Лекция 7

ДИСКРИМИНАНТНЫЙ АНАЛИЗ

Коэффициенты канонической дискриминантной функции

Один из методов поиска наилучшей дискриминации данных заключается в нахождении такой канонической дискриминантной функции ***d***, которая бы максимизировала отношение межгрупповой вариации к внутригрупповой

$$\lambda = \mathbf{B}(d)/\mathbf{W}(d) \qquad (2)$$

где **B** - межгрупповая и **W** внутригрупповая матрицы рассеяния наблюдаемых переменных от средних.

В некоторых работах вместо **W** используют матрицу рассеяния **T** объединенных данных.

Коэффициенты канонической дискриминантной функции

Рассмотрим ***максимизацию отношения*** (2) ***для произвольного числа классов***.

Введем следующие обозначения:

g – число классов;

p – число дискриминантных переменных;

n_k – число наблюдений в ***k***-й группе;

n – общее число наблюдений по всем группам;

x_{ikm} – величина переменной ***i*** для ***m***-го наблюдения в ***k***-й группе;

$\bar{x}_{ik}$ – средняя величина переменной ***i*** в ***k***-й группе;

$\bar{x}_i$ – среднее значение переменной ***i*** по всем группам;

$T(u, v)$ – общая сумма перекрестных произведений для переменных ***u*** и ***v***;

$W(u, v)$ – внутригрупповая сумма перекрестных произведений для переменных ***u*** и ***v***.

$t_{ij} - T(x_i, x_j)$; $w_{ij} - W(x_i, x_j)$.

В модели дискриминации должны соблюдаться следующие условия:

1) число групп: $g \geq 2$;
2) число объектов в каждой группе: $n_i \geq 2$;
3) число дискриминантных переменных: $0 < p < (n - 2)$;
4) дискриминантные переменные измеряются в интервальной шкале;
5) дискриминантные переменные линейно независимы;
6) ковариационные матрицы групп примерно равны;
7) дискриминантные переменные в каждой группе подчиняются многомерному нормальному закону распределения.

Коэффициенты канонической дискриминантной функции

Рассмотрим задачу максимизации отношения (2) когда имеются g групп.

Оценим сначала информацию, характеризующую степень различия между объектами по всему пространству точек, определяемому переменными групп.

Для этого вычислим матрицу рассеяния **T**, которая равна сумме квадратов отклонений и попарных произведений наблюдений от общих средних $\bar{x}_i$, $i = 1, \ldots, p$ по каждой переменной.

Элементы матрицы **T** определяются выражением

$$t_{ij} = \sum_{k=1}^{g} \sum_{m=1}^{n} (x_{ikm} - \bar{x}_i)(x_{jkm} - \bar{x}_j) \qquad (3)$$

07.05.2020

Продолжение темы «Дискриминантный анализ»

Коэффициенты канонической дискриминантной функции

В (3): $$\bar{x}_i = (1/n)\sum_{k=1}^{g} n_i \bar{x}_{ik},\ i = 1,\ldots,p$$

$$\bar{x}_{ik} = (1/n_i)\sum_{m=1}^{n_k} x_{ikm},\ i = 1,\ldots,p;\ k = 1,\ldots,g$$

Запишем это выражение в матричной форме. Обозначим ***p***-мерную случайную векторную переменную ***k***-ой группы следующим образом

$$X_k = \{x_{ikm}\}\ \ i = 1,\ldots,p,\ \ k = 1,\ldots,g,\ \ m = 1,\ldots,n_k$$

Тогда объединенная ***p***-мерная случайная векторная переменная всех групп будет иметь вид

$$\mathbf{X} = [\mathbf{X}_1 \mathbf{X}_2 \ldots \mathbf{X}_g]$$

Коэффициенты канонической дискриминантной функции

Общее среднее этой *p*-мерной случайной векторной переменной будет равен вектору средних отдельных признаков

$$\bar{\mathbf{x}} = [\bar{x}_1 \bar{x}_2 \ldots \bar{x}_p]$$

Матрица рассеяния от среднего при этом запишется в виде

$$\mathbf{T} = \sum_{k=1}^{g} (\mathbf{X}_k - \bar{\mathbf{x}})(\mathbf{X}_k - \bar{\mathbf{x}})'$$

Если использовать векторную переменную объединенных переменных **X**, то матрица **T определится по формуле**

$$\mathbf{T} = (\mathbf{X} - \bar{\mathbf{x}})(\mathbf{X} - \bar{\mathbf{x}})'$$

Коэффициенты канонической дискриминантной функции

- Матрица **T** содержит полную информацию о распределении точек по пространству переменных.
- Диагональные элементы представляют собой сумму квадратов отклонений от общего среднего и показывают как ведут себя наблюдения по отдельно взятой переменной.
- Внедиагональные элементы равны сумме произведений отклонений по одной переменной на отклонения по другой.

- Если разделить матрицу **T** на (n – 1), то получим ковариационную матрицу.

- Для проверки условия линейной независимости переменных полезно рассмотреть вместо **T** нормированную корреляционную матрицу.

- Для измерения степени разброса объектов внутри групп рассмотрим матрицу **W**, которая отличается от **T** только тем, что ее элементы определяются векторами средних для отдельных групп, а не вектором средних для общих данных.

- Элементы внутригруппового рассеяния определятся выражением

$$w_{ij} = \sum_{k=1}^{g} \sum_{m=1}^{n_k} (x_{ikm} - \bar{x}_{ik})(x_{jkm} - \bar{x}_{jk})$$

Запишем это выражение в матричной форме.

Данным g групп будут соответствовать векторы средних

$$\bar{\mathbf{x}}_1 = [\bar{x}_{11} \bar{x}_{21} \ldots \bar{x}_{p1}],$$

$$\ldots$$

$$\bar{\mathbf{x}}_g = [\bar{x}_{1g} \bar{x}_{2g} \ldots \bar{x}_{pg}].$$

Тогда матрица внутригрупповых вариаций запишется в виде

$$\mathbf{W} = \sum_{k=1}^{g} (\mathbf{X}_k - \bar{\mathbf{x}}_k)(\mathbf{X}_k - \bar{\mathbf{x}}_k)'$$

Если разделить каждый элемент матрицы **W** на (***n***–g), то получим оценку ковариационной матрицы внутригрупповых данных.

- Когда центроиды различных групп совпадают, то элементы матриц **T** и **W** будут равны.

- Если же центроиды групп различные, то разница

$$\mathbf{B} = \mathbf{T} - \mathbf{W} \qquad (8)$$

будет определять межгрупповую сумму квадратов отклонений и попарных произведений.

- Если расположение групп в пространстве различается (т.е. их центроиды не совпадают), то степень разброса наблюдений внутри групп будет меньше межгруппового разброса.

- Элементы матрицы **B** можно вычислить и по данным средних

$$b_{ij} = \sum_{k=1}^{g} n_k (\bar{x}_{ik} - \bar{x}_i)(\bar{x}_{jk} - \bar{x}_j), \; i, j = 1, \ldots, p \qquad (9)$$

- Матрицы **W** и **B** содержат всю основную информацию о зависимости внутри групп и между группами.
- ***Для лучшего разделения наблюдений на группы*** нужно подобрать коэффициенты дискриминантной функции из условия максимизации отношения межгрупповой матрицы рассеяния к внутригрупповой матрице рассеяния при условии ортогональности дискриминантных плоскостей.
- Тогда нахождение коэффициентов дискриминантных функций сводится к решению задачи о собственных значениях и векторах.

- **Это утверждение можно сформулировать так**: если спроектировать g групп ***p***-мерных выборок на (g – 1) пространство, порожденное собственными векторами

$$(v_{1k}, \dots, v_{pk}), k = 1, \dots, g-1,$$

то отношение (2) будет максимальным, т.е. ***рассеивание между группами будет максимальным при заданном внутригрупповом рассеивании***.

- Если бы мы захотели спроектировать g выборок на прямую при условии максимизации наибольшего рассеивания между группами, то следовало бы использовать собственный вектор $(v_{11}, \dots, v_{1k})$ соответствующий максимальному собственному числу λ_1.

При этом дискриминантные функции можно получать: по ***нестандартизованным*** и ***стандартизованным коэффициентам.***

Нестандартизованные коэффициенты

Пусть $\lambda_1 \geq \ldots \geq \lambda_p$ и $\mathbf{v}_1,\ldots,\mathbf{v}_p$ соответственно собственные значения и векторы. Тогда условие (2) в терминах собственных чисел и векторов запишется в виде

$$\lambda = \frac{\sum_k b_{jk} v_j v_k}{\sum_k w_{jk} v_j v_k}$$

что влечет равенство $\sum_k (b_{jk} - \lambda w_{jk}) v_k = 0$

или в матричной записи $(\mathbf{B} - \lambda_i \mathbf{W})\mathbf{v}_i = 0, \quad \mathbf{v}_i' \mathbf{W} \mathbf{v}_j = \delta_{ij}$ (10)

где δ_{ij} символ Кронекера.

Таким образом, решение уравнения $|\,\mathbf{B} - \lambda\mathbf{W}\,| = 0$ позволяет нам определить компоненты собственных векторов, соответствующих дискриминантным функциям.

Нестандартизованные коэффициенты

Если **B** и **W** невырожденные матрицы, то собственные корни уравнения

$$|\mathbf{B}-\lambda\mathbf{W}|=0 \quad \text{такие же, как и у} \quad |\mathbf{W}^{-1}\mathbf{B}-\lambda\mathbf{I}|=0$$

Решение системы уравнений (10) можно получить путем использования *разложения Холецкого* **LL′** матрицы $\mathbf{W}^{-1}$ и решения задачи о собственных значениях

$$(\mathbf{L}'\mathbf{B}\mathbf{L}-\lambda_i\mathbf{I})\mathbf{v}_i=0, \quad \mathbf{v}_i'\mathbf{v}_j=\delta_{ij}$$

- Каждое решение, которое имеет свое собственное значение λ_i и собственный вектор $\mathbf{v}_i$, соответствует одной дискриминантной функции.
- Компоненты собственного вектора $\mathbf{v}_i$ можно использовать в качестве коэффициентов дискриминантной функции.
- Однако при таком подходе начало координат не будет совпадать с главным центроидом.

Нестандартизованные коэффициенты

Для того, чтобы начало координат совпало с главным центроидом нужно нормировать компоненты собственного вектора

$$\boldsymbol{\beta}_i = \mathbf{v}_i \sqrt{n-g}, \quad \beta_0 = -\sum_{i=1}^{p} \beta_i \bar{x}_i \qquad (11)$$

- Нормированные коэффициенты (11) получены по нестандартизованным исходным данным, поэтому они называются ***нестандартизованными.***
- Нормированные коэффициенты приводят к таким дискриминантным значениям, единицей измерения которых является стандартное квадратичное отклонение.
- При таком подходе каждая ось в преобразованном пространстве сжимается или растягивается таким образом, что соответствующее дискриминантное значение для данного объекта представляет собой число стандартных отклонений точки от главного центроида.

Стандартизованные коэффициенты

можно получить двумя способами:

1) по формуле (11), если исходные данные были приведены к стандартной форме;
2) преобразованием нестандартизованных коэффициентов к стандартизованной форме:

$$\mathbf{c}_i = \boldsymbol{\beta}_i \sqrt{\frac{w_{ii}}{n-g}} \qquad (12)$$

где w_{ii} – сумма внутригрупповых квадратов i-й переменной, определяемой по формуле (5).

- Стандартизованные коэффициенты полезно применять для уменьшения размерности исходного признакового пространства переменных.
- Если абсолютная величина коэффициента для данной переменной для всех дискриминантных функций мала, то эту переменную можно исключить, тем самым сократив число переменных.

Структурные коэффициенты

определяются коэффициентами взаимной корреляции между отдельными переменными и дискриминантной функцией. Если относительно некоторой переменной абсолютная величина коэффициента велика, то вся информация о дискриминантной функции заключена в этой переменной.

Структурные коэффициенты полезны при классификации групп.

Структурный коэффициент можно вычислить и для переменной в пределах отдельно взятой группы.

Тогда получаем ***внутригрупповой структурный коэффициент***, который вычисляется по формуле:

$$s_{ij} = \sum_{k=1}^{p} r_{ik} c_{kj} = \sum_{k=1}^{p} \frac{w_{ik} c_{kj}}{\sqrt{w_{ii} w_{jj}}}$$

где S_{ij} – внутригрупповой структурный коэффициент для ***i***-ой переменной и ***j***-ой функции; r_{ik} – внутригрупповые структурные коэффициенты корреляции между переменными ***i*** и ***k***; c_{kj} – стандартизованные коэффициенты канонической функции для переменной ***k*** и функции ***j***.

Структурные и стандартизованные коэффициенты

- *Структурные коэффициенты* по своей информативности несколько отличаются от стандартизованных коэффициентов.
- *Стандартизованные коэффициенты* показывают вклад переменных в значение дискриминантной функции. Если две переменные сильно коррелированы, то их стандартизованные коэффициенты могут быть меньше по сравнению с теми случаями, когда используется только одна из этих переменных.
- Такое распределение величины стандартизованного коэффициента объясняется тем, что при их вычислении учитывается влияние всех переменных.
- Структурные же коэффициенты являются парными корреляциями и на них не влияют взаимные зависимости прочих переменных.

Коэффициент канонической корреляции

Другой характеристикой, позволяющей оценить полезность дискриминантной функции является ***коэффициент канонической корреляции*** r_i.

Каноническая корреляция является мерой связи между двумя множествами переменных. Максимальная величина этого коэффициента равна 1.

Будем считать, что группы составляют одно множество, а другое множество образуют дискриминантные переменные.

Коэффициент канонической корреляции для *i*-ой дискриминантной функции определяется формулой:

$$r_i = \sqrt{\frac{\lambda_i}{1+\lambda_i}} \qquad (14)$$

Остаточная дискриминация

- Так как дискриминантные функции находятся по выборочным данным, они нуждаются в *проверке статистической значимости*.
- Дискриминантные функции представляются аналогично главным компонентам. Поэтому для проверки этой значимости можно воспользоваться *критерием*, аналогичным дисперсионному критерию в методе главных компонент.
- Этот критерий оценивает *остаточную дискриминантную способность*, под которой понимается способность различать группы, если при этом исключить информацию, полученную с помощью ранее вычисленных функций.
- Если остаточная дискриминация мала, то не имеет смысла дальнейшее вычисление очередной дискриминантной функции.

Остаточная дискриминация

Полученная статистика носит название «Λ-*статистики Уилкса*» и вычисляется по формуле:

$$\Lambda = \prod_{i=k+1}^{g} (1/(1+\lambda_i)) \qquad (15)$$

где $\boldsymbol{k}$ – число вычисленных функций.

Чем меньше эта статистика, тем значимее соответствующая дискриминантная функция.

Остаточная дискриминация

Величина

$$\chi^2 = -[n - ((p+g)/2) - 1]\ln \Lambda_k, \quad k = 0, 1, \ldots, g-1$$

имеет хи–квадрат распределение с $(p - k)(g - k - 1)$ степенями свободы.

Вычисления проводятся в следующем порядке:

1. Находим значение критерия χ^2 при $\boldsymbol{k}=0$. Значимость критерия подтверждает существование различий между группами. Кроме того, это доказывает, что первая дискриминантная функция значима и имеет смысл ее вычислять.
2. Определяем первую дискриминантную функцию и проверяем значимость критерия при $\boldsymbol{k}=1$. Если критерий значим, то вычисляем вторую дискриминантную функцию и продолжаем процесс до тех пор, пока не будет исчерпана вся значимая формация.

Классифицирующие функции

- Ранее было рассмотрено получение канонических дискриминантных функций при известной принадлежности объектов к тому или иному классу.
- Основное внимание уделялось ***определению числа и значимости этих функций, и использованию их для объяснения различий между классами***. Все сказанное относилось к интерпретации результатов ДА.
- Однако наибольший интерес представляет ***задача предсказания класса, которому принадлежит некоторый случайно выбранный объект***.
- Эту задачу можно решить, используя информацию, содержащуюся в дискриминантных переменных. Существуют различные способы классификации.

Классифицирующие функции

- В процедурах классификации могут использоваться как сами дискриминантные переменные, так и канонические дискриминантные функции.
- В первом случае применяется *метод максимизации различий между классами* для получения функции классификации, различие же классов на значимость не проверяется и, следовательно, дискриминантный анализ не проводится.
- Во втором случае для классификации используются непосредственно дискриминантные функции и проводится более глубокий анализ.

Контрольные вопросы

Вариант 1

1. Разведочный анализ данных: предпосылки, суть и цели.

2. Количественный анализ – суть и предназначение. Фазы количественной обработки данных.

Вариант 2

1. Принцип «мягких вычислений» - суть и значение для ИАД.

2. Качественный анализ - суть и предназначение

Вариант 3

1. Data Mining: определение понятия и требования к знаниям.

2. Понятие корреляционно-регрессионного анализа данных. Этапы КРА.

Вариант 4

1. Стадии ИАД.

2. Дисперсионный анализ. Постановка задачи.

Вариант 5

1. Основные задачи Data Minig.

2. Однофакторный дисперсионный анализ. Основные этапы и соотношения.