37. Дискриминантный анализ. Классификация с использованием вероятностно-статистических методов (расстояние Махаланобиса и апостериорная вероятность). Простые классифицирующие функции и их применение для классификации.

Дискриминантный анализ — это раздел многомерного статистического анализа, предназначенный для решения задач классификации объектов в ситуациях, когда **группы (классы) известны заранее**. В отличие от кластерного анализа, этот метод относится к обучению «с учителем», так как строится на основе обучающей выборки, где принадлежность каждого объекта к определенной группе уже определена.

Расстояние Махаланобиса

Расстояние Махаланобиса представляет собой меру близости конкретного объекта к центроиду (средней точке) каждой из известных групп в многомерном пространстве признаков.

В тех случаях, когда переменные коррелированны, измерены в разных единицах и имеют разные стандартные отклонения, бывает трудно определить понятие «расстояние». В 1963г. Махаланобис предложил использовать обобщенную меру расстояния, которая устраняет эти трудности:

$$d^2(X/G_k) = (x - \mu_k)^T S_*^{-1}(x - \mu_k) \qquad (23)$$

Здесь $d^2(X/G_k)$ — квадрат расстояния от объекта X (данный объект) до центра класса k; μ_k — вектор средних значений переменных для класса k; S_* — обобщенная ковариационная матрица внутригрупповых ковариаций.

После вычисления d^2 для каждого объекта, его относят в класс с наименьшим значением d^2. Это класс, чей типичный профиль по дискриминантным переменным больше похож на профиль для этого объекта.

- **Особенности метрики:** В отличие от обычного евклидова расстояния, расстояние Махаланобиса **учитывает корреляционные связи** между переменными и их разную изменчивость (дисперсии). Это делает его более точным инструментом для анализа сложных статистических данных, где признаки взаимозависимы.
- **Правило классификации:** Объект относится к той группе, **расстояние Махаланобиса до центра которой минимально**,.
- **Практическая реализация:** В статистических программах (например, Statistica) результаты часто представляются в виде матрицы квадратов расстояний Махаланобиса от объектов до центров всех выделенных групп.

Апостериорная вероятность

Апостериорная вероятность — это вероятность того, что объект принадлежит к определенному классу, вычисленная **после (post) получения и анализа значений его признаков**.

Обозначим $P(X/G_k)$ вероятность того, что объект X действительно принадлежит классу k. Оценить эту величину можно как долю объектов в классе G_k, которые отстоят от центра дальше, чем наш объект:

$$P(X/G_k) = \frac{\widetilde{m_k}}{n_k} \tag{24}$$

где $\widetilde{m_k}$ — число объектов в классе k, расположенных от центра «дальше», чем X; n_k — число объектов в классе k, причем $n_k = m_k + \widetilde{m_k}$, а m_k — число объектов в классе k, расположенных к центру «ближе», чем X.

Для объекта X сумма данных вероятностей не равна 1, так как кассы могут перекрываться.

Апостериорная вероятность того, что объект X является членом класса k рассчитывается по формуле:

$$P_{\text{ап}}(G_k/X) = \frac{P(X/G_k) \cdot P(G_k)}{\sum_{k=1}^{g} P(X/G_k) \cdot P(G_k)} \tag{25}$$

где $P(G_k)$ — априорная вероятность принадлежности объектов к классу G_k; $P(X/G_k)$ — условная вероятность того, что объект X является членом класса k (доля объектов в классе G_k, отстоящих от центра «дальше», чем X).

- **Использование априорных данных:** Для вычисления апостериорных значений необходимо задать априорные вероятности (предположения о частоте классов до начала исследования). Они могут быть:
  1. **Равными** для всех групп.
  2. **Пропорциональными** фактическим размерам групп в обучающей выборке.
- **Правило классификации:** Модель классифицирует объект в ту группу, для которой **апостериорная вероятность является максимальной**,.
- **Значимость:** Данный подход позволяет не просто распределить объекты, но и оценить степень уверенности в правильности классификации для каждого конкретного случая.

1.5.3. Простые классифицирующие функции

Фишер (1936) был первым, кто предположил, что классификация должна проводиться с помощью линейной комбинации дискриминантных переменных. Он предложил применять линейную комбинацию, которая максимизирует различия между классами, но минимизирует дисперсию внутри классов. Таким образом, для каждого класса определяется своя линейная комбинация, которая и называется простой классифицирующей функцией. Она имеет следующий вид:

$$h_k = b_{k0} + b_{k1}X_1 + b_{k2}X_2 + \cdots + b_{kp}X_p \tag{26}$$

где h_k – значение простой классифицирующей функции для класса k, а b_{ki} - коэффициенты, которые необходимо определить. Объект относится к классу, для которого значение h максимально. Коэффициенты классифицирующих функций определяются по формулам:

$$b_{ki} = (n - g)\sum_{j=1}^{p} u_{ij}X_{jk}, \tag{27}$$

где u_{ij} - элемент матрицы, обратной к внутригрупповой матрице W сумм попарных произведений. Постоянный член определяется с помощью выражения

$$b_{k0} = -0,5\sum_{j=1}^{p} b_{kj}X_{jk}. \tag{28}$$

Коэффициенты простых классифицирующих функций обычно не интерпретируются, так как они не стандартизованы и каждому классу соответствует своя функция. Точные значения функции роли не играют: необходимо знать лишь, для какого класса это значение наибольшее. Именно к нему объект ближе всего.

Суть классифицирующих функций

Классифицирующие функции представляют собой линейные комбинации независимых переменных (признаков), которые позволяют вычислить «показатель близости» объекта к каждой из групп. Для каждого класса j строится свое отдельное уравнение.

Основная формула функции для группы j:

$$S_j = c_j + w_{1j}x_1 + w_{2j}x_2 + \cdots + w_{pj}x_p$$

Где:

- S_j — результирующее значение (классификационный показатель) для группы j;
- c_j — константа для j-й группы;
- $w_{1j}, \ldots, w_{pj}$ — весовые коэффициенты признаков, рассчитанные для данной группы;
- $x_1, \ldots, x_p$ — фактические значения признаков классифицируемого объекта.

Применение для классификации

Процесс классификации с использованием этих функций выглядит следующим образом:

1. **Расчет показателей:** Значения признаков нового объекта (x_i) подставляются в уравнения функций для **каждой** из имеющихся групп.

2. **Сравнение результатов:** Для каждого объекта получается набор значений $S_1, S_2, \ldots, S_k$.
3. **Принятие решения:** Объект относится к той группе, для которой значение классифицирующей функции оказалось **максимальным**.

Математически это означает, что объект имеет наибольшую вероятность принадлежности именно к этой группе.

Особенности и условия применения

- **Оценка качества:** Эффективность разделения групп этими функциями проверяется с помощью статистики **Лямбда Уилкса (Λ)**. Значения Λ, близкие к 0, указывают на четкое различие между группами, а близкие к 1 — на плохую дискриминацию.
- **Допущения:** Метод предполагает, что независимые переменные имеют многомерное нормальное распределение, а ковариационные матрицы групп примерно равны.
- **Практическое использование:** В программных пакетах (например, Statistica) расчет коэффициентов w_{ij} и констант c_j происходит автоматически на основе обучающей выборки.