ДИСКРИМІНАНТНИЙ АНАЛІЗ ПРИ УМОВІ НОРМАЛЬНОГО РОЗПОДІЛУ ПОКАЗНИКІВ

Єрмолаєва Вікторія, Анатолій Плічко

Науковий керівник: д. фіз.-мат. наук, професор Плічко А.М.

У статті розглянуто історію становлення та основні поняття дискримінантного аналізу. Детально описаний дискримінантний аналіз при умові нормального розподілу показників. Наведено приклад розв'язання задачі дискримінантного аналізу при цій умові. Детально описано метод класифікації на основі вказаного методу.

Ключові слова: *дискримінантний аналіз, кореляція, дисперсія, класифікація об'єктів.*

DISCRIMINATIVE ANALYSIS UNDER CONDITION OF NORMAL DISTRIBUTION OF INDICATORS

Yermolaieva Viktoriia, Anatolii Plichko

Scientific adviser: Doctor of Phys. and Math. sciences, Professor Plichko A.M.

The history of formation and basic concepts of discriminant analysis are considered in the article. The concept of discriminant analysis under the condition of normal distribution of indicators is described in detail. An example of solving the problem of discriminant analysis under this condirion is given. The method of classification on the basis of the specified method is described in detail.

Keywords: *discriminant analysis, correlation, dispersion, classification of objects.*

Мета статті: визначити що являє собою дискримінантний аналіз, навести приклад задачі при умові нормального розподілу показників, а також пояснити для чого необхідний саме такий метод дискримінантного аналізу.

Основні результати дослідження. Дискримінантний аналіз – один з методів багатовимірного аналізу, метою якого є класифікація об'єктів, тобто віднесення об'єкта до однієї з відомих груп деяким оптимальним способом (наприклад, розбиття сукупності підприємств на кілька однорідних груп за значеннями будь-яких показників виробничо-господарської діяльності).

Методи дискримінантного аналізу розроблялися починаючи з кінця 1950-х рр. такими вченими, як Прасанта Чандра Махаланобіс (індійський економіст і статистик, 1893-1972), Гарольд Готелінґ (американський економіст і статистик, 1895-1973), Рональд Фішер (англійський статистик, біолог-еволюціоніст, генетик, 1890-1962), і іншими.

Відміною властивістю дискримінантного аналізу як методу класифікації є те, що досліднику заздалегідь відомі число груп, на які потрібно розбити розглянуту сукупність об'єктів, і їх властивості; відомо також, що об'єкт напевно належить до однієї з певних груп (але до якої саме - невідомо).

У відповідності з властивостями дискримінантного аналізу виникають завдання двох типів:

1. опису відмінностей між класами;
2. класифікації об'єктів, що не входили до первісної навчальної вибірки.

Для вирішення першого завдання будуються канонічні дискримінантні функції, які дозволяють з максимальною ефективністю розділити класи.

Для вирішення другого завдання (класифікації об'єктів, що не входили до первісної навчальної вибірки) обчислюються відстані від кожного нового об'єкта, що підлягає класифікації, до геометричного центру (центру ваги) кожного класу.

Дискримінантний аналіз висуває суворі вимоги до вихідних даних: в моделі повинно бути не менше двох класів, в кожному класі – не менше двох об'єктів з навчальної вибірки, число дискримінантних змінних не повинно перевищувати обсяг навчальної вибірки, дискримінантні змінні повинні бути кількісними і лінійно незалежними. Для кожного класу потрібні приблизна рівність коваріаційних матриць, а також багатовимірна нормальність розподілу.

Для того щоб використовувати дискримінантний аналіз необхідно дотримуватися певних умов, бо без цього результати не матимуть значення.

1. Досліджувана сукупність повинна мати нормальний розподіл. Передбачається, що аналізовані змінні представляють вибірку з багатовимірного нормального розподілу. Відзначимо, однак, що нехтування умовою нормальності зазвичай не є фатальним в тому сенсі, що результуючі критерії значимості все ще заслуговують довіри, але все ж найбільш достовірні результати виходять в разі, коли величина має нормальний розподіл.
2. Однорідність дисперсій. Передбачається, що матриці дисперсій змінних однорідні, тобто мають близькі значення. Як і раніше, малі відхилення не фатальні, однак перш ніж зробити остаточні висновки при важливих дослідженнях, непогано звернути увагу на внутрішньогрупові матриці дисперсій і кореляцій. Зокрема, можна побудувати матричну діаграму розсіювання, вельми корисну для цієї мети.
3. Кореляції між середніми і дисперсіями. Некоректність застосування критеріїв значимості виникає через можливу залежність між середніми за сукупностями і дисперсіями (або стандартними відхиленнями) між собою. Якщо є велика мінливість в сукупності з високими середніми в декількох змінних, то ці високі середні ненадійні. Однак критерії значимості ґрунтуються на об'єднаних дисперсіях, тобто, на середній дисперсії за всіма сукупностями. Тому критерії значимості для відносно великих середніх (з великими дисперсіями) будуть засновані на відносно менших об'єднаних дисперсіях і будуть помилково вказувати на статистичну значущість. На практиці цей варіант може статися також, якщо одна з досліджуваних сукупностей містить кілька екстремальних викидів, які сильно впливають на середні і, таким чином, збільшують мінливість. Для визначення такого випадку слід вивчити описові статистики, тобто середні і стандартні відхилення або дисперсії для таких кореляцій.
4. Завдання з погано обумовленою матрицею. Інше припущення в дискримінантному аналізі полягає в тому, що змінні, які використовуються для дискримінації між сумами, не є повністю надлишковими. При обчисленні результатів дискримінантного аналізу відбувається обернення матриці дисперсій змінних в моделі. Якщо одна із змінних лінійно залежна від інших змінних, то

відповідна матриця називається погано обумовленою. Наприклад, якщо одна змінна є сумою трьох інших, то це відіб'ється також і в моделі, і розглянута матриця буде погано обумовленою.

5. Значення толерантності. Щоб уникнути поганої обумовленості матриць, необхідно постійно перевіряти так звані значення толерантності для кожної змінної.

Приклад використання дискримінантного аналізу при умові нормального розподілу показників.

Маємо дві генеральні сукупності *X* та *Y*, які мають трьохвимірний закон розподілу з невідомими, але рівними коваріаційними матрицями. Із них взяті навчальні вибірки з об'ємами n_1 в *X* та n_2 в *Y*.

$$X = \begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ \dots & \dots & \dots \\ x_{n_1 1} & x_{n_1 2} & x_{n_1 3} \end{pmatrix}$$

$$Y = \begin{pmatrix} y_{11} & y_{12} & y_{13} \\ y_{21} & y_{22} & y_{23} \\ \dots & \dots & \dots \\ y_{n_2 1} & y_{n_2 2} & y_{n_2 3} \end{pmatrix}$$

Ціллю дискримінантного аналізу є віднесення нового спостереження (рядку матриці *Z*) або до *X*, або до *Y*.

$$Z = \begin{pmatrix} z_{11} & z_{12} & z_{13} \\ z_{21} & z_{22} & z_{23} \\ \dots & \dots & \dots \\ z_{i1} & z_{i2} & z_{i3} \end{pmatrix}$$

Для розв'язання задачі, за навчальними вибірками знайдемо вектори середніх.

$$\bar{X} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \bar{x}_3 \end{pmatrix} \quad \text{і} \quad \bar{Y} = \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \bar{y}_3 \end{pmatrix}$$

1. Визначимо оцінки коваріаційних матриць

$$S_x = \{S_{ki}\}_x \quad \text{і} \quad S_y = \{S_{ki}\}_y;$$

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

Знайдемо елемент матриці S_x :

$$S_{kj}(x) = \frac{1}{n_1}\sum_{i=1}^{n_1}\left(x_{ij} - \bar{x}_j\right)\left(x_{ik} - \bar{x}_k\right) = \overline{x_j x_k} - \bar{x}_j\bar{x}_k; \quad j,k = 1,2,3,$$

де $\bar{x}_i$ та $\bar{x}_k$ – середні значення.

2. Підрахуємо незміщену оцінку сумарної коваріаційної матриці

$$\hat{S} = \frac{1}{n_1 + n_2 - 2}\left(n_1 S_x + n_2 S_y\right).$$

3. Знайдемо обернену матрицю.
4. Підрахуємо вектор оцінок коефіцієнтів дискримінантної функції

$$a = \hat{S}^{-1}(\bar{X} - \bar{Y}).$$

5. Підрахуємо оцінки векторів значень дискримінантної функції для матриць вхідних даних $\widehat{U}_x = Xa$, $\widehat{U}_y = Ya$.
6. Підрахуємо середні значення оцінок дискримінантної функції

$$\bar{\hat{u}}_x = \frac{1}{n_1}\sum_{i=1}^{n_1}\hat{u}_{xi}, \quad \bar{\hat{u}}_y = \frac{1}{n_2}\sum_{i=1}^{n_2}\hat{u}_{yi}.$$

7. Знайдемо константу

$$\hat{C} = \frac{1}{2}(\bar{\hat{u}}_x + \bar{\hat{u}}_y)$$

Дискримінантну функцію для *v*-того спостереження, який підлягає дискримінації, отримаємо з рівності $\hat{u}_v = z_{v_1}a_1 + z_{v_2}a_2 + z_{v_3}a_3$.

Якщо $\hat{u}_v \geq \hat{C}$, то *v*-те спостереження потрібно віднести до сукупності *X*, якщо $\hat{u}_v < \hat{C}$, то *v*-те спостереження потрібно віднести до сукупності *Y*.

Приклад дискримінантного аналізу.

Діяльність кожного виробничого об’єднання галузі оцінювались за такими трьома показниками:

а) середньорічна вартість основних виробничих фондів (ОВФ);

б) середньооблікова чисельність промислово-виробничого персоналу (ПВП);

в) балансовий прибуток.

В галузі виокремлені дві групи: передова, яка складається з чотирьох об'єднань, та інша, яка включає в себе п'ять об'єднань. Дані представлені у таблиці.

Галузі було передано об'єднання *Z*, яке має такі показники:

вартість ОВФ - 55,451;
чисельність ПВП - 9,592 тис. людей;
балансовий прибуток - 12,840.

Потрібно визначити чи можна віднести нове об'єднання до передової групи виробництв галузі.

Розв'язання

Показники / Група об'єднань	Вартість ОПФ	Чисельність ППП	Балансовий прибуток
Передова	224,228	17,115	22,981
	151,827	14,904	21,481
	147,313	13,627	28,669
	152,253	10,545	10,199
Решта	46,757	4,428	11,124
	29,033	5,510	6,091
	52,134	4,214	11,842
	37,050	5,527	11,873
	63,979	4,211	12,860

$$X = \begin{pmatrix} 224{,}228 & 17{,}115 & 22{,}981 \\ 151{,}827 & 14{,}904 & 21{,}481 \\ 147{,}313 & 13{,}627 & 28{,}669 \\ 152{,}253 & 10{,}545 & 10{,}199 \end{pmatrix}$$

$$Y = \begin{pmatrix} 46{,}751 & 4{,}428 & 11{,}124 \\ 29{,}033 & 5{,}510 & 6{,}091 \\ 52{,}134 & 4{,}214 & 11{,}842 \\ 37{,}050 & 5{,}527 & 11{,}873 \\ 63{,}979 & 4{,}211 & 12{,}860 \end{pmatrix}$$

де $n_1 = n_x = 4$; $n_2 = n_y = 5$;

рядок матриці *Z*: $Z^T = (55{,}451\ \ 9{,}592,\ \ 12{,}840)$.

2. Отримаємо вектори середніх

$$\bar{X} = \begin{pmatrix} 168{,}92025 \\ 14{,}04775 \\ 20{,}8325 \end{pmatrix}; \qquad \bar{Y} = \begin{pmatrix} 45{,}7926 \\ 4{,}778 \\ 10{,}758 \end{pmatrix}.$$

3. Отримаємо оцінку коваріаційних матриць

$$S_x = \begin{pmatrix} 1025{,}61 & 55{,}66575 & 28{,}94475 \\ & 5{,}6468625 & 10{,}27365 \\ & & 44{,}879675 \end{pmatrix};$$

$$S_y = \begin{pmatrix} 145{,}8666 & -6{,}60952 & 22{,}78694 \\ & 0{,}371782 & -0{,}902484 \\ & & 5{,}750302 \end{pmatrix}.$$

4. Отримаємо незміщену оцінку сумарної коваріаційної матриці

$$\hat{S} = \frac{1}{4+5-2}(4S_x + 5S_y);$$

$$\hat{S} = \begin{pmatrix} 690{,}25328 & 27{,}087914 & 32{,}816242 \\ & 3{,}4923371 & 5{,}2260257 \\ & & 29{,}752887 \end{pmatrix}.$$

5. Знайдемо обернену матрицю

$$\hat{S}^{-1} = \begin{pmatrix} 0{,}0020945371 & -0{,}017349116 & 0{,}00073714 \\ & 0{,}53214303 & -0{,}07433441 \\ & & 0{,}04565381 \end{pmatrix}.$$

6. Знайдемо вектор оцінки коефіцієнтів дискримінації

$$a = \hat{S}^{-1}(\bar{X} - \bar{Y}) = \hat{S}^{-1}\begin{pmatrix} 123{,}12765 \\ 9{,}26975 \\ 10{,}0745 \end{pmatrix} - \begin{pmatrix} 0{,}10449979 \\ 2{,}0475006 \\ -0{,}13634981 \end{pmatrix}.$$

7. Розрахуємо оцінки дискримінантної функції

$$\widehat{U}_x = Xa = \begin{pmatrix} 55{,}346433 \\ 43{,}457381 \\ 39{,}3990544 \\ 36{,}113833 \end{pmatrix}; \qquad \widehat{U}_y = \begin{pmatrix} 12{,}437003 \\ 13{,}486817 \\ 12{,}46277 \\ 13{,}571031 \\ 13{,}555623 \end{pmatrix}.$$

8. Визначимо середнє значення оцінок дискримінантної функції

$$\bar{\hat{u}}_x = 43{,}577047; \qquad \bar{\hat{u}}_y = 13{,}102648.$$

9. Отримаємо константу

$$\hat{C} = \frac{1}{2}(43{,}577047 + 13{,}102648) = 28{,}339847.$$

10. Визначимо можливість включення нового об'єднання до групи передових. Оскільки матриця цього об'єднання складається лише з одного рядка, то

$$\hat{u}_z = a_1 z_1 + a_2 z_2 + a_3 z_3,$$

$$\hat{u}_z = 0{,}10449979 \cdot 55{,}451 + 2{,}0478006 \cdot 9{,}952 - 0{,}13634981 \cdot 12{,}840 \approx 23{,}69$$

Середнє значення дискримінантної функції менше ніж константа

$$\hat{u}_z < \hat{C};$$

$$23{,}69 < 28{,}34.$$

Висновок: об'єднання Z не можна віднести до передової групи.

Висновки. Отже, головним призначенням дискримінантного аналізу є класифікація даних. За його допомогою можна визначити, які змінні найкраще розділяють об'єкти чи процеси, що досліджуються, на дві чи більше груп. Рішення про належність або неналежність до групи приймається на основі значення дискримінантної функції. Окрім цього, дискримінантний аналіз можна використовувати й для виявлення властивостей або факторів, які об'єднують досліджуванні об'єкти в наперед задані класи. Ці властивості називають «групуючими».

Список використаних джерел

1. С.А.Айвязян, В.М.Бухштабер и др. «Прикладная статистика. Классификация и снижение размерности». Москва, 1989
2. А.М. Дубров, В.С. Мхитарян, Л.И. Трошин, Многомерные статистические методы, Москва, "Финансы и статистика", 2003