групи, а об'єкти, розташовані нижче цієї прямої, ближче до центра другої множини, тобто ставляться до другої групи. Якщо границя між групами обрана так, як сказано вище, те сумарна ймовірність помилкової класифікації мінімальна.

Розглянемо приклад використання дискримінантного аналізу для проведення багатомірної класифікації об'єктів. При цьому в якості навчальних будемо використовувати спочатку дві вибірки, що належать двом класам, а потім узагальнимо алгоритм класифікації на випадок k класів.

Приклад 4.1. Є дані (табл. 4.1) по двох групах промислових підприємств машинобудівного комплексу:

X_1 - фондовіддача основних фондів, грн.;

X_2 - витрати на 1 грн. виробленої продукції, коп.;

X_3 -витрати сировини й матеріалів на 1 грн. продукції, коп.

Таблиця 4.1

Вихідні дані

	Номер підприємства	X_1	X_2	X_3
1-я група	1	0,50	94,0	8,50
	2	0,67	75,4	8,79
	3	0,68	85,2	9,10
	4	0,55	98,8	8,47
2-я група	5	1,52	81,5	4,95
	6	1,20	93,8	6,95
	7	1,46	86,5	4,70

Необхідно провести класифікацію чотирьох нових підприємств, що мають наступні значення вихідних змінних:

1-і підприємство: $x_1 = 1{,}07, \quad x_2 = 93{,}5, \quad x_3 = 5{,}30,$

2-і підприємство: $x_1 = 0{,}99, \quad x_2 = 84{,}0, \quad x_3 = 4{,}85,$

3-і підприємство: $x_1 = 0{,}70, \quad x_2 = 76{,}8, \quad x_3 = 3{,}50,$

4-і підприємство: $x_1 = 1{,}24, \quad x_2 = 88{,}0, \quad x_3 = 4{,}95.$

Для зручності запишемо значення вихідних змінних для кожної групи підприємств у вигляді матриць X_1 і X_2:

$$X_1 = \begin{pmatrix} 0{,}50 & 94{,}0 & 8{,}50 \\ 0{,}67 & 75{,}4 & 8{,}79 \\ 0{,}68 & 85{,}2 & 9{,}10 \\ 0{,}55 & 98{,}8 & 8{,}47 \end{pmatrix}, \qquad X_2 = \begin{pmatrix} 1{,}52 & 81{,}5 & 4{,}95 \\ 1{,}20 & 93{,}8 & 6{,}95 \\ 1{,}46 & 86{,}5 & 4{,}70 \end{pmatrix}.$$

Розрахуємо середнє значення кожної змінної в окремих групах для визначення положення центрів цих груп:

I гр. $\bar{x}_{11} = 0{,}60, \quad \bar{x}_{21} = 88{,}4, \quad \bar{x}_{31} = 8{,}72;$

II гр. $\bar{x}_{12} = 1{,}39, \quad \bar{x}_{22} = 87{,}3, \quad \bar{x}_{32} = 5{,}53.$

Дискримінантна функція $f(x)$ даному випадку має вигляд:

$$f(x) = a_1 x_1 + a_2 x_2 + a_3 x_3. \tag{4.8}$$

Коефіцієнти a_1, a_2 и a_3 обчислюються по формулі:

$$A = S_*^{-1}\left(\overline{X}_1 - \overline{X}_2\right),$$

де $\overline{X}_1, \overline{X}_2$ - вектори середніх у першій і другій групах;

A - вектор коефіцієнтів;

S_* - матриця, зворотна спільній коваріаційній матриці.

Для визначення спільної коваріаційної матриці S_* потрібно розрахувати матриці S_1 й S_2. Кожний елемент цих матриць являє собою різницю між відповідним значенням вихідної змінної x_{ij} й середнім значенням цієї змінної в даній групі $\bar{x}_{ik}$ (k — номер групи):

$$S_1 = \begin{pmatrix} 0{,}0238 & -2{,}2460 & 0{,}0698 \\ -2{,}2460 & 318{,}76 & -5{,}958 \\ 0{,}0698 & -5{,}958 & 0{,}2602 \end{pmatrix}; \quad S_2 = \begin{pmatrix} 0{,}0579 & -2{,}0450 & -0{,}4033 \\ -2{,}0450 & 76{,}530 & 13{,}2580 \\ -0{,}4033 & 13{,}258 & 3{,}0417 \end{pmatrix}.$$ Тоді спільна коваріаційна матриця S_*, буде дорівнювати:

$$S_* = \frac{1}{n_1 + n_2 - 2}\left(S_1 + S_2\right),$$

де n_1, n_2 - число об'єктів 1-й і 2-й групи;

$$S_* = \frac{1}{(4+3-2)}\begin{pmatrix} 0{,}0817 & -4{,}291 & -0{,}3335 \\ -4{,}291 & 395{,}290 & 7{,}300 \\ -0{,}3335 & 7{,}300 & 3{,}3019 \end{pmatrix} =$$

$$= \begin{pmatrix} 0{,}01634 & -0{,}8582 & -0{,}0667 \\ -0{,}8582 & 79{,}058 & 1{,}460 \\ -0{,}0667 & 1{,}460 & 0{,}6604 \end{pmatrix}.$$

Зворотна матриця S_*^{-1} буде дорівнювати:

$$S_*^{-1} = \begin{pmatrix} 339{,}970 & -3{,}190 & 27{,}290 \\ -3{,}190 & 0{,}043 & -0{,}227 \\ 27{,}290 & -0{,}227 & 8{,}380 \end{pmatrix}.$$

Звідси знаходимо вектор коефіцієнтів дискримінантної функції по формулі:

$$A = S_*^{-1}\left(\overline{X}_1 - \overline{X}_2\right) = \begin{pmatrix} 339{,}970 & -3{,}190 & 27{,}290 \\ -3{,}190 & 0{,}043 & -0{,}227 \\ 27{,}290 & -0{,}227 & 8{,}380 \end{pmatrix} \cdot \begin{pmatrix} -0{,}79 \\ 1{,}10 \\ 3{,}19 \end{pmatrix} = \begin{pmatrix} -185{,}03 \\ 1{,}84 \\ 4{,}92 \end{pmatrix},$$

тобто $a_1 = -185{,}03, \quad a_2 = 1{,}84, \quad a_3 = 4{,}92.$

Підставимо отримані значення коефіцієнтів у формулу (4.8) і розрахуємо значення дискримінантної функції для кожного об'єкта.

Для 1-ї множини:

$$\begin{cases} f_{11} = 0{,}5(-185{,}03) + 94 \cdot 1{,}84 + 8{,}5 \cdot 4{,}92 = 122{,}265; \\ f_{12} = 0{,}67(-185{,}03) + 75{,}4 \cdot 1{,}84 + 8{,}79 \cdot 4{,}92 = 58{,}0127; \\ f_{13} = 0{,}68(-185{,}03) + 85{,}2 \cdot 1{,}84 + 9{,}1 \cdot 4{,}92 = 75{,}7196; \\ f_{14} = 0{,}55(-185{,}03) + 98{,}8 \cdot 1{,}84 + 8{,}47 \cdot 4{,}92 = 121{,}6979; \end{cases}$$

$$\overline{f}_1 = 94{,}4238.$$

Для 2-ї множини:

$$\begin{cases} f_{21} = 1{,}52 - 185{,}03 + 81{,}5 \cdot 1{,}84 + 4{,}95 \cdot 4{,}92 = -106{,}9316; \\ f_{22} = 1{,}20 - 185{,}03 + 93{,}8 \cdot 1{,}84 + 6{,}95 \cdot 4{,}92 = -15{,}25; \\ f_{23} = 1{,}46 - 185{,}03 + 86{,}5 \cdot 1{,}84 + 4{,}7 \cdot 4{,}92 = -87{,}8598; \end{cases}$$

$$\overline{f_2} = -70{,}0138.$$

Тоді константа дискримінації C буде дорівнювати:

$$C = \frac{1}{2}(94{,}4238 - 70{,}0138) = 12{,}205.$$

Після одержання константи дискримінації можна перевірити правильність розподілу об'єктів у вже існуючих двох класах, а також провести класифікацію нових об'єктів.

Розглянемо, наприклад, об'єкти з номерами 1, 2, 3, 4. Для того щоб віднести ці об'єкти до однієї із двох множин, розрахуємо для них значення дискримінантних функцій (по трьох змінним):

$$f_1 = -185{,}03 \cdot 1{,}07 + 1{,}84 \cdot 93{,}5 + 4{,}92 \cdot 5{,}30 = 0{,}1339;$$
$$f_2 = -185{,}03 \cdot 0{,}99 + 1{,}84 \cdot 84{,}0 + 4{,}92 \cdot 4{,}85 = -4{,}7577;$$
$$f_3 = -185{,}03 \cdot 0{,}70 + 1{,}84 \cdot 76{,}8 + 4{,}92 \cdot 3{,}50 = 29{,}0110;$$
$$f_4 = -185{,}03 \cdot 1{,}24 + 1{,}84 \cdot 88{,}0 + 4{,}92 \cdot 4{,}95 = -43{,}1632.$$

Таким чином, об'єкти 1, 2 і 4 відносяться до другого класу, а об'єкт 3 відноситься до першого класу, тому що $f_1 < c,\ f_2 < c,\ f_3 > c,\ f_4 < c$.

Тепер розглянемо класифікацію при наявності k навчальних вибірок.

При необхідності можна проводити розбивку множини об'єктів на k класів (при $k > 2$). У цьому випадку потрібно розрахувати k дискримінантних функцій, тому що класи будуть відокремлюватися один від одного індивідуальними поділяючими поверхнями. На рис. 4.3 показаний випадок із трьома множинами й трьома дискримінантними змінними.

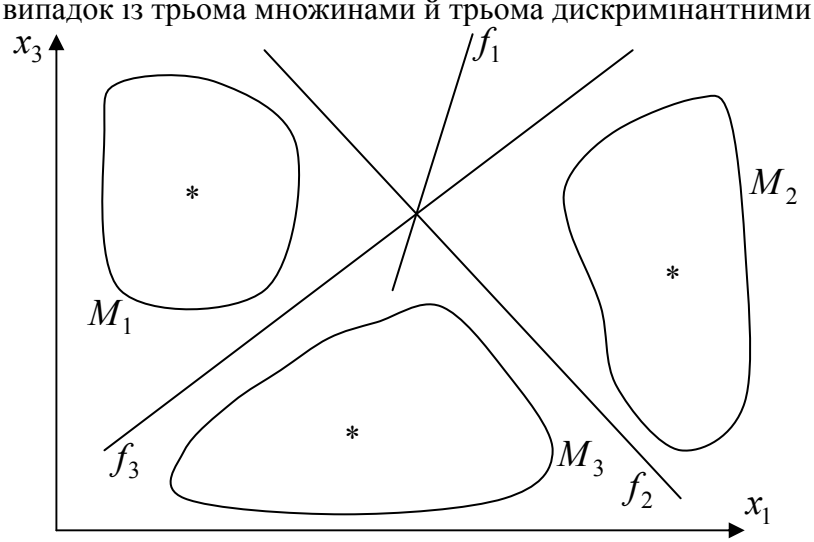

Рис. 3. Три класи об'єктів і поділяючі їх прямі:
f_1 - перша, f_2 - друга, f_3 - третя дискримінантні функції.

Таким чином, ми бачимо, що зміна числа змінних сильно впливає на результат дискримінантного аналізу. Щоб судити про доцільність включення (видалення) дискримінантной змінної, звичайно використовують спеціальні статистичні критерії, що дозволяють оцінити значимість погіршення або поліпшення розбивки після включення (видалення) кожної з відібраних змінних.

4. Взаємозв'язок між дискримінантними змінними й дискримінантними функціями

Для оцінки внеску окремої змінної в значення дискримінантної функції доцільно користуватися стандартизованими коефіцієнтами дискримінантної функції. Стандартизовані коефіцієнти можна розрахувати двома шляхами:

стандартизувати значення вихідних змінних таким чином, щоб їхні середні значення дорівнювали нулю, а дисперсії - одиниці;

обчислити стандартизовані коефіцієнти виходячи зі значень коефіцієнтів у нестандартній формі:

$$b_j = a_j \sqrt{\frac{W_{jj}}{p - m}},$$

де p - загальне число вихідних змінних;

m - число груп;

W_{jj} - елементи матриці коваріацій:

$$W_{jj} = \sum_{k=1}^{m}\sum_{i=1}^{n_k}\left(x_{ikj} - \bar{x}_{kj}\right)\left(x_{ikj} - \bar{x}_{kj}\right),$$

де i - номер спостереження;

j - номер змінної;

k - номер класу;

n_k - кількість об'єктів в k-му класі.

Стандартизовані коефіцієнти застосовують у тих випадках, коли потрібно визначити, яка з використовуваних змінних вносить найбільший вклад у величину дискримінантної функції. У прикладі із двома класами, розглянутому вище, дискримінантна функція мала вигляд:

$$f = -185{,}03X_1 + 1{,}84X_2 + 4{,}92X_3.$$

Отже, найбільший внесок у величину дискримінантної функції вносить змінна X_1.

Визначимо значення стандартизованих коефіцієнтів і запишемо нове значення дискримінантної функції:

$$f' = -211{,}32Z_1 - 99{,}26Z_2 - 12{,}48Z_3,$$

де $z_{ij} = \dfrac{x_{ij} - \bar{x}_j}{\sigma_{x_j}}$.

Стандартизовані коефіцієнти дискримінантної функції теж показують визначальний вплив першої змінної на величину дискримінантної функції.

Крім визначення внеску кожної вихідної змінної в дискримінантну функцію, можна проаналізувати й ступінь кореляційної залежності між ними.

Для оцінки тісноти зв'язку між окремими змінними й дискримінантними функціями служать коефіцієнти кореляції, які називаються структурними коефіцієнтами. По величині структурних коефіцієнтів судять про зв'язок між змінними й дискримінантними функціями. Структурні коефіцієнти дозволяють також якщо буде потреба привласнити ім'я кожній функції. Вони можуть бути розраховані в цілому по всій сукупності об'єктів (*R*) і для кожного класу окремо $\left(R_{x_j} f_k\right)$.

Різні знаки в структурних коефіцієнтів можна інтерпретувати в такий спосіб. Вихідні змінні, що мають різний напрямок зв'язку з дискримінантною функцією, тобто позитивні або негативні структурні коефіцієнти, будуть орієнтувати об'єкти в різних напрямках, видаляючи або наближаючи їх до центрів відповідних класів. З даного приклада видно, що змінна X_1 й функція f_1 мають коефіцієнт -0,036. Це значить, що при збільшенні значень X_1 функція f_1 зменшується. Допустимо, усі різниці $\left(f_1 - f_l\right) > 0 \left(l = 2,...,k\right)$ для i-го спостереження, значить його варто віднести до першого класу. Якщо в класифікуемих об'єктів значення змінної X_1 будуть зростати, то значення функції f_1 для цих об'єктів будуть зменшуватися, що приведе до віддалення їх від центра першого класу. Зрештою X_1 досягне в *р-го* об'єкта «критичного» значення, якому буде відповідати нерівність $\left(f_1 - f_l\right) < 0$, тобто i-й об'єкт уже не потрапить у перший клас. Аналогічні міркування проводяться й для позитивних структурних коефіцієнтів.

Контрольні запитання

1. Що таке дискримінантний аналіз?
2. На які дві групи можна розбити всі процедури дискримінантного аналізу?

3. Які ознаки називаються дискримінантними змінними?
4. Які допущення приймаються в дискримінантному аналізі?
5. Яким повинне бути співвідношення числа об'єктів спостереження й числа дискримінантних змінних?
6. Як визначається канонічна дискримінантна функція?
7. Яким чином визначають коефіцієнти дискримінантної функції?
8. Як визначається границя, що розділяє розглянуті групи?
9. Яким буде алгоритм використання дискримінантного аналізу для проведення багатомірної класифікації об'єктів?
10. Як зміна числа змінних впливає на результат дискримінантного аналізу?
11. На підставі чого судять про доцільність включення (видалення) дискримінантної змінної?
12. Як розраховуються стандартизовані коефіцієнти дискримінантної функції?
13. У яких випадках застосовують стандартизовані коефіцієнти дискримінантної функції?
14. Як оцінюється тіснота зв'язку між окремими змінними й дискримінантними функціями?

Тема 5. Аналіз даних методами нечіткої кластерізації

1. Постановка задачі нечіткої кластерізації
2. Алгоритм розв’язування задачі нечіткої кластеризації
3. Виконання алгоритму FCM в системі MATLAB
4. Приклад реалізації алгоритму FCM

Після вивчення лекції студент повинен:

Знати:

- задачі, які вирішують методи нечіткої кластерізації;
- методику проектування і побудови систем інтелектуального аналізу даних на основі методів нечіткої кластеризації.

Вміти:

- виконати групування первинних даних;
- визначити критерії якості, цільову функцію, значення якої дозволять зіставити різні схеми класифікації.

1. Постановка задачі нечіткої кластерізації

Концептуальний зв'язок між кластерним аналізом і теорією нечітких множин оснований на тому, що при розв’язуванні задач структуризації складних систем більшість класів, що формуються, «розмиті» за своєю природою. Тому, найбільш адекватну відповідь слід шукати не на питання «Чи належить елемент до того чи іншого класу?», а на питання «В якій степені даний елемент належить класу, що розглядається?».

Методи нечіткої кластеризації вводять до розгляду нечіткі кластери і відповідні їм функції належності, які приймають значення з інтервалу $[0,1]$.

Таким чином, задача нечіткої кластеризації полягає у тому, що необхідно знайти нечітке розбиття або нечітке покриття множини елементів сукупності, що досліджується. Задача зводиться до знаходження степенем належності елементів множини нечітким кластерам (класам).

Нехай початкова (досліджувана) сукупність даних представляє собою скінчену множину елементів $A = \{a_1, a_2, ..., a_n\}$, яке ще називається множиною об’єктів кластеризації. Вводиться також скінчена множина ознак або атрибутів об’єктів $P = \{p_1, p_2, ..., p_q\}$, кожний з яких являє собою деяку характеристику елементів множини A.

Далі, пропонується, що для всіх елементів множини об'єктів кластеризації виміряли всі ознаки множини P, і кожен елемент множини $a_i \in A$, представлений вектором $x_i = \{x_i^1, x_i^2, ..., x_i^q\}$, де $x_i^j \in R$ - дійсне значення ознаки $p_j \in P$ для об'єкту $a_i \in A$.

Взагалі, проблема кількісного вимірювання ознак кожного об'єкта з сукупності – нетривіальна і самостійна задача. Процес вимірювання ознак може бути реалізований в різних шкалах, кожна з яких характеризується допустимим перетворенням даних. В зв'язку з цим, визначають різні типи шкал:

- шкала найменувань: об'єкту ставиться у відповідність деякий символ або номер, який лише відокремлює одне значення ознаки від іншого; прикладом таких ознак є стать людини –(м, ж) або найменування міст (Київ, Житомир, Луганськ,...); допустимим відображенням множини об'єктів у множину символів є бієктивне відображення;
- порядкова шкала: разом з відповідною множиною символьних ознак об'єктів ця шкала дозволяє встановити відношення порядку відносно цієї ознаки; тоді об'єкту ставиться у відповідність деяке число, яке грає роль його оцінки в балах; допустимим відображенням є монотонне зростаюче відображення або функція між двома множинами значень ознак; приклад – оцінки на іспитах;
- інтервальна шкала: крім порядку елементів по ознакам ця шкала встановлює рівність інтервалів значень цієї ознаки; об'єкту, як правило, ставиться у відповідність число, яке дорівнює значенню цієї ознаки; допустимим перетворенням тут є довільна лінійна зростаюча функція між двома множинами значень ознак; характерною ознакою такої шкали є відсутність абсолютного нуля; приклад – температура в шкалах Цельсія;
- шкала відношень: в доповнення до рівності інтервалів додає ще рівність відношень значень ознаки, що розглядається; об'єкту ставиться у відповідність деяке число, яке дорівнює значенню цієї ознаки; допустимим відображенням є довільна лінійна зростаюча функція, яка проходить через нуль; приклад – відстань в метрах, швидкість в км/ч.

Множину ознак слід обирати таким чином, щоб всі $x_i^j \in R$ були виміряні в шкалах відношень чи інтервалів. Саме в такому випадку результати нечіткої кластеризації мають змістовну інтерпретацію, яка адекватна проблемі знаходження нечітких кластерів.

Вектори значень ознак $x_i = \{x_i^1, x_i^2, ..., x_i^q\}$ зручно представляти у вигляді матриці даних *D* розмірності (*n*q*), кожний рядок якої представляє собою значення вектору x_i.

Отже, **задача нечіткого кластерного аналізу формулюється наступним чином**: на основі даних матриці D визначити таке нечітке розбиття $R(A) = \{A_k \mid A_k \subseteq A\}$ або нечітке покриття $J(A) = \{A_k \mid A_k \subseteq A\}$ множини A на задане число нечітких кластерів $A_k (k \in \{2,...,c\})$, яке доставляє екстремум деякій цільовій функції $F(R(A))$ серед всіх нечітких розбиттів чи екстремум цільової функції $F(J(A))$ серед всіх можливих нечітких покриттів.

Для конкретизації задачі ще слід уточнити вигляд цільової функції та тип шуканих нечітких кластерів.

Одним з видів конкретизації цієї задачі є використання спеціальної функції fcm системи MATLAB, який оснований на алгоритмі розв'язування методом нечітких с-середніх.

Для уточнення вигляду цільової функції $F(J(A))$ вводяться деякі додаткові поняття. По-перше, пропонується, що шукані нечіткі кластери представляють собою нечіткі множини $A_k (k \in \{2,...,c\})$, які є нечітким покриттям початкової множини об'єктів кластеризації A, для якої має місце наступна умова:

$$\sum_{k=1}^{c} \mu_{A_k}(a_i) = 1, (\forall a_i \in A), \tag{5.1}$$

де c – загальна кількість нечітких кластерів $A_k (k \in \{2,...,c\})$, яке вважається попередньо заданим.

Далі для кожного кластеру вводяться так звані типові представники або **центри** v_k шуканих нечітких кластерів $A_k (k \in \{2,...,c\})$, які розраховуються за наступною формулою:

$$v_j^k = \frac{\sum_{i=1}^{n} (\mu_{A_k}(a_i))^m x_i^j}{\sum_{i=1}^{n} (\mu_{A_k}(a_i))^m}, (\forall k \in \{2,...,c\}, \forall p_j \in P), \qquad (5.2)$$

де m – деякий параметр, який має назву **експоненційна вага** і дорівнює деякому дійсному числу (m>1). Кожний з центрів кластерів є вектором $v_k = (v_k^1, v_k^2, ..., v_k^q)$ в деякому q-вимірному нормованому просторі ознак, який ізоморфний R^q, якщо всі ознаки виміряні по шкалі відношень.

В якості цільової функції будемо розглядати суму квадратів зважених відхилень координат об'єктів кластеризації від центрів нечітких кластерів:

$$\mathrm{F}(\mathrm{A_k}, \mathrm{v_k^j}) = \sum_{\mathrm{i=1}}^{\mathrm{n}} \sum_{\mathrm{k=1}}^{\mathrm{c}} (\mu_{\mathrm{A_k}}(\mathrm{a_i}))^{\mathrm{m}} \sum_{\mathrm{j=1}}^{\mathrm{q}} (\mathrm{x_i^j} - \mathrm{v_k^j})^2 . \qquad (5.3)$$

Чим більше елементів містить множина А, тим менше значення слід вибирати для m>1.

Тоді задача нечіткої кластеризації полягає у наступному:

для заданої матриці даних *D*, кількості нечітких кластерів $c \in N, c > 1$, параметра *m*, визначити матрицю *U* значень функції належності об'єктів кластерізації $a_i \in A$ нечітким кластерам $A_k (k \in \{2,...,c\})$, які доставляють мінімум цільової функції (5.3) і задовольняють обмеженням (5.1)-(5.2), а також додатковим обмеженням:

$$\sum_{i=1}^{n} \mu_{A_k}(a_i) > 0, (\forall k = \{2,...,c\})$$
$$\mu_{A_k}(a_i) \geq 0 (\forall k = \{2,...,c\}, a_i \in A) \qquad (5.4).$$

Умови (4) виключають появу пустих нечітких кластерів в шуканій нечіткій кластеризації. Таким чином, мінімізація цільової функції (5.3) мінімізує відхилення всіх об'єктів кластеризації від центрів нечітких кластерів пропорційно значенням функцій належності цих об'єктів відповідним нечітким кластерам.

Ця функція не є випуклою, а тому задача кластеризації в загальному випадку відноситься до багатоекстремальних задач нелінійного програмування.

2. Алгоритм розв'язування задачі нечіткої кластеризації

Основні ідеї алгоритму для розв'язування задачі нечіткої кластеризації були запропоновані Дж.К.Данном у 1974р. Цей алгоритм спочатку отримав назву нечіткого алгоритму fuzzyISODATA. У 1980 році Дж.К. Беджек теоретично довів збіжність цього алгоритму, потім він же узагальнив цей алгоритм на випадок довільних нечітких множин даних і запропонував для цього алгоритму назву нечітких середніх FCM, Fuzzy-C-Means. Саме під такою назвою алгоритм реалізований в системі MATLAB.

Алгоритм FCM має ітеративний характер послідовного покращення деякого початкового нечіткого розбиття $R(A) = \{A_k \mid A_k \subseteq A\}$, яке задається користувачем або формується автоматично за деяким евристичним правилом. На кожному кроці ітерації прораховуються значення функцій належності нечітких кластерів і їх типових представників.

Алгоритм FCM завершує роботу у випадку, коли відбудеться наперед задане число ітерацій, або, коли мінімальна абсолютна різниця між значеннями функцій належності на двох послідовних ітераціях не стане менше деякого наперед заданого значення.

Формально алгоритм FCM представляється у вигляді наступних кроків:

1. попередньо необхідно задати наступні значення: кількість шуканих нечітких кластерів c, максимальну кількість ітерацій алгоритмів $s \in N$, параметр збіжності алгоритму $\varepsilon \in R_+$, а також екпоненційну вагу для цільової функції і центрів кластерів m>1. В якості початкового розбиття на першій ітерації алгоритму для матриці даних D задати деяке нечітке розбиття $R(A) = \{A_k \mid A_k \subseteq A\}$ на c непустих нечітких кластерів, які описуються сукупністю функцій належності $\mu_k(a_i), \forall k \in \{2,...,c\}, \forall a_i \in A$.

2. для поточного нечіткого розбиття $R(A) = \{A_k \mid A_k \subseteq A\}$ за формулою (5.2) розрахувати центри нечітких кластерів $v_k^j, (\forall k = \{2,...,c\}, \forall p_j \in P$ і значення цільової функції (5.3). Кількість виконаних ітерацій покласти 1.

3. сформувати нове нечітке розбиття $R'(A) = \{A_k \mid A_k \subseteq A\}$ множини об'єктів кластеризації A на c непусті нечіткі кластери, які характеризуються сукупністю функцій належності $\mu_k^1(a_i), \forall k \in \{2,...,c\}, \forall a_i \in A$, що визначаються за формулою:

$$\mu_k^1(a_i) = \left(\sum_{l=1}^{c} \left(\frac{\left(\sum_{j=1}^{q} (x_i^j - v_k^j)^2 \right)^{\frac{1}{2}}}{\left(\sum_{j=1}^{q} (x_i^j - v_l^j)^2 \right)^{\frac{1}{2}}} \right)^{\frac{2}{m-1}} \right)^{-1}, \forall k \in \{2,...,c\}, \forall a_i \in A$$

4. При цьому, якщо для деякого $k \in \{2,...,c\}$ і деякого $a_i \in A$ значення $\sum_{j=1}^{q} (x_i^j - v_k^j)^2 = 0$, тоді для відповідного нечіткого кластеру A_k беремо $\mu_k^1(a_i) = 1$, а для інших кластерів $A_l (\forall l = \{2,...,c\}, l \neq k)$ беремо $\mu_l^1(a_i) = 0$. Якщо ж таких значень $k \in \{2,...,c\}$ виявиться декілька для $a_i \in A$, тоді евристично беремо $\mu_k^1(a_i) = 1$ для меншого з них, а для інших $\mu_l^1(a_i) = 0$.

5. Для нового нечіткого розбиття $R'(A) = \{A_k \mid A_k \subseteq A\}$ за формулою (5.2) розраховуємо центри нечітких кластерів і значення цільової функції (5.3).

6. Якщо кількість виконаних ітерацій більше за s або модуль різниці між попереднім і новим значенням цільової функції менше за $\varepsilon \in R_+$, тоді в якості результату прийняти нечітке розбиття $R'(A) = \{A_k \mid A_k \subseteq A\}$ і завершити виконання алгоритму. Інакше, вважати поточним розбиттям $R(A) = R'(A)$ і перейти на крок 3, збільшивши на 1 кількість виконаних ітерацій.

В результаті виконання алгоритм зводиться до деякого локально-оптимального розбиття $R^*(A)$, яке описується сукупністю функцій належності $\mu_k(a_i)$, а також центрами нечітких класів $v_k = (v_k^1, v_k^2, ..., v_k^q)$.

3. Виконання алгоритму FCM в системі MATLAB

Функція fcm може бути викликана в одному з наступних форматів:

$[center, U, obj\_fcn] = fcm(data, cluster\_n)$ або

$[center, U, obj\_fcn] = fcm(data, cluster\_n, options)$.

Вхідними аргументами цієї функції є

- data: матриця початкових даних *D*, *i*-тий рядок якої являє собою інформацію про об'єкт нечіткої кластеризації $a_i \in A$ у формі вектора $x_i = \{x_i^1, x_i^2, ..., x_i^q\}$;
- cluster_n: число шуканих кластерів $c \in N, c > 1$.

Вихідними аргументами цієї функції є

- center: матриця центрів шуканих нечітких кластерів, кожний рядок якої являє собою координати центру одного з нечітких кластерів в формі вектора $v_k = (v_k^1, v_k^2, ..., v_k^q)$;
- *U*: матриця значень функцій належності шуканого нечіткого розбиття $\mu_k(a_i), \forall k \in \{2,...,c\}, \forall a_i \in A$;
- obj_fun: значення цільової функції (3) на кожній з ітерацій роботи алгоритму.

Функція fcm() може бути викликана з додатковими аргументами options, які введенні для управління процесом кластеризації, а також для зміни критерію останова роботи алгоритму і/або відображення інформації на екрані монітора.

Ці додаткові аргументи мають наступні значення:

- option (1): експоненційна вага m для розрахунків матриці нечіткого розбиття U (за замовченням $m = 2$);
- option (2): максимальне число ітерацій s (за замовченням це значення дорівнює 100);
- option (3): параметр збіжності алгоритму ε (за замовченням це значення дорівнює 0.00001);
- option (4): інформація про поточну ітерацію, яка відображається на екрані монітора (за замовченням, це значення 1).

Якщо будь-яке зі значень додаткових аргументів дорівнює NaN (не число), тоді для цього аргументу використовується значення за замовченням.

4. Приклад реалізації алгоритму FCM

В якості прикладу застосування нечіткої кластеризації розглянемо множину даних, які містяться в системі MATLAB і використовуються в якості текстової сукупності об'єктів нечіткої кластеризації. Ці дані являють собою матрицю D розмірності 140x2 і містяться у файлі fcmdata.dat, який поставляється разом зі MATLAB В даному випадку матриця D відповідає 140 об'єктам, для кожного з який виконане вимірювання за двома ознаками, що є дуже зручним для візуалізації результатів нечіткої кластеризації в двовимірному просторі на площині.

1. Для візуалізації цих даних слід виконати наступні команди:

Load fcmdata.dat

plot(fcmdata(:,1), fcmdata(:,2), 'o')

На екрані з'явиться графічне зображення, яке представлене на рис. 5.1.

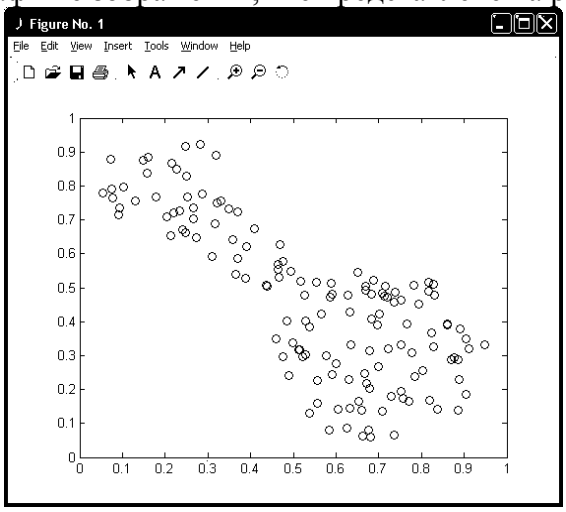

Рис. 5.1. Зображення точок матриці D з файлу fcmdata.dat

2. Далі слід викликати функцію fcm, наприклад, з наступним форматом:

[center, U, obj_fcn]=fcm(fcmdata, 2)

Потім слід подивитись результати виконання процедури нечіткої кластеризації:

- координати центрів класів, тобто, матрицю center;
- належність кожної сукупності даних до класів – матрицю U;
- значення функції цілі – obj_fcn.

Взагалі, процедуру нечіткої кластеризації можна записати у вигляді командного М-файла з наступним текстом:

```
load fcmdata.dat
plot(fcmdata(:,1), fcmdata(:,2),'o')
[center, U, obj_fcn]= fcm(fcmdata, 2);
maxU=max(U);
index1 = find(U(1,:)== maxU);
index2 = find(U(2,:)== maxU);
line(fcmdata(index1,1), fcmdata(index1,2),'linestyle','none',…
'marker', 'x', 'color', 'g');
line(fcmdata(index2,1), fcmdata(index2,2),'linestyle','none',…
'marker', 'x', 'color', 'r');
hold on
plot( center(1,1), center(1,2),'ko', 'markersize',10, 'LineWidth', 2)
plot( center(2,1), center(2,2),'ko', 'markersize',10, 'LineWidth', 2)
```

Результатом цієї програми буде розбиття даних на два кластера, візуалізація якого зображена на рис.5.2.

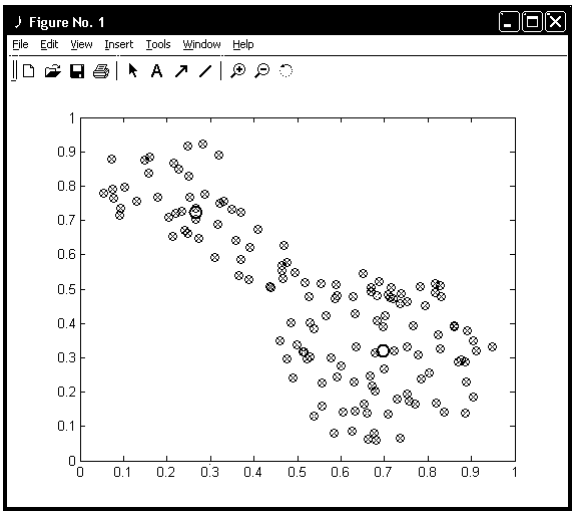

Рис. 5.2. Результат роботи програми нечіткої кластерізації

3. Після роботи програми в системі MATLAB можна перевірити значення матриць center та U, набравши їх назву в командному рядку і натиснувши Enter.

4. Крім того, в програмі можна використати наступний формат запису:

[center, U, obj_fcn]= fcm(fcmdata, 2, [2.5 1000 0.000001 1]);

В цьому випадку експоненційна вага 2.5, максимальне число ітерацій 1000, параметр збіжності $\varepsilon = 0.000001$. Порівняльний аналіз показує практичну ідентичність графіків – результатів використання обох форматів функції fcm, що дозволяє зробити висновок про відповідність отриманих результатів нечіткої кластеризації.

5. Для розв'язування задачі нечіткої кластеризації в системі MATLAB можна використовувати графічний інтерфейс, який викликається за допомогою команди **findcluster**. Ця програма може використовувати або метод с-середніх або метод субтрактивної кластеризації (substractive clustering, який викликається і окремо за допомогою команди **subclust**). Останній використовується тоді, коли не можна заздалегідь встановити число кластерів с на кроці 6.

Формат виклику графічного інтерфейсу: **findcluster або findcluster('file.dat')**.

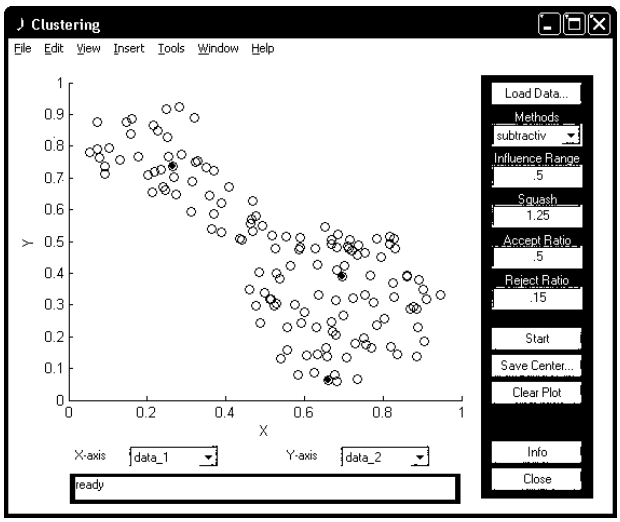

Рис 5.3. Вікно роботи графічного інтерфейсу нечіткої кластеризації для алгоритму субтрактивної кластеризації

В даному вікні можна завантажити файл даних Load Data, обрати метод кластеризації Methods, обрати необхідні значення параметрів і натиснути кнопку Start.

6. Нехай, заздалегідь невідома кількість кластерів c. Тоді слід використати метод субтрактивної кластеризації. Ідея цього методу полягає у тому, що кожна точка даних пропонується в якості центра потенційного кластеру. Далі слід вирахувати деяку міру можливості кожної точки даних представляти центр кластеру. Ця кількісна міра основана на оцінці густини точок навколо відповідного центра кластера.

Цей алгоритм, який є узагальненням методу кластеризації Р. Ягера, заснований на виконанні наступних кроків:

1) вибрати точку даних з максимальним потенціалом для представлення центру першого кластеру

2) забрати всі точки даних в околі центру даного кластеру, величина якої задається параметром **radii**, щоб визначити наступний нечіткий кластер і координати його центру.

Далі, ці дві процедури повторюються до тих пір, допоки всі точки даних не будуть лежати в границях околів радіуса **radii** навколого шуканих кластерів.

Функція командного рядка

[C, S] = subclust (X, radii, xBounds, options)

знаходить центри таких кластерів.

При цьому матриця X містить об’єкти кластеризації, кожний рядок якої відповідає координатам окремої точки даних. Параметр radii являє собою вектор, компоненти якого приймають значення з інтервалу [0,1] і задають діапазон розрахунку центрів кластерів по кожній з розказ вимірювань об’єктів. Робиться припущення, що всі дані знаходяться в деякому гіперкубі. Взагалі, малі значення параметрів radii призводять до знаходження малого числа великих по кількістю точок кластерів. Найкращих результатів можна очікувати при значенні radii між 0.2 і 0.5.

Аргумент xBounds являє собою матрицю розміру (2xq), яка визначає засіб відображення матриці даних X в деякому одиничному гіперкубі. Тут q – кількість ознак. Цей аргумент є необов’язковим, якщо матриця X вже нормована. Перший рядок цієї матриці містить мінімальні значення інтервалу вимірювання кожної ознаки, а другий рядок – максимальне значення вимірювання кожної ознаки.

Для зміни значень, які встановлені по замовченню, можна використати параметр options, компоненти вектора якого можуть приймати наступні значення:

- options(1) = guashFactor – параметр, який використовується в якості коефіцієнту для множення значень radii з ціллю зменшення впливу потенціалу граничних точок, які розглядаються як частина даного кластеру (за замовченням це значення дорівнює 1.25);

- options(2) = acceptRatio – параметр, який встановлює потенціал як частину потенціалу першого кластеру, вище якого інша точка даних не може розглядатся в якості центра іншого кластеру (за замовченням це значення 0.5);
- option(3) = rejectRation - параметр, який встановлює потенціал як частину потенціалу першого кластеру, нижче якого інша точка даних не може розглядатся в якості центра іншого кластеру (за замовченням це значення 0.15);
- options(4) = verbose – якщо значення цього параметра не дорівнює 0, тоді на екран монітору виводиться інформація про виконання процесу кластеризації (за замовченням це значення 0).

Функція subclust повертає матрицю C значень координат центрів нечітких кластерів. При цьому кожний рядок цієї матриці містить координати одного центру кластеру. Вектор S містить σ-значень, які визначають діапазон впливу центра кластеру по кожній з розглянутих ознак. При цьому, всі центра кластерів мають однакову множину σ-значень.

Для **прикладу** розглянемо наступну послідовність команд:

Load fcmdata.dat

[C, S] = subclust(fcmdata, [0.5 0.5], [], [1.25 0.5 0.15 1])

Для кожної ознаки вводяться радіуси околів – 0.5 і 0.5.

Як видно з рис. 3. для даної сукупності даних функція suclust має знайти три кластери.

Для розв’язування задачі субтрактивної кластеризації може використовуватися і розглянутий раніше графічний інтерфейс, який викликається командою **findcluster**.

Таким чином, система MATLAB дозволяє розв’язувати задачі нечіткої кластеризації двома способами: за допомогою функцій командного рядка і за допомогою графічного інтерфейсу кластеризації. Результати нечіткої кластеризації мають наближений характер і мають служити для попередньої структуризація вхідної (початкової) інформації про систему, що вивчається. Тобто, провівши кластерізацію, можна отримати і зберегти знання про систему у вигляді структуризації початкової інформації.

Контрольні запитання

1. Чим обумовлена необхідність розвитку методів кластерного аналізу і їхнього використання?
2. На чому заснований концептуальний зв'язок між кластерним аналізом і теорією нечітких множин?
3. У чому полягає задача нечіткої кластеризації?
4. Ким і коли були запропоновані основні ідеї алгоритму для розв’язування задачі нечіткої кластеризації?
5. Які задачі дозволяють вирішувати методи нечіткої кластеризації?
6. У вигляді яких послідовних кроків можна представити алгоритм FCM?

3. ПЛАНИ ПРАКТИЧНИХ (СЕМІНАРСЬКИХ) ЗАНЯТЬ, САМОСТІЙНОЇ РОБОТИ

Тематика та плани практичних занять

Практичне заняття 1. «Основні поняття методів багатомірної класифікації»

Питання для дискусії:

1. У якій формі можуть представлятися вихідні дані в задачах класифікації об'єктів?
2. Що таке навчальна вибірка?
3. Які кінцеві прикладні цілі ставить перед собою дослідник при проведенні класифікації?
4. Які фактори називаються типоутворюючими?
5. Як будується комбінаційне угруповання?
6. Яка ідея покладена в основу методу відбору найбільш інформативних ознак-детермінантів?

Практичні заняття 2-3. «Кластерний аналіз. Ієрархічні методи класифікації»

Питання для дискусії:

1. Чим обумовлена необхідність розвитку методів кластерного аналізу і їхнього використання?
2. Які задачі дозволяють вирішувати методи кластерного аналізу?
3. У чому суть агломеративних методів кластерного аналізу?
4. У чому суть дивізимних методів кластерного аналізу?
5. Чому вибір способу обчислення відстані між об'єктами є вузловим моментом дослідження?
6. Які відстані між об'єктами найбільш часто використовуються в задачах кластерного аналізу?
7. Які показники можуть бути використані в якості мір подібності?
8. У чому полягає сутність ієрархічних агломеративних методів?

Приклади розв'язання типових задач:

Задача 1

За вихідним даними, представленим у табл. 1.1, потрібно провести класифікацію шести промислових підприємств $(n = 6)$ по двом показникам: $x^{(1)}$ - рентабельність (%), $x^{(2)}$ - продуктивність праці (тис. грн/чіл.)

Таблиця 1.1

Характеристики аналізованих підприємств

№ підприємства	1	2	3	4	5	6
$x^{(1)}$	23,4	17,5	9,7	18,2	6,6	8,0
$x^{(2)}$	9,1	5,2	5,5	9,4	7,5	5,7

Класифікацію проведемо по ієрархічному агломеративному алгоритму з використанням звичайного й зваженого ($w_1 = 0{,}75$ і $w_2 = 0{,}25$) евклидова відстані, а також принципів: «найближчого» і «далекого» сусіда, центру ваги й середньому зв'язку.

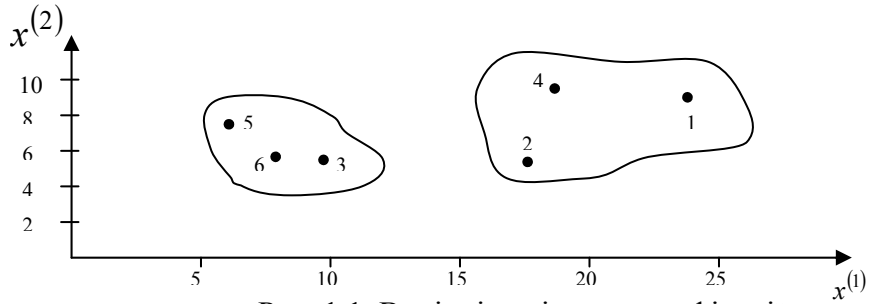

Рис. 1.1. Вихідні дані для класифікації.

На підставі попереднього якісного аналізу можна висунути припущення, що по відомих характеристиках підприємства 1, 2, 4 належать до одній типологічній групі, а підприємства 3, 5, 6 – до іншої, що узгодиться з розташуванням шести спостережень на площині (мал. 1.1).

1) Проведемо класифікацію, вибравши при звичайній евклідовій відстані принцип «найближчого сусіда». Відповідно звичайній евклідовій метриці відстань між спостереженнями 1 і 2 рівно:

$$d_{1,\,2} = \sqrt{(23{,}4 - 17{,}5)^2 + (9{,}1 - 5{,}2)^2} = 7{,}07\,.$$

При цьому очевидно, що $d_{1,\,1} = 0$. Знаходимо відстані між усіма шістьома спостереженнями й будуємо матрицю відстаней:

$$D_1 = \begin{pmatrix} 0 & 7{,}07 & 14{,}6 & 5{,}21 & 16{,}88 & 15{,}77 \\ 7{,}07 & 0 & 7{,}81 & 4{,}26 & 11{,}14 & 9{,}51 \\ 14{,}16 & 7{,}81 & 0 & 9{,}35 & 3{,}69 & 1{,}71 \\ 5{,}21 & 4{,}26 & 9{,}35 & 0 & 11{,}75 & 10{,}85 \\ 16{,}88 & 11{,}14 & 3{,}69 & 11{,}75 & 0 & 2{,}28 \\ 15{,}77 & 9{,}51 & 1{,}71 & 10{,}85 & 2{,}28 & 0 \end{pmatrix}.$$

З матриці відстаней D_1 випливає, що об'єкти 3 і 6 найбільш близькі $\left(d_{3,\,6} = d_{6,\,3} = 1{,}71\right)$, тому об'єднаємо їх в один кластер. Після об'єднання об'єктів одержимо п'ять кластерів: S_1, S_2, $S_{(3,6)}$, S_4, S_5.

Відстань між кластерами будемо знаходити за принципом «найближчого сусіда», скориставшись формулою перерахування. Так, відстань між кластером S_1 і кластером $S_{(3,6)}$ рівно:

$$d_{1,\,(3,\,6)} = d\left(S_1,\, S_{(3,\,6)}\right) = \frac{1}{2} d_{1,\,3} + \frac{1}{2} d_{1,\,6} - \frac{1}{2}\left|d_{1,\,3} - d_{1,\,6}\right| =$$

$$= \frac{1}{2}(14{,}16 + 15{,}77) - \frac{1}{2}\left|14{,}16 - 15{,}77\right| = 14{,}16.$$

Як видне з розрахунку, відстань $d_{1,\,(3,\,6)}$ дорівнює відстані від об'єкта 1 до найближчого до нього об'єкта, що входить у кластер $S_{(3,\,6)}$, тобто $d_{1,\,(3,\,6)} = d_{1,\,3} = 14{,}16$. Провівши аналогічні розрахунки, одержимо матрицю відстаней:

$$D_2 = \begin{pmatrix} 0 & 7{,}07 & 14{,}16 & 5{,}21 & 16{,}88 \\ 7{,}07 & 0 & 7{,}81 & 4{,}26 & 11{,}14 \\ 14{,}16 & 7{,}81 & 0 & 9{,}35 & 3{,}69 \\ 5{,}21 & 4{,}26 & 9{,}35 & 0 & 11{,}75 \\ 16{,}88 & 11{,}14 & 3{,}69 & 11{,}75 & 0 \end{pmatrix}.$$

З матриці відстаней D_2 випливає, що найбільш близькі кластери $S_{(3,\,6)}$ й S_5 $\left(d_{5,(3,\,6)} = 3{,}69\right)$. Після їхнього об'єднання маємо чотири кластери: S_1, S_2, $S_{(3,\,5,\,6)}$, S_4.

Будуємо нову матрицю спостережень:

$$D_3 = \begin{pmatrix} 0 & 7{,}07 & 14{,}16 & 5{,}21 \\ 7{,}07 & 0 & 7{,}81 & 4{,}26 \\ 14{,}16 & 7{,}81 & 0 & 9{,}35 \\ 5{,}21 & 4{,}26 & 9{,}35 & 0 \end{pmatrix}.$$

Об'єднаємо спостереження 2 і 4, що мають найменшу відстань ($d_{2,\,4} = 4{,}26$). Після об'єднання маємо три кластери: S_1, $S_{(2,\,4)}$ і $S_{(3,\,5,\,6)}$. Обчислимо нову матрицю спостережень:

$$D_4 = \begin{pmatrix} 0 & 5{,}21 & 14{,}16 \\ 5{,}21 & 0 & 7{,}81 \\ 14{,}16 & 7{,}81 & 0 \end{pmatrix}.$$

Об'єднаємо кластери S_1 й $S_{(2,\,4)}$, відстань між якими, згідно з матрицею D_4, мінімально ($d_{1,\,(2,\,4)} = 5{,}21$). У результаті цього одержимо два кластери: $S_{(1,\,2,\,4)}$ і $S_{(3,\,5,\,6)}$. Матриця відстаней буде мати вигляд:

$$D_5 = \begin{pmatrix} 0 & 7{,}81 \\ 7{,}81 & 0 \end{pmatrix}.$$

З матриці D_5 випливає, що на відстані $d_{(1,\,2,4),\,(3,\,5,\,6)} = 7{,}81$ всі шість спостережень поєднуються в один кластер.

На підставі графічної вистави результатів кластерного аналізу можна зробити висновок, що найкращим є розбивка шести підприємств на два кластери: $S_{(1,\,2,\,4)}$ і $S_{(3,\,5,\,6)}$, коли гранична відстань перебуває в інтервалі $5{,}21 \le d_{пор} \le 7{,}81$.

Задача 2

Проведемо класифікацію, вибравши при звичайній евклідовій відстані принцип «далекого сусіда».

Як і в першому випадку, будемо використовувати звичайне євклідово відстань, тому матриця D_1 залишиться без зміни. Згідно агломеративному алгоритму в один кластер поєднуються об'єкти 3 і 6, як найбільше близькі $\left(d_{3,\,6} = 1{,}71\right)$. Після об'єднання маємо п'ять кластерів: S_1, S_2, $S_{(3,\,6)}$, S_4 і S_5.

У вигляді того, що відстань між кластерами вимірюємо за принципом «далекого сусіда», у формулі перерахування ухвалюємо $\delta = 1/2$, а не $-1/2$, як у першому випадку. Тоді, наприклад, відстань між кластером S_1 і кластером $S_{(3,\,6)}$ визначається по формулі:

$$d_{1,\,(3,\,6)} = d\left(S_1,\, S_{(3,\,6)}\right) = \frac{1`}{2} d_{1,\,3} + \frac{1}{2} d_{1,\,6} + \frac{1}{2}\left|d_{1,\,3} - d_{1,\,6}\right| =$$

$$= \frac{1}{2}(14{,}16 + 15{,}77) + \frac{1}{2}|14{,}16 - 15{,}77| = 15{,}77.$$

Таким чином, відстань $d_{1,\,(3,\,6)}$ дорівнює відстані від об'єкта 1 до найбільш віддаленого від нього об'єкта, що входить у кластер $S_{(3,\,6)}$, тобто $d_{1,\,(3,\,6)} = d_{1,\,6} = 15{,}77$.

Аналогічно розглядаються всі інші елементи матриці відстаней:

$$D_2 = \begin{pmatrix} 0 & 7{,}07 & 15{,}77 & 5{,}21 & 16{,}88 \\ 7{,}07 & 0 & 9{,}51 & 4{,}26 & 11{,}14 \\ 15{,}77 & 9{,}51 & 0 & 10{,}85 & 3{,}69 \\ 5{,}21 & 4{,}26 & 10{,}85 & 0 & 11{,}75 \\ 16{,}88 & 11{,}14 & 3{,}69 & 11{,}75 & 0 \end{pmatrix}.$$

Згідно з матрицею D_2 найбільш близькими кластерами будуть $S_{(3,6)}$ і S_5, $d_{(3,\,6),\,5} = 3{,}69$. Після їхнього об'єднання маємо чотири кластери: S_1, S_2, $S_{(3,\,5,\,6)}$ і S_4. Будуємо матрицю відстаней D_3, скориставшись принципом далекого сусіда:

$$D_3 = \begin{pmatrix} 0 & 7{,}07 & 16{,}88 & 5{,}21 \\ 7{,}07 & 0 & 11{,}14 & 4{,}26 \\ 16{,}88 & 11{,}14 & 0 & 11{,}75 \\ 5{,}21 & 4{,}26 & 11{,}75 & 0 \end{pmatrix}.$$

Об'єднаємо об'єкти 2 і 4 в один кластер, як найбільше близькі (згідно з матрицею D_3), $d_{2,\,4} = 4{,}26$. Після об'єднання маємо три кластери: S_1, $S_{(2,\,4)}$ і $S_{(3,\,5,\,6)}$. Будуємо нову матрицю D_4 за принципом «далекого сусіда»:

$$D_4 = \begin{pmatrix} 0 & 7{,}07 & 16{,}88 \\ 7{,}07 & 0 & 11{,}75 \\ 16{,}88 & 11{,}75 & 0 \end{pmatrix}.$$

Об'єднаємо кластери S_1 й $S_{(2,\,4)}$, відстань між якими $d_{1,\,(2,\,4)} = 7{,}07$ мінімально, і одержимо два кластери: $S_{(1,\,2,\,4)}$ і $S_{(3,\,5,\,6)}$, відстань між якими визначається по матриці:

$$D_5 = \begin{pmatrix} 0 & 16{,}88 \\ 16{,}88 & 0 \end{pmatrix}$$

і рівно $d_{(1,\,2,\,4),\,(3,\,5,\,6)} = 16{,}88$.

Графічні результати класифікації представлені на мал. 1.2.

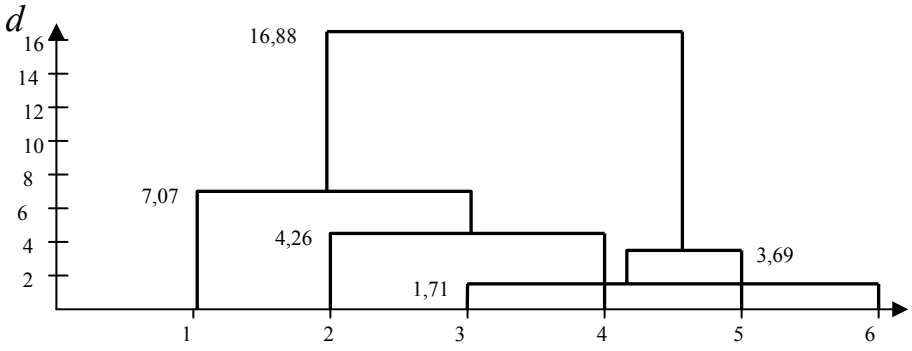

Рис. 1.2. Дендрограма (звичайне євклідово відстань, далекий сусід).

Як і в попередньому випадку найкращим є розбивка підприємств на два кластери (мал. 1.2): $S_{(1,\,2,\,4)}$ і $S_{(3,\,5,\,6)}$, після передостаннього кроку класифікації, коли інтервал виміру відстані об'єднання найбільша $7{,}07 \le d_{пор} \le 16{,}88$.

Практичні заняття 4-5. «Кластерний аналіз. Ітеративні методи класифікації»

Питання для дискусії:

1. Яким чином проводиться класифікація об'єктів методом k-середніх?
2. До виконання яких кроків зводяться обчислювальні процедури більшості ітеративних методів класифікації?
3. У чому складається істотна відмінність методу пошуку згущень від інших ітеративних методів класифікації?
4. У чому полягає суть ітеративного алгоритму типу «форель»?
5. Якими способами вибирається радіус сфери для пошуку локальних згущень точок у методі пошуку згущень?
6. Для якої мети використовуються критерії якості класифікації?

Приклади розв'язання типових задач:

За вихідним даними, представленим у табл. 2.1, потрібно провести класифікацію шести промислових підприємств $(n = 6)$ по двом показникам: $x^{(1)}$ - рентабельність (%), $x^{(2)}$ - продуктивність праці (тис. грн/чіл.)

Таблиця 2.1

Характеристики аналізованих підприємств

№ підприємства	1	2	3	4	5	6
$x^{(1)}$	23,4	17,5	9,7	18,2	6,6	8,0
$x^{(2)}$	9,1	5,2	5,5	9,4	7,5	5,7

На попередньому занятті класифікація аналізованих підприємств була проведена за допомогою ієрархічних методів класифікації. Тепер проведемо класифікацію підприємств за допомогою методу k-середніх і зрівняємо отримані результати.

Для початку з n-об'єктів досліджуваної сукупності відбираються випадковим образом або задаються дослідником виходячи з якихось апріорних міркувань k-об'єктів еталонів. Оскільки згідно результатів розв'язку цього завдання на практичному занятті №1 одержали

розбивку підприємств на два кластери, то виберемо два об'єкти - еталона з найбільш високими й низькими показниками, наприклад, підприємства 1 і 5.

Запишемо вихідні значення еталонів і ваг:

$$\left.\begin{array}{l} E_1^0 = X_1 = (23{,}4;\ 9{,}1);\ w_1^0 = 1 \\ E_2^0 = X_2 = (6{,}6;\ 7{,}5);\ w_2^0 = 1 \end{array}\right\} \text{ - нулевая итерация}.$$

На *першому кроці* беремо другий об'єкт і визначаємо його відстань до кожного з еталонів по евклідовій метриці:

$$d_{21} = \sqrt{(17{,}5 - 23{,}4)^2 + (5{,}2 - 9{,}1)^2} = 7{,}072;$$

$$d_{22} = \sqrt{(17{,}5 - 6{,}6)^2 + (5{,}2 - 7{,}5)^2} = 11{,}14.$$

Отже, розглянутий об'єкт повинен бути приєднаний до першого еталона й перший еталон буде перелічений, а другий не міняється:

$$E_1^1 = \frac{w_1^0 \cdot E_1^0 + X_2}{w_1^0 + 1},$$

$$w_1^1 = w_1^0 + 1 = 2,\ E_2^1 = E_2^0,\ w_2^1 = w_2^0,$$

де X_2 - вектор значень змінних для другого об'єкта,

E_1^1 - перелічене значення еталона;

$$E_1^1 = \left(\frac{23{,}4 + 17{,}5}{2}; \frac{9{,}1 + 5{,}2}{2}\right) = (20{,}45;\ 7{,}15).$$

На *другому кроці* перевіряємо, до якого еталона ближче всього перебуває третій об'єкт:

$$d_{31} = \sqrt{(9{,}7 - 20{,}45)^2 + (5{,}5 - 7{,}15)^2} = 10{,}876;$$

$$d_{32} = \sqrt{(9{,}7 - 6{,}6)^2 + (5{,}5 - 7{,}5)^2} = 3{,}689.$$

Тому що $d_{32} = \min\{d_{31}; d_{32}\}$, отже, третій об'єкт приєднується до другого еталона, цей еталон перераховується й вага його збільшується:

$$E_2^2 = \left(\frac{9{,}7 + 6{,}6}{2}; \frac{5{,}5 + 7{,}5}{2}\right) = (8{,}15;\ 6{,}5),$$

$$w_2^2 = w_2^1 + 1 = 2; \quad E_1^2 = E_1^1; \quad E_2^2 = E_2^1; \quad w_1^2 = w_1^1.$$

На *третьому кроці* перевіряємо, до якого еталона ближче всього перебуває четвертий об'єкт:

$$d_{41} = \sqrt{(18{,}2 - 20{,}45)^2 + (9{,}4 - 7{,}15)^2} = 3{,}182;$$

$$d_{42} = \sqrt{(18{,}2 - 8{,}15)^2 + (9{,}4 - 6{,}5)^2} = 10{,}46.$$

Отже, четвертий об'єкт приєднується до першого еталона, вага якого стає рівним:

$$E_1^3 = \left(\frac{20{,}45 \cdot 2 + 18{,}2}{3}; \frac{7{,}15 \cdot 2 + 9{,}4}{3}\right) = (19{,}7;\ 7{,}9);$$

$$w_1^3 = w_1^2 + 1 = 3; \quad E_2^3 = E_2^2; \quad w_2^3 = w_2^2.$$

На *четвертому кроці* перевіряємо, до якого еталона ближче всього перебуває шостий об'єкт:

$$d_{61} = \sqrt{(8{,}0 - 19{,}7)^2 + (5{,}7 - 7{,}9)^2} = 11{,}905;$$

$$d_{62} = \sqrt{(8{,}0 - 8{,}15)^2 + (5{,}7 - 6{,}5)^2} = 0{,}814.$$

Шостий об'єкт приєднується до другого еталона, вага якого стає рівним:

$$E_2^4 = \left(\frac{8{,}15 \cdot 2 + 8}{3}; \frac{6{,}5 \cdot 2 + 5{,}7}{3} \right) = (8{,}1;\ 6{,}233);$$

$$w_1^4 = w_1^3; \quad w_2^4 = w_2^3 + 1 = 3; \quad E_1^4 = E_1^3.$$

Таким чином, отримана розбивка аналізованих підприємств на два кластери: (1, 2, 4) і (3, 5, 6).

Після того як переглянуті всі об'єкти, крім першого й п'ятого, процес «зациклюється», тобто по тому ж правилу здійснюються перегляд і приєднання до відповідного до еталона кожного із шести об'єктів. При цьому відбувається перерахування еталонів і триває нарощування їх ваг. Результати розрахунків представлені в табл. 2.

Таблиця 2

Параметричні дані кластеризації об'єктів методом k-середніх

Номер ітерації	*Еталони і їх ваги*	
	1	*2*
0	$E_1^0 = (23{,}4;\ 9{,}1);\ w_1^0 = 1$	$E_2^0 = (6{,}6;\ 7{,}5);\ w_2^0 = 1$
1	$E_1^1 = (20{,}45;\ 7{,}15);\ w_1^1 = 2$	$E_2^1 = (6{,}6;\ 7{,}5);\ w_2^1 = 1$
2	$E_1^2 = (20{,}45;\ 7{,}15);\ w_1^2 = 2$	$E_2^2 = (8{,}15;\ 6{,}5);\ w_2^2 = 2$
3	$E_1^3 = (19{,}7;\ 7{,}9);\ w_1^3 = 3$	$E_2^3 = (8{,}15;\ 6{,}5);\ w_2^3 = 2$
4	$E_1^4 = (19{,}7;\ 7{,}9);\ w_1^4 = 3$	$E_2^4 = (8{,}1;\ 6{,}233);\ w_2^4 = 3$
5	$E_1^5 = (20{,}625;\ 8{,}2);\ w_1^5 = 4$	$E_2^5 = (8{,}1;\ 6{,}233);\ w_2^5 = 3$
6	$E_1^6 = (20{,}0;\ 7{,}6);\ w_1^6 = 5$	$E_2^6 = (8{,}1;\ 6{,}233);\ w_2^6 = 3$
7	$E_1^7 = (20{,}0;\ 7{,}6);\ w_1^7 = 5$	$E_2^7 = (8{,}5;\ 6{,}05);\ w_2^7 = 4$
8	$E_1^8 = (19{,}7;\ 7{,}9);\ w_1^8 = 6$	$E_2^8 = (8{,}5;\ 6{,}05);\ w_2^8 = 4$
9	$E_1^9 = (19{,}7;\ 7{,}9);\ w_1^9 = 6$	$E_2^9 = (8{,}12;\ 6{,}34);\ w_2^9 = 5$
10	$E_1^{10} = (19{,}7;\ 7{,}9);\ w_1^{10} = 6$	$E_2^{10} = (8{,}1;\ 6{,}23);\ w_2^{10} = 6$

Отже, на цьому процес завершується, тому що наступна розбивка (ітерації 5 - 10) дали такий же результат розділення, як і попереднє (ітерації 1 - 4).

Утворено два кластери: $S_1\{1,\ 2,\ 4\}$, $S_2\{3,\ 5,\ 6\}$. Обчислюємо центри ваги отриманих кластерів (у загальному випадку ці центри можуть не збігатися із центрами еталонів):

$$C_1 = \left(\frac{23{,}4 + 17{,}5 + 18{,}2}{3}; \frac{9{,}1 + 5{,}2 + 9{,}4}{3} \right) = (19{,}7;\ 7{,}9);$$

$$C_2 = \left(\frac{9{,}7 + 6{,}6 + 8{,}0}{3}; \frac{5{,}5 + 7{,}5 + 5{,}7}{3} \right) = (8{,}1;\ 6{,}23).$$

Після цього будується остаточна розбивка: кожна багатомірна крапка ставиться до того кластера, центр якого ближче всіх до цієї крапки.

Для нашого прикладу визначаємо по черзі відстані всіх об'єктів $(X_1, X_2, X_3, X_4, X_5, X_6)$ до центрів двох кластерів (табл. 3).

Таблиця 3

Відстані до центрів класів

Центри кластерів	Об'єкти					
	1	2	3	4	5	6
C_1	3,89	3,48	10,28	2,12	13,11	11,91
$З_2$	15,57	9,46	1,76	10,59	1,97	0,54

Як видне з табл. 3, підтверджується отримана розбивка на два кластери: $S_1\{1,\ 2,\ 4\}$, $S_2\{3,\ 5,\ 6\}$. На цьому алгоритм завершується.

Таким чином, класифікація аналізованих підприємств із використанням методу k-середніх дала такий же результат, як і класифікація ієрархічними методами кластерного аналізу.

Практичні заняття 6-7. «Дискримінантний аналіз»

Питання для дискусії:

1. Які ознаки називаються дискримінантними змінними?
2. Які допущення приймаються в дискримінантному аналізі?
3. Яким повинне бути співвідношення числа об'єктів спостереження й числа дискримінантних змінних?
4. Як визначається канонічна дискримінантна функція?
5. Яким чином визначають коефіцієнти дискримінантної функції?
6. Як визначається границя, що розділяє розглянуті групи?
7. Яким буде алгоритм використання дискримінантного аналізу для проведення багатомірної класифікації об'єктів?
8. Як зміна числа змінних впливає на результат дискримінантного аналізу?

Типові задачі:

Аналізується діяльність 8 регіонів України по трьом економічним показникам, що характеризують сільськогосподарське виробництво в цих регіонах в 2000 г (табл. 3.1.): урожайність зернових культур (у господарствах усіх категорій; з 1 га; ц), середній річний удій молока від однієї корови (у сільськогосподарських підприємствах; кг.) і виробництво м'яса на душу населення (у господарствах усіх категорій; у забійній вазі; кг/чіл).

Як видно з табл. 3.1, з регіонів виділено 2 групи: передова (X) і відстаюча (Y). Для регіонів, які підлягають дискримінації (Z), потрібно обчислити значення дискримінантної функції й провести їхню класифікацію.

Метою дискримінантного аналізу є віднесення нового спостереження (рядка матриці Z) або до X, або к. Y.

Таблиця 3.1

Вихідні дані

Регіони / Показники		Урожайність зернових культур	Середній річний удій молока від однієї корови	Виробництво м'яса на душу населення
Передовий, X	Вінницька обл.	23,6	3015	52
	Київська обл.	24,7	2830	64
	Черкаська обл.	27,2	2920	52
Відстаючий, Y	Запорізька обл.	15,5	2432	37
	Луганська обл.	11,8	2305	20
	Миколаївська обл.	14,7	2487	38
Z	Волинська обл.	19,4	2558	41
	Чернівецька обл.	24,2	3188	31

Запишемо вихідні дані у вигляді матриць X і Y:

$$X = \begin{pmatrix} 23,6 & 3015 & 52 \\ 24,7 & 2830 & 64 \\ 27,2 & 2920 & 52 \end{pmatrix}; \quad Y = \begin{pmatrix} 15,5 & 2432 & 37 \\ 11,8 & 2305 & 20 \\ 14,7 & 2487 & 38 \end{pmatrix}.$$

При розв'язку завдань по навчальних вибірках спочатку визначають їхні вектори середніх:

$$\overline{X} = \begin{pmatrix} \overline{x}_1 \\ \overline{x}_2 \\ \overline{x}_3 \end{pmatrix} = \begin{pmatrix} 25,2 \\ 2922 \\ 56 \end{pmatrix}; \quad \overline{Y} = \begin{pmatrix} \overline{y}_1 \\ \overline{y}_2 \\ \overline{y}_3 \end{pmatrix} = \begin{pmatrix} 14 \\ 2408 \\ 31,7 \end{pmatrix}. \tag{3.1}$$

Використовуючи дані (3.1) можна знайти коваріаційні матриці:

$$S_x = \left(s_{ki}\right)_x \text{ и. } S_y = \left(s_{ki}\right)_y$$

Елементи матриці S_x визначаються вираженням:

$$s_{ki}(x) = \frac{1}{n_1}\sum_{i=1}^{n_1}\left(x_{ij} - \overline{x}_j\right)\left(x_{ik} - \overline{x}_k\right) = \overline{x_j x_k} - \overline{x}_j \overline{x}_k, \tag{3.2}$$

де $j,\ k = 1,\ 2,\ 3$;

$$\overline{x}_j = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{ij}.$$

Елементи матриці S_y визначаються по аналогічному вираженню. Визначимо елементи коваріаційних матриць S_x і S_y відповідно для передових і відстаючих регіонів згідно формули (3.2):

$$s_{11}(x) = \frac{23,6 \cdot 23,6 + 24,7 \cdot 24,7 + 27,2 \cdot 27,2}{3} - 25,2 \cdot 25,2 = 0,59;$$

$$s_{12}(x)=\frac{23{,}6\cdot 3015+24{,}7\cdot 2830+27{,}2\cdot 2920}{3}-25{,}2\cdot 2922=141{,}4;$$

$$s_{13}(x)=\frac{23{,}6\cdot 52+24{,}7\cdot 64+27{,}2\cdot 52}{3}-25{,}2\cdot 56=3{,}73;$$

$$s_{22}(x)=\frac{3015\cdot 3015+2830\cdot 2830+2920\cdot 2920}{3}-2922\cdot 2922=3758;$$

$$s_{23}(x)=\frac{3015\cdot 52+2830\cdot 64+2920\cdot 52}{3}-2922\cdot 56=-385{,}3;$$

$$s_{33}(x)=\frac{52\cdot 52+64\cdot 64+52\cdot 52}{3}-56\cdot 56=32;$$

$$s_{11}(y)=\frac{15{,}5\cdot 15{,}5+11{,}8\cdot 11{,}8+14{,}7\cdot 14{,}7}{3}-14\cdot 14=2{,}53;$$

$$s_{12}(y)=\frac{15{,}5\cdot 2432+11{,}8\cdot 2305+14{,}7\cdot 2487}{3}-14\cdot 2408=106;$$

$$s_{13}(y)=\frac{15{,}5\cdot 37+11{,}8\cdot 20+14{,}7\cdot 38}{3}-14\cdot 31{,}7=12{,}2;$$

$$s_{22}(y)=\frac{2432\cdot 2432+2305\cdot 2305+2487\cdot 2487}{3}-2408\cdot 2408=5808;$$

$$s_{23}(y)=\frac{2432\cdot 37+2305\cdot 20+2487\cdot 38}{3}-2408\cdot 31{,}7=529{,}7;$$

$$s_{33}(y)=\frac{37\cdot 37+20\cdot 20+38\cdot 38}{3}-31{,}7\cdot 31{,}7=66{,}1.$$

У результаті розрахунків отримані наступні коваріаційні матриці:

$$S_x=\begin{pmatrix}0{,}59 & 141{,}4 & 3{,}73\\ & 3758 & -385{,}3\\ & & 32\end{pmatrix};\quad S_y=\begin{pmatrix}2{,}53 & 106 & 12{,}2\\ & 5808 & 529{,}7\\ & & 66{,}1\end{pmatrix}.$$

Далі знайдемо сумарну коваріаційну матрицю:

$$S_*=\frac{1}{n_1+n_2-2}\left(n_1S_x+n_2S_y\right)=\frac{1}{3+3-2}\left(3S_x+3S_y\right)=\begin{pmatrix}2{,}34 & 185{,}5 & 11{,}9\\ & 7174 & 108{,}3\\ & & 73{,}6\end{pmatrix},$$

а потім зворотну їй матрицю

$$S_*^{-1}=\begin{pmatrix}-0{,}27723 & 0{,}00664 & 0{,}03505\\ & -0{,}00002 & -0{,}00105\\ & & 0{,}00946\end{pmatrix}.$$

Обчислимо вектор коефіцієнтів дискримінантної функції:

$$A=S_*^{-1}\left(\overline{X}-\overline{Y}\right)=S_*^{-1}\begin{pmatrix}11{,}2\\ 514\\ 24{,}3\end{pmatrix}=\begin{pmatrix}1{,}16\\ 0{,}039\\ 0{,}083\end{pmatrix}.$$

Після цього розрахуємо вектори дискримінантної функції для матриць вихідних даних:

$$\hat{U}_x = X \cdot A = \begin{pmatrix} 23,6 & 3015 & 52 \\ 24,7 & 2830 & 64 \\ 27,2 & 2920 & 52 \end{pmatrix} \cdot \begin{pmatrix} 1,16 \\ 0,039 \\ 0,083 \end{pmatrix} = \begin{pmatrix} 149,3 \\ 144,3 \\ 149,7 \end{pmatrix};$$

$$\hat{U}_y = Y \cdot A = \begin{pmatrix} 15,5 & 2432 & 37 \\ 11,8 & 2305 & 20 \\ 14,7 & 2487 & 38 \end{pmatrix} \cdot \begin{pmatrix} 1,16 \\ 0,039 \\ 0,083 \end{pmatrix} = \begin{pmatrix} 115,9 \\ 105,2 \\ 117,2 \end{pmatrix}.$$

Визначимо середні значення векторів дискримінантної функції:

$$\overline{\hat{u}_x} = \frac{1}{n_1} \sum_{i=1}^{n_1} \hat{U}_{x_i} = 147,8;$$

$$\overline{\hat{u}_y} = \frac{1}{n_2} \sum_{i=1}^{n_2} \hat{U}_{y_i} = 112,8,$$

а потім константу дискримінації:

$$\hat{C} = \frac{1}{2}(147,8 + 112,8) = 130,3.$$

Щоб визначити, до якої групи ставляться регіони, що підлягають дискримінації, розрахуємо для них дискримінантні функції

$$\hat{U}_1(Z) = A_1 Z_1 + A_2 Z_2 + A_3 Z_3 = 1,16 \cdot 19,4 + 0,039 \cdot 2558 + 0,083 \cdot 41 = 125,7;$$

$$\hat{U}_2(Z) = 1,16 \cdot 24,2 + 0,039 \cdot 3188 + 0,083 \cdot 31 = 155.$$

Зрівняємо отримані дискримінантні функції з константою дискримінації:

1) $\hat{U}_1(Z) = 125,7 < \hat{C} = 130,3$, тому Волинську область не можна віднести до передових регіонів;

2) $\hat{U}_2(Z) = 155 > \hat{C} = 130,3$, тому Чернівецьку область слід віднести до передових регіонів.

КРИТЕРІЇ ОЦІНЮВАННЯ РЕЗУЛЬТАТИВНОСТІ РОБОТИ СТУДЕНТІВ НА ПРАКТИЧНОМУ ЗАНЯТТІ

- міра повноти застосування практичних прийомів і методів аналізу теоретичних положень і концепцій учбової дисципліни;
- міра повноти застосування придбаних студентами умінь і навичок використання сучасних теоретичних методів в рішенні конкретних практичних завдань;
- міра повноти застосування творчого професійного мислення;
- міра повноти використання професійних знань в учбових умовах - оволодіння термінологією відповідної дисципліни;
- міра повноти оволодіння уміннями і навичками постановки і рішення проблем і завдань.

ПЛАНИ САМОСТІЙНОЇ РОБОТИ

Завдання для самостійної роботи

Самостійна робота студента включає: опрацювання навчального матеріалу, підготовку до практичних занять, виконання індивідуальних завдань, підготовку до усіх видів контролю.

№ з/п	Види, зміст самостійної роботи (назва теми)	Кількість годин	
		денне	заочне
1	Проробка конспекту лекції, вивчення рекомендованої літератури по темах, які викладаються на лекціях	15	25
2	Підготовка до практичних занять	25	35
3	Підготовка до поточного контролю	20	25
4	Підготовка до підсумкового контролю (екзамену)	38	41
5	Виконання індивідуального завдання (ІНДЗ) з курсу	20	20
	Разом	**118**	**146**

Питання для самостійного опрацювання

Тема 1. Основні поняття методів багатомірної класифікації

1. Що розуміється під класифікацією об'єктів?
2. У чому полягає проблема класифікації об'єктів за багатомірними даними?
3. У якій формі можуть представлятися вихідні дані в задачах класифікації об'єктів?
4. Що таке навчальна вибірка?
5. Які кінцеві прикладні цілі ставить перед собою дослідник при проведенні класифікації?
6. Які фактори називаються типоутворюючими?
7. Як будується комбінаційне угруповання?
8. Яка ідея покладена в основу методу відбору найбільш інформативних ознак-детермінантів?
9. Дати характеристику класифікації як необхідного попереднього етапу статистичної обробки багатомірних даних.
10. Як використовується класифікація в задачах планування вибіркових обстежень?
11. Як класифікуються задачі розбивки об'єктів на однорідні групи залежно від наявності апріорної й попередньої вибіркової інформації?

Тема 2. Кластерний аналіз. Ієрархічні методи класифікації

1. Що таке кластерний аналіз?
2. Яка мета кластерного аналізу?
3. Чим обумовлена необхідність розвитку методів кластерного аналізу і їхнього використання?
4. Які задачі дозволяють вирішувати методи кластерного аналізу?
5. У чому суть агломеративних методів кластерного аналізу?
6. У чому суть дивізимних методів кластерного аналізу?
7. Чому вибір способу обчислення відстані між об'єктами є вузловим моментом дослідження?
8. Які відстані між об'єктами найбільш часто використовуються в задачах кластерного аналізу?
9. Які показники можуть бути використані в якості мір подібності?

10. У чому полягає сутність ієрархічних агломеративних методів?
11. Як будується дендограма?
12. Які відстані між групами об'єктів є найбільш уживаними?
13. Чому якість проведення кластеризації залежить від алгоритму об'єднання в ієрархічних агломеративних методах?

Тема 3. Кластерний аналіз. Ітеративні методи класифікації

1. У чому полягає суть ітеративних методів кластерного аналізу?
2. Яким чином в ітеративних методах задаються початкові умови?
3. Яким чином проводиться класифікація об'єктів методом k -середніх?
4. До виконання яких кроків зводяться обчислювальні процедури більшості ітеративних методів класифікації?
5. У чому складається істотна відмінність методу пошуку згущень від інших ітеративних методів класифікації?
6. У чому полягає суть ітеративного алгоритму типу «форель»?
7. Якими способами вибирається радіус сфери для пошуку локальних згущень точок у методі пошуку згущень?
8. Для якої мети використовуються критерії якості класифікації?
9. Які найпоширеніші функціонали якості використовуються в кластерному аналізі?
10. Які найпростіші прийоми дозволяють судити про якість розбивки об'єктів на кластери?

Тема 4. Дискримінантний аналіз

1. Що таке дискримінантний аналіз?
2. На які дві групи можна розбити всі процедури дискримінантного аналізу?
3. Які ознаки називаються дискримінантними змінними?
4. Які допущення приймаються в дискримінантному аналізі?
5. Яким повинне бути співвідношення числа об'єктів спостереження й числа дискримінантних змінних?
6. Як визначається канонічна дискримінантна функція?
7. Яким чином визначають коефіцієнти дискримінантної функції?
8. Як визначається границя, що розділяє розглянуті групи?
9. Яким буде алгоритм використання дискримінантного аналізу для проведення багатомірної класифікації об'єктів?
10. Як зміна числа змінних впливає на результат дискримінантного аналізу?
11. На підставі чого судять про доцільність включення (видалення) дискримінантної змінної?
12. Як розраховуються стандартизовані коефіцієнти дискримінантної функції?
13. У яких випадках застосовують стандартизовані коефіцієнти дискримінантної функції?

Тема 5. Аналіз даних методами нечіткої кластеризації

1. Чим обумовлена необхідність розвитку методів кластерного аналізу і їхнього використання?
2. На чому заснований концептуальний зв'язок між кластерним аналізом і теорією нечітких множин?
3. У чому полягає задача нечіткої кластеризації?
4. Ким і коли були запропоновані основні ідеї алгоритму для розв’язування задачі нечіткої кластеризації?
5. Які задачі дозволяють вирішувати методи нечіткої кластеризації?
6. У вигляді яких послідовних кроків можна представити алгоритм FCM?

ІНДИВІДУАЛЬНІ ЗАВДАННЯ

Індивідуальні розрахункові завдання є обов'язковою частиною самостійної роботи студента.

Написання індивідуального розрахункового завдання передбачає збір, узагальнення та аналіз статистичних даних за обраним напрямом дослідження.

Студенти обирають проблемну ситуацію із запропонованих у переліку або за власним бажанням, збирають необхідні дані, здійснюють аналіз даних із використання методів ті інструментів, що були розглянуті продовж лекційних занять та роблять відповідні висновки.

ІРГЗ оцінюються за критеріями:

- самостійності виконання;
- логічності та послідовності викладення матеріалу;
- деталізації плану;
- повноти та глибини розкриття теми, проблемної ситуації, аналітичної частини;
- наявності ілюстрацій (таблиці, рисунки, схеми і т. д.);
- кількості використаних джерел;
- використання статистичної інформації, додаткових літературних джерел та ресурсів мережі Internet;
- відображення практичного досвіду;
- обґрунтованості висновків;
- якості оформлення, презентації та захисту індивідуального розрахункового завдання.

Критерії оцінювання здобувачів вищої освіти За виконання індивідуально-розрахункової роботи

Індивідуально-розрахункова робота студента складається з двох частин: перша – опрацювання теоретичного питання, друга – виконання практичного завдання. Максимальна оцінка за виконання завдань індивідуально-розрахункової роботи – 20 балів.

Виконання завдань оцінюється за такими **критеріями**:

1) теоретичне питання:
   - повнота й ґрунтовність викладу;
   - аргументованість тверджень;
   - суб'єктне усвідомлення змісту;
   - термінологічна коректність;

2) практичне завдання:
   - технологічна грамотність;
   - методична грамотність;
   - обґрунтованість висновків;
   - правильність оформлення.

4. ПИТАННЯ, ЗАДАЧІ, ЗАВДАННЯ АБО КЕЙСИ ДЛЯ ПОТОЧНОГО ТА ПІДСУМКОВОГО КОНТРОЛЮ ЗНАНЬ І ВМІНЬ ЗДОБУВАЧІВ ВИЩОЇ ОСВІТИ, ДЛЯ КОНТРОЛЬНИХ РОБІТ, ПЕРЕДБАЧЕНИХ НАВЧАЛЬНИМ ПЛАНОМ, ПІСЛЯАТЕСТАЦІЙНОГО МОНІТОРИНГУ НАБУТИХ ЗНАНЬ І ВМІНЬ З НАВЧАЛЬНОЇ ДИСЦИПЛІНИ

4.1. ПИТАННЯ, ЗАДАЧІ, ЗАВДАННЯ АБО КЕЙСИ ДЛЯ ПОТОЧНОГО ТА ПІДСУМКОВОГО КОНТРОЛЮ ЗНАНЬ І ВМІНЬ ЗДОБУВАЧІВ ВИЩОЇ ОСВІТИ

ПИТАННЯ ДЛЯ УСНОГО ПОТОЧНОГО КОНТРОЛЮ

Поточний контроль за темою
Тема 1. Основні поняття методів багатомірної класифікації – 2 бали

1. Що розуміється під класифікацією об'єктів?
2. У чому полягає проблема класифікації об'єктів за багатомірними даними?
3. У якій формі можуть представлятися вихідні дані в задачах класифікації об'єктів?
4. Що таке навчальна вибірка?
5. Які кінцеві прикладні цілі ставить перед собою дослідник при проведенні класифікації?
6. Які фактори називаються типоутворюючими?
7. Як будується комбінаційне угруповання?
8. Яка ідея покладена в основу методу відбору найбільш інформативних ознак-детермінантів?
9. Дати характеристику класифікації як необхідного попереднього етапу статистичної обробки багатомірних даних.
10. Як використовується класифікація в задачах планування вибіркових обстежень?
11. Як класифікуються задачі розбивки об'єктів на однорідні групи залежно від наявності апріорної й попередньої вибіркової інформації?

Поточний контроль за темою
Тема 2. Кластерний аналіз. Ієрархічні методи класифікації – 2 бали

Питання для контролю знань:

1. Що таке кластерний аналіз?
2. Яка мета кластерного аналізу?
3. Чим обумовлена необхідність розвитку методів кластерного аналізу і їхнього використання?
4. Які задачі дозволяють вирішувати методи кластерного аналізу?
5. У чому суть агломеративних методів кластерного аналізу?
6. У чому суть дивізимних методів кластерного аналізу?
7. Чому вибір способу обчислення відстані між об'єктами є вузловим моментом дослідження?
8. Які відстані між об'єктами найбільш часто використовуються в задачах кластерного аналізу?
9. Які показники можуть бути використані в якості мір подібності?
10. У чому полягає сутність ієрархічних агломеративних методів?
11. Як будується дендограма?
12. Які відстані між групами об'єктів є найбільш уживаними?
13. Чому якість проведення кластеризації залежить від алгоритму об'єднання в ієрархічних агломеративних методах?

Поточний контроль за темою
Тема 3. Кластерний аналіз. Ітеративні методи класифікації – 2 бали

Питання для контролю знань:

1. У чому полягає суть ітеративних методів кластерного аналізу?
2. Яким чином в ітеративних методах задаються початкові умови?
3. Яким чином проводиться класифікація об'єктів методом k -середніх?
4. До виконання яких кроків зводяться обчислювальні процедури більшості ітеративних методів класифікації?
5. У чому складається істотна відмінність методу пошуку згущень від інших ітеративних методів класифікації?
6. У чому полягає суть ітеративного алгоритму типу «форель»?
7. Якими способами вибирається радіус сфери для пошуку локальних згущень точок у методі пошуку згущень?
8. Для якої мети використовуються критерії якості класифікації?
9. Які найпоширеніші функціонали якості використовуються в кластерному аналізі?
10. Які найпростіші прийоми дозволяють судити про якість розбивки об'єктів на кластери?

Поточний контроль за темою
Тема 4. Дискримінантний аналіз – 2 бали

Питання для контролю знань:

1. Що таке дискримінантний аналіз?
2. На які дві групи можна розбити всі процедури дискримінантного аналізу?
3. Які ознаки називаються дискримінантними змінними?
4. Які допущення приймаються в дискримінантному аналізі?
5. Яким повинне бути співвідношення числа об'єктів спостереження й числа дискримінантних змінних?
6. Як визначається канонічна дискримінантна функція?
7. Яким чином визначають коефіцієнти дискримінантної функції?
8. Як визначається границя, що розділяє розглянуті групи?
9. Яким буде алгоритм використання дискримінантного аналізу для проведення багатомірної класифікації об'єктів?
10. Як зміна числа змінних впливає на результат дискримінантного аналізу?
11. На підставі чого судять про доцільність включення (видалення) дискримінантної змінної?
12. Як розраховуються стандартизовані коефіцієнти дискримінантної функції?
13. У яких випадках застосовують стандартизовані коефіцієнти дискримінантної функції?

Поточний контроль за темою
Тема 5. Аналіз даних методами нечіткої кластеризації – 2 бали

Питання для контролю знань:

1. Чим обумовлена необхідність розвитку методів кластерного аналізу і їхнього використання?
2. На чому заснований концептуальний зв'язок між кластерним аналізом і теорією нечітких множин?
3. У чому полягає задача нечіткої кластеризації?
4. Ким і коли були запропоновані основні ідеї алгоритму для розв’язування задачі нечіткої кластеризації?
5. Які задачі дозволяють вирішувати методи нечіткої кластеризації?
6. У вигляді яких послідовних кроків можна представити алгоритм FCM?

КРИТЕРІЇ ОЦІНЮВАННЯ ЗДОБУВАЧІВ ВИЩОЇ ОСВІТИ ЗА ВІДПОВІДІ НА ПИТАННЯ ПОТОЧНОГО КОНТРОЛЮ

Завданням поточного контролю є перевірка розуміння та засвоєння певного матеріалу, вироблених навичок проведення розрахункових робіт, умінь самостійно опрацьовувати тексти, здатності осмислити зміст теми чи розділу, умінь публічно чи письмово представити певний матеріал (презентація).

Якісними критеріями оцінювання виконання завдань поточного контролю є:

1. Повнота відповіді або виконання завдання:

- елементарна;
- фрагментарна;
- повна;
- неповна.

2. Рівень сформованості логічних умінь:

- елементарні дії;
- операція, правило, алгоритм;
- правила визначення понять;
- формулювання законів і закономірностей;
- структурування суджень, доводів, описів.

ТЕСТОВІ ЗАВДАННЯ ДЛЯ ПРОМІЖНОГО КОНТРОЛЮ ЗНАНЬ

Варіант 1 (приклад)

1. По наведених ненормованих даних визначити звичайну євклідову відстань між підприємствами №№ 1,2.

а) 7,2
б) 9,6
в) 10,8
г) 14,4
Відповідь: 1) а; 2) б; 3) в; 4) г.

2. Метод мінімальної локальної відстані («найближчого сусіда») полягає в:

а) $d(S_l, S_m) = \min\limits_{\substack{x_i \in S_l \\ x_j \in S_m}} d_{ij}$; б) $d(S_l, S_m) = \max\limits_{\substack{x_i \in S_l \\ x_j \in S_m}} d_{ij}$

в) $d(S_l, S_m) = d(\bar{x}_l, \bar{x}_m)$; г) $d_{cp} = (S_l, S_m) = \dfrac{1}{n_l n_m} \sum\limits_{x_i \in S_l} \sum\limits_{x_j \in S_m} d(x_i, x_j)$

д) $V_k = \sum\limits_{i=1}^{n_k} \sum\limits_{j=1}^{p} (x_{ij} - \bar{x}_{jk})^2$

Відповідь: 1) а; 2) б; 3) в; 4) г; 5) д.

3. Метод k-середніх це:

а) агломеративний метод
б) дивізимний метод
в) ітеративний метод
Відповідь: 1) а; 2) б; 3) в.

4. Звичайна євклідова відстань розраховується по формулі:

а) $d_{ij} = \sqrt{\sum\limits_{k=1}^{m} w_k (x_{ik} - x_{jk})^2}$

б) $d_{ij} = \sqrt{\sum\limits_{k=1}^{m} (x_{ik} - x_{jk})^2}$

в) $d_{ij} = \sum\limits_{k=1}^{m} |x_{ik} - x_{jk}|$

г) $d_{ij} = \max |x_{ik} - x_{jk}|$

д) $d_{ij} = \left(\sum\limits_{k=1}^{m} |x_{ik} - x_{jk}|^p \right)^{1/r}$

Відповідь: 1) а; 2) б; 3) в; 4) г; 5) буд.

5. Утворенням кластерів приблизно рівних розмірів з мінімальною внутрішньо-кластерною варіацією характеризується:

а) Метод Уорда

б) Метод k-середніх

в) Метод «далекого сусіда»

г) Метод «найближчого сусіда»

Відповідь: 1) а; 2) б; 3) в; 4) г.

Задача. Проведіть нормування ознак №№ 1,2

Ознаки \ Підприємства	"Металургійний комбінат "Азовсталь"	"Кримський содовий завод"	"Азовкабель"	"Об'єднання "Склопластик"	"Ізюмський машинобудівний завод"	"Нікопольський трубний завод"	"Нафтопереробний комплекс "Галичина"	"Компанія "Дніпро"	"АК "Свема"	"Нікопольський завод феросплавів"
	1	2	3	4	5	6	7	8	9	10
1	36,8	33,2	34,1	37,7	35,4	36,8	33,0	36,6	32,2	36,2
2	136	122	133	138	131	136	126	134	125	133
3	9,4	6,6	7,4	10	9,3	8,6	4,0	3,5	6,6	0,9
4	0,15	0,48	0,62	0,32	0,41	0,35	0,51	0,45	0,5	0,9
5	1,91	0,88	1,09	2,62	1,45	1,78	1,94	1,67	1,32	1,89

Критерії оцінювання здобувачів вищої освіти
За виконання робіт поточного і проміжного контролю

У процесі поточного контролю здійснюється перевірка запам'ятовування та розуміння програмного матеріалу, набуття вміння і навичок конкретних розрахунків та обґрунтувань.

Шкала оцінювання за відповіді на питання поточного контролю та тестування і практичні завдання проміжного контролю знань

Поточний контроль, самостійна робота, індивідуальні завдання							
Поточне оцінювання					Проміжний контроль (тестування+задача)		Разом
Розділ 1			Розділ 2				
Т1	Т2	Т3	Т4	Т5	тести	задача	
2	2	2	2	2	20	10	40

5. ЗАВДАННЯ ПІДСУМКОВОГО КОНТРОЛЮ – СЕМЕСТРОВОГО ЕКЗАМЕНУ
(чотирирівнева шкала оцінювання)

Контрольні питання на екзамен

1. Поняття класифікації. Формалізація загальної задачі класифікації.
2. Комбінаційні угруповання та їх безперервне узагальнення.
3. Проста типологізація.
4. Зв'язна упорядкована типологізація.
5. Структурна типологізація.
6. Класифікація динамічних траєкторій розвитку систем.
7. Типологізація математичних задач класифікації.
8. Загальні поняття про кластерний аналіз.
9. Характеристика методів кластерного аналізу.
10. Відстань між об'єктами та міри близькості: евклідова відстань, зважена евклідова відстань.
11. Відстань між об'єктами та міри близькості: Хемінгова відстань, відстань Чебишева, ступінна відстань.
12. Міри схожості для об'єднання двох кластерів.
13. Метод “найближчого сусіда”.
14. Метод “дальнього сусіда”.
15. Метод середнього зв'язку.
16. Метод медіанного зв'язку.
17. Методи ієрархічного кластерного аналізу: метод одиночного зв'язку.
18. Методи ієрархічного кластерного аналізу: метод повних зв'язків.
19. Методи ієрархічного кластерного аналізу: метод середнього зв'язку.
20. Методи ієрархічного кластерного аналізу: метод Уорда.
21. Загальні поняття про ітеративні методи кластерного аналізу.
22. Математичне описання методу k-середніх.
23. Модифікації методу k-середніх.
24. Алгоритм обчислювальних процедур ітеративних методів класифікації.
25. Модифікації методу пошуку згущень.
26. Оцінка сталості угруповань об'єктів за методом пошуку згущень.
27. Основні поняття про функціонали або критерії якості.
28. Найбільш поширені функціонали якості: сума квадратів відстаней до центрів класів, сума внутрішньокласових відстаней між об'єктами, сумарна внутрішньокласова дисперсія.
29. Загальні положення дискримінантного аналізу.
30. Дискримінантні змінні та шкала відношень.
31. Геометрична інтерпретація дискримінантної функції та дискримінантних змінних.
32. Розрахунок коефіцієнтів дискримінантної функції.
33. Алгоритм класифікації при наявності двох та більше навчаючих вибірок.
34. Взаємозв'язок між дискримінантними змінними та дискримінантними функціями.
35. Постановка задачі нечіткої кластерізації.
36. Концептуальний зв'язок між кластерним аналізом і теорією нечітких множин.
37. Методи нечіткої кластеризації.

***Підсумковий контроль** за курсом* - у формі екзамену.

До складання екзамену допускають студентів, що мають задовільну кількість балів із складених тестів з основних навчальних елементів змістовних модулів, написання та захисту індивідуального науково-дослідного завдання та інших завдань передбачених програмою дисципліни.

Екзамен проводиться у відповідності до екзаменаційних білетів, які містять чотири завдання (два теоретичних питання, одне практичне завдання - задача і тести) (рис. 1). Вони дають можливість здійснити оцінювання знань студента за дисципліною.

Харківський національний університет імені В. Н.Каразіна

Факультет економічний
Спеціальність 051 «Економіка»
Освітня програма «Бізнес-аналітика та міжнародна статистика»
«Економічна аналітика та статистика»
Семестр 2
Форма навчання денна / заочна
Рівень вищої освіти (освітньо-кваліфікаційний рівень): магістр

Навчальна дисципліна: «**Методи класифікації даних в пакеті Statistica**»

ЕКЗАМЕНАЦІЙНИЙ БІЛЕТ № 1*

1. Дати загальну характеристику методів кластерного аналізу.
2. Проведіть нормування ознак №№ 4,2.

	ВАТ "Металургійний комбінат "Азовсталь"	ВАТ "Кримський содовий завод"	ВАТ "Азовкабель"	ВАТ "Об'єднання "Склопластик"	ВАТ "Джанкойський машинобудівний завод"	ВАТ "Нікопольський південнотрубний завод"	ВАТ "Нафтопереробний комплекс "Галичина"	ВАТ "Компанія "Дніпро"	ВАТ "АК "Свема"	ВАТ "Нікопольський завод феросплавів"
1	36,8	33,2	34,1	37,7	35,4	36,8	33,0	36,6	32,2	36,2
2	136	122	133	138	131	136	126	134	125	133
3	9,4	6,6	7,4	10	9,3	8,6	4,0	3,5	6,6	0,9
4	0,15	0,48	0,62	0,32	0,41	0,35	0,51	0,45	0,5	0,9

3. За наведеними ненормованими даними визначте Евклідові відстані між підприємствами №№ 1,4: а) 7,2; б) 9,6; в) 10,8; г) 14,4.
Відповідь: 1) а; 2) б; 3) в; 4) г.
4. Метод k-середніх це: а) агломеративний метод; б) дивізимний метод; в) ітеративний метод.
Відповідь: 1) а; 2) б; 3) в.

Затверджено на засіданні кафедри статистики, обліку та аудиту
Протокол №____ від „____” _____ 20____року

Завідувач кафедри _____ **Володимир СОБОЛЄВ**
(підпис) (прізвище та ініціали)
Екзаменатор _____ **Олексій КОРЕПАНОВ**
(підпис) (прізвище та ініціали)

***** За завдання 1 – 10 балів, за завдання 2 – 20 балів, за завдання 3, 4 – по 5 балів. Всього – **40 балів**.

Рис. 1. Зразок екзаменаційного білету.

Критерії оцінювання на екзамені

Оцінювання знань студента проводиться за чотирирівневою шкалою (відмінно, добре, задовільно, незадовільно). За екзамен студент може отримати максимум 40 балів:

1. Для отримання оцінки «відмінно» (35-40 балів) студент повинен:
- укластися у встановлений строк підготовки відповіді;
- викласти теоретичний матеріал чітко, коротко, зв'язно й обґрунтовано;
- навести вірне рішення задачі та тестів.

2. Для одержання оцінки «добре» (25-34 бала) студент повинен:
- укластися у встановлений строк підготовки відповіді;
- викласти теоретичний матеріал зв'язно й обґрунтовано;
- навести вірне рішення задачі;
- можливі помилки у відповідях на тести.

3. Для отримання оцінки «задовільно» (15-24 бала) студент повинен:
- викладати теоретичний матеріал у доступній для розуміння формі;
- можливі помилки при розв'язанні задачі та в тестах.

4. Оцінку «незадовільно» (1-14 балів) отримують студенти, відповіді яких можуть бути оцінені нижче вимог, сформульованих у попередніх пунктах.

Кожне завдання заліку оцінюється окремо. Загальна оцінка дорівнює сумі оцінок за всі завдання (види контролю). Якщо одна з оцінок «незадовільно», то загальна оцінка не може бути вищою за «задовільно».

Підсумкова оцінка з навчальної дисципліни визначається як загальна оцінка, яка враховує оцінки з кожного виду контролю (оцінки поточного та проміжного контролю за роботу протягом семестру, індивідуального завдання та оцінка за результатами підсумкового контролю).

У відповідності до набраних студентом балів оцінка знання матеріалу проводиться за чотирирівневою системою згідно з Методикою переведення показників успішності знань студентів.

Шкала оцінювання

Сума балів за всі види навчальної діяльності протягом семестру	Оцінка за національною шкалою
	для **чотирирівневої** шкали оцінювання
90-100	відмінно
80-89	добре
70-79	
60-69	задовільно
50-59	
1-49	незадовільно

6. КРИТЕРІЇ ОЦІНЮВАННЯ ЗНАНЬ СТУДЕНТІВ ТА РОЗПОДІЛ БАЛІВ

КРИТЕРІЇ ОЦІНЮВАННЯ (ЗАСОБИ ДІАГНОСТИКИ)

Критерії оцінювання результативності роботи студентів при виконанні самостійної роботи

Якісними критеріями оцінювання виконання індивідуальних завдань студентами є:

1.Повнота виконання завдання:

- Елементарна;
- Фрагментарна;
- Повна;
- Неповна.

2. Рівень самостійності студента

- під керівництвом викладача;
- консультація викладача;
- самостійно.

3. Сформованість навчально-інформаційних умінь (роботи з підручником, володіння різними способами читання, складання плану, рецензій, конспекту, вміння користуватися бібліотекою, спостереження, експеримент тощо)

4. Сформованість навчально-інтелектуальних умінь (визначення понять, аналіз, синтез, порівняння, класифікація, систематизація, узагальнення, абстрагування, вміння відповідати на запитання, виконувати творчі завдання тощо);

5. Рівень сформованості фахових методичних вмінь (вміння застосовувати на практиці набуті знання):

- низький – володіння умінням здійснювати первинну обробку навчальної інформації без подальшого її аналізу;
- середній – уміння вибирати відомі способи дій для виконання фахових завдань;
- достатній – застосовує набуті знання у стандартних практичних ситуаціях;
- високий – володіння умінням творчо-пошукової діяльності.

Критерії оцінювання здобувачів вищої освіти за відповіді на питання поточного контролю

Завданням поточного контролю є перевірка розуміння та засвоєння певного матеріалу, вироблених навичок проведення розрахункових робіт, умінь самостійно опрацьовувати тексти, здатності осмислити зміст теми чи розділу, умінь публічно чи письмово представити певний матеріал (презентація).

Якісними критеріями оцінювання виконання завдань поточного контролю є:

1. *Повнота відповіді або виконання завдання:*
   - елементарна;
   - фрагментарна;
   - повна;
   - неповна.
2. *Рівень сформованості логічних умінь:*
   - елементарні дії;
   - операція, правило, алгоритм;
   - правила визначення понять;
   - формулювання законів і закономірностей;
   - структурування суджень, доводів, описів.

Критерії оцінювання здобувачів вищої освіти за виконання індивідуальної розрахункової роботи

Індивідуально-розрахункова робота студента складається з двох частин: перша – опрацювання теоретичного питання, друга – виконання практичного завдання. Максимальна оцінка за виконання завдань індивідуальної розрахункової роботи – 20 балів.

Виконання завдань оцінюється за такими критеріями:

1) теоретичне питання:
   - повнота й ґрунтовність викладу;
   - аргументованість тверджень;
   - суб'єктне усвідомлення змісту;
   - термінологічна коректність;
2) практичне завдання:
   - технологічна грамотність;
   - методична грамотність;
   - обґрунтованість висновків;
   - правильність оформлення.

Шкала оцінювання індивідуальної розрахункової роботи

Кількість балів	Теоретичне питання	Практичне завдання
15-20	Повне засвоєння та суб'єктне усвідомлення матеріалу. Твердження чітко аргументовані. Продемонстровано термінологічну грамотність	Продемонстровано методичну й технологічну грамотність. Методичні рішення обґрунтовано. Оформлення відповідає вимогам.
10-14	Повне засвоєння матеріалу, але недостатнє суб'єктне його усвідомлення. Нечітка аргументація тверджень. Часткова термінологічна некоректність.	Наявність незначних методичних і технологічних помилок, а також помилок в оформленні роботи. Методичні рішення обґрунтовано.
5-9	Часткове засвоєння матеріалу, суб'єктне його не усвідомлення. Аргументація відсутня. Термінологічна неграмотність.	Наявність значної кількості методичних і технологічних помилок, а також в оформленні роботи.
0-4	Теоретичний матеріал не засвоєно. Аргументація відсутня. Термінологічна неграмотність.	Методична й технологічна неграмотність. Неправильне оформлення роботи.

Критерії оцінювання на екзамені

Оцінювання знань студента проводиться за чотирирівневою шкалою (відмінно, добре, задовільно, незадовільно). За екзамен студент може отримати максимум 40 балів:

1. Для отримання оцінки «відмінно» (35-40 балів) студент повинен:
   - укластися у встановлений строк підготовки відповіді;
   - викласти теоретичний матеріал чітко, коротко, зв'язно й обґрунтовано;
   - навести вірне рішення задачі та тестів.
2. Для одержання оцінки «добре» (25-34 бала) студент повинен:
   - укластися у встановлений строк підготовки відповіді;
   - викласти теоретичний матеріал зв'язно й обґрунтовано;
   - навести вірне рішення задачі;
   - можливі помилки у відповідях на тести.

3. Для отримання оцінки «задовільно» (15-24 бала) студент повинен:

- викладати теоретичний матеріал у доступній для розуміння формі;
- можливі помилки при розв'язанні задачі та в тестах.

4. Оцінку «незадовільно» (1-14 балів) отримують студенти, відповіді яких можуть бути оцінені нижче вимог, сформульованих у попередніх пунктах.

Кожне завдання екзамену оцінюється окремо. Загальна оцінка дорівнює сумі оцінок за всі завдання (види контролю). Якщо одна з оцінок «незадовільно», то загальна оцінка не може бути вищою за «задовільно».

Шкала оцінювання екзаменаційної роботи (всього 40 балів):

– за правильну відповідь на теоретичне питання завдання 1 студент одержує 10 балів;

– за правильно виконане завдання 2 (розрахункова задача) – 20 балів;

– за правильно виконані завдання 3, 4 (тести) студент одержує по 5 балів.

Підсумкова оцінка з навчальної дисципліни визначається як загальна оцінка, яка враховує оцінки з кожного виду контролю (оцінки поточного та проміжного контролю за роботу протягом семестру, індивідуального завдання та оцінка за результатами підсумкового контролю).

Зведена шкала оцінювання роботи студентів з дисципліни «Методи класифікації даних в пакеті Statistica»

Види робіт	Максимум балів
Поточне оцінювання	10
Проміжний контроль (тестування)	30
Індивідуальне завдання	20
РАЗОМ	**60**
Екзамен/залік	**40**
ВСЬОГО	**100**

У відповідності до набраних студентом балів оцінка знання матеріалу проводиться за чотирирівневою/дворівневою системою згідно з Методикою переведення показників успішності знань студентів.

Шкала оцінювання

Сума балів за всі види навчальної діяльності протягом семестру	Оцінка за національною шкалою
	для чотирирівневої шкали оцінювання
90-100	відмінно
80-89	добре
70-79	
60-69	задовільно
50-59	
1-49	незадовільно

СХЕМА НАРАХУВАННЯ БАЛІВ

Оцінювання знань, вмінь та навичок студентів включає ті види занять, які згідно з програмою навчальної дисципліни "Методи класифікації даних в пакеті Statistica" передбачають лекційні, практичні заняття, самостійну роботу та виконання індивідуального науково-дослідного завдання.

Перевірка та оцінювання знань студентів проводиться в наступних формах:

- поточне оцінювання роботи і знань студентів під час практичних занять;
- складання проміжного контролю знань за розділами (тестування);
- оцінювання виконання та захист індивідуального науково-дослідного завдання;
- складання екзамену.

Структура засобів контролю та розподіл балів із дисципліни "Методи класифікації даних в пакеті Statistica" наведена в табл.

Узагальнена схема нарахування балів (денна форма навчання)

Поточний контроль, самостійна робота, індивідуальні завдання									Екзаменаційна робота	Сума
Поточне оцінювання					Проміжний контроль (тестування)	Контрольна робота, передбачена навчальним планом	Індивідуальне завдання	Разом		
Розділ 1			Розділ 2							
Т1	Т2	Т3	Т4	Т5						
2	2	2	2	2	30	-	20	60	40	**100**

Т1, Т2 ... – теми розділів.

Узагальнена схема нарахування балів (заочна форма навчання)

Поточний контроль, самостійна робота, індивідуальні завдання									Екзаменаційна робота	Сума
Поточне оцінювання					Проміжний контроль (тестування)	Контрольна робота, передбачена навчальним планом	Індивідуальне завдання	Разом		
Розділ 1			Розділ 2							
Т1	Т2	Т3	Т4	Т5						
2	2	2	2	2	30	-	20	60	40	**100**

Т1, Т2 ... – теми розділів.

Поточне оцінювання знань студентів здійснюється під час проведення практичних і має на меті перевірку рівня підготовленості студента до виконання конкретної роботи. Об'єктами поточного контролю є:

- активність та результативність роботи студента протягом семестру над вивченням програмного матеріалу дисципліни;
- відвідування занять;
- виконання індивідуального розрахункового завдання;
- складання проміжного контролю знань (тестування).

Контроль систематичного виконання самостійної роботи та активності на практичних заняттях проводиться за такими критеріями:

- розуміння, ступінь засвоєння теорії та методології проблем, що розглядаються;
- ступінь засвоєння фактичного матеріалу навчальної дисципліни;
- ознайомлення з рекомендованою літературою, а також із сучасною літературою з питань, що розглядаються;
- уміння поєднувати теорію з практикою при розгляді практичних ситуацій, розв'язанні задач, проведенні розрахунків при виконанні індивідуальних завдань, та завдань, винесених на розгляд в аудиторії;

– оволодіння методами економіко-статистичної обробки даних з використанням комп'ютерних технологій;

– логіка, структура, стиль викладу матеріалу в письмових роботах і при виступах в аудиторії, вміння обґрунтовувати свою позицію, здійснювати узагальнення інформації та робити висновки.

Оцінювання знань студента під час виконання завдань для самостійної роботи проводиться за чотирирівневою шкалою.

Оцінка «відмінно» ставиться за умови відповідності виконаного завдання студента або його усної відповіді до всіх зазначених критеріїв. Відсутність тієї чи іншої складової знижує оцінку.

При оцінюванні практичних занять увага приділяється також їх якості та самостійності, своєчасності здачі виконаних завдань викладачу (згідно з графіком навчального процесу). Якщо якась із вимог не буде виконана, то оцінка буде знижена.

***Проміжний контроль* (*тестування*)** рівня знань передбачає виявлення опанування студентом лекційного матеріалу та вміння застосування його для вирішення практичної ситуації і проводиться у вигляді тестування. При цьому тестове завдання може містити як запитання, що стосуються суто теоретичного матеріалу, так і запитання, спрямовані на вирішення невеличкого практичного завдання.

Проміжний тестовий контроль проводиться один раз на семестр. Загальна тривалість тестів – 1,5 години. Поточне тестування складається з 10 тестів і 1 практичної задачі. Одна правильна відповідь на кожен з тестів дорівнює 2 балам, задача – 10 балів. Тестове завдання містить запитання одиничного і множинного вибору різного рівня складності.

Тести можуть бути застосовані як з метою контролю, так і для закріплення теоретичних знань і практичних навичок.

Тести для проміжного контролю обираються із загального переліку тестів за відповідними темами.