35. Дискриминантный анализ: постановка задачи, исходные данные. Канонические дискриминантные функции, их значимость, их использование для классификации объектов.

1. Постановка задачи и исходные данные

Дискриминантный анализ — это раздел многомерного статистического анализа, целью которого является решение задач различения (дискриминации) объектов наблюдения по определенным признакам.

В зависимости от целей выделяют две группы задач:

1. **Интерпретация межгрупповых различий:** определение того, можно ли отличить один класс от другого по данному набору характеристик, и какие из них наиболее информативны.
2. **Классификация:** нахождение правила (функции), позволяющего отнести новый объект к одной из заранее известных групп.

Исходные данные («Вход» задачи)

Для проведения анализа исследователь должен иметь:

1. **Матрицу данных** X типа «объект-свойство» размерности $n \times p$:

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

   где x_{ij} — значение j-го признака для i-го объекта.
2. **Обучающие выборки:** g групп (классов), где $g \geq 2$. Число объектов в каждой обучающей выборке k должно быть не менее двух ($n_k \geq 2$).

Обозначения

- g — число классов;
- p — число дискриминантных переменных (признаков);
- n_i — число объектов класса i;
- n — общее число объектов.

Требования к дискриминантным переменным

1. Измеряются в интервальной шкале или шкале отношений.
2. Должны быть линейно независимыми.
3. Распределены по многомерному нормальному закону.
4. Их число не должно превосходить общее число наблюдений за вычетом двух ($0 < p < n - 2$).
5. Предполагается приблизительное равенство ковариационных матриц для каждого класса.

2. Канонические дискриминантные функции

Для разделения классов переходят от исходных признаков к **дискриминантным функциям**. В случае линейного дискриминантного анализа каноническая функция имеет вид линейной комбинации переменных:

$$f(x) = a_0 + \sum_{j=1}^{p} a_j x_j$$

где a_j — коэффициенты, показывающие вклад переменной в разделение.

Геометрическая интерпретация и вывод коэффициентов (для 2-х классов)

Задача сводится к определению новой оси координат, такой, чтобы проекции объектов разных классов на эту ось были максимально разделены (центры классов $\overline{f}_1$ и $\overline{f}_2$ максимально удалены друг от друга), а разброс внутри классов был минимальным.

Вектор коэффициентов A определяется из условия максимизации отношения межгрупповой вариации к внутригрупповой.

1. **Внутригрупповая вариация (W):**
   Рассматривается как сумма квадратов отклонений значений функции от среднего по группе.

$$W = A^T(n_1 + n_2 - 2)S_* A$$

   где S_* — объединенная ковариационная матрица:

$$S_* = \frac{1}{n_1 + n_2 - 2}(Y_1^T Y_1 + Y_2^T Y_2)$$

   (Y_k — матрица центрированных значений признаков в k-й группе).
2. **Межгрупповая вариация (V):**
   Определяется как квадрат расстояния между средними значениями функций двух групп:

$$V = (\overline{f}_1 - \overline{f}_2)^2 = A^T(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T A$$

   где μ_k — вектор средних значений признаков для группы k.
3. **Критерий Фишера (F):**
   Необходимо найти такой вектор A, чтобы:

$$F = \frac{V}{W} \to \max$$

   Для нахождения максимума берутся частные производные по вектору коэффициентов и приравниваются к нулю: $\frac{\partial F}{\partial A} = 0$.
4. **Решение:**
   В результате дифференцирования для случая двух классов вектор коэффициентов равен:

$$A = S_*^{-1}(\mu_1 - \mu_2)$$

Обобщенный случай (более 2-х классов)

Для g групп задача сводится к решению обобщенной задачи на собственные значения и собственные векторы с условием нормировки:

$$\begin{cases} (V - \lambda W)A = 0, \\ A^T A = 1 \end{cases}$$

где A — искомый вектор коэффициентов, λ — собственное число.

Количество канонических функций m определяется как:

$$m = \min\{g - 1, p\}$$

3. Значимость дискриминантных функций

Каждая полученная функция характеризуется собственным числом λ_i. Чем больше λ_i, тем большей разделительной способностью обладает функция. Для оценки значимости используются следующие критерии:

1. Относительное процентное содержание (τ_i)

Показывает, насколько одна функция «сильнее» других:

$$\tau_i = \frac{\lambda_i}{\sum_{j=1}^{m} \lambda_j} \cdot 100\%$$

2. Коэффициент канонической корреляции (r_j^*)

Мера связи между группами и дискриминантной функцией (аналог η^2 в дисперсионном анализе). Чем ближе к 1, тем лучше разделение:

$$r_j^* = \sqrt{\frac{\lambda_j}{1 + \lambda_j}}$$

3. Лямбда-статистика Уилкса (Λ)

Оценивает **остаточную дискриминантную способность** (способность различать группы без учета информации от уже вычисленных функций).

$$\Lambda = \prod_{i=k+1}^{g} \frac{1}{1 + \lambda_i}$$

где k — число уже вычисленных функций.

- $\Lambda \approx 0$ — высокое различение.
- $\Lambda \approx 1$ — низкое различение.

4. Критерий Хи-квадрат (χ^2)

Используется для проверки статистической значимости Λ-статистики (проверка гипотезы о том, что оставшиеся функции не дают значимого разделения).

$$\chi^2_{\text{набл}} = -\left[n - \frac{p+g}{2}\right] \ln \Lambda_k$$

Если $\chi^2_{\text{набл}} > \chi^2_{\text{кр}}$, то дискриминация значима, и имеет смысл вычислять следующую функцию.

4. Использование функций для классификации объектов

Процедура классификации с использованием дискриминантных функций (для случая 2-х классов):

1. Вычисляются коэффициенты A и значения дискриминантной функции для каждого объекта обучающей выборки.
2. Находятся средние значения функции для каждой группы (центроиды): $\overline{f}_1$ и $\overline{f}_2$.
3. Определяется **константа дискриминации** C (разделяющая граница):

$$C = 0,5(\overline{f}_1 + \overline{f}_2)$$

4. Для нового объекта X_γ рассчитывается значение функции:

$$f_\gamma = X_\gamma \cdot A$$

5. **Правило классификации:**
   - Если $f_\gamma > C$, объект относится к первой группе.
   - Если $f_\gamma < C$, объект относится ко второй группе.
     (При условии, что $\overline{f}_1 > \overline{f}_2$, иначе знаки меняются).

Для случая $g > 2$:
Рассчитываются $g - 1$ функций. Пространство разбивается гиперплоскостями. Рассчитываются

константы разделения C_{ij} между парами центроидов. Новый объект относится к тому классу, в интервал которого попадает значение его дискриминантной функции.