

Хохлов Юрий Степанович

yskhokhlov@yandex.ru

Многомерный фиксированный статистический анализ.

ЛЕКЦИЯ 1

Введение ~~Тема~~

Опр: Вероятностное пр-во

Опр: Случайная величина и ее распределение

Опр: Случайный вектор и его распределение, матем. ожидание, дисперсия, ковариация, коэфф. корреляции и св-ва.

Глава 1 Тисслертово пр-во случайных величин

Тисслертово пр-во фиксированное пр-во: $\Omega, \mathcal{A}, P$

$\mathcal{L}_2$ есть пр-во случайных величин ξ :

$$E(|\xi|^2) < \infty$$

Задача: $\mathcal{L}_2$ -линейное пр-во

Р: (если * не число, вынесется из $\mathcal{L}_2$)

(+ из пер-ва Коши-Бунковского)

Покажем, что в нем можно ввести скалярный произв.

Определим на L_2 среднюю функционал

$$\forall \xi_1, \xi_2 \in L_2 \quad (\xi_1, \xi_2) := E(\xi_1 \xi_2) \quad (1)$$

Задача: Определенный по ф-ле (1) функционал, обладает следующими св-вами: $(\xi_1, \xi_2) \in L_2$

$$① (\xi_1, \xi_1) \geq 0, \quad (\xi_1, \xi_1) = 0 \Leftrightarrow \xi_1 \stackrel{\text{почти наверное}}{=} 0$$

$$② (\xi_1, \xi_2) = (\xi_2, \xi_1) \text{ (симметрично) (с вер-ю 1)}$$

$$③ (c_1 \xi_1 + c_2 \xi_2, \xi_3) = c_1 (\xi_1, \xi_3) + c_2 (\xi_2, \xi_3) \quad \left. \begin{array}{l} \text{(св-во линейности)} \end{array} \right\} \Rightarrow$$

$\Rightarrow (\xi_1, \xi_2)$ - скалярное произв в L_2

Опр: Нормой случайной величины $\xi \in L_2$ называется число $\|\xi\|$, по опр $= \sqrt{(\xi, \xi)} = \sqrt{E(|\xi|^2)}$ (2)

Опр: Послед-ть случайных величин $\xi_n \in L_2$ с.с. в среднем квадратическом (с.к.)

к случайной величине ξ_0 , если

$$\|\xi_n - \xi_0\|^2 = E(|\xi_n - \xi_0|^2) \rightarrow 0 \quad n \rightarrow \infty \quad (3)$$

Замечание: С функционал. анале докажем, что L_2 полно относительно сходимости в с.к., т.е. если

$$E(|\xi_n - \xi_0|^2) \rightarrow 0, \quad n \rightarrow \infty, \quad m \rightarrow \infty, \quad \xi_1, \dots, \xi_n \in L_2$$

$E(|\xi_n - \xi_0|^2) \rightarrow 0 \quad n \rightarrow \infty \Rightarrow L_2$ - гильберт
 (линейное пр-во со
 каждой прецв порожд
 сходим отн этого скет прецв,
 и если оно полное $\rightarrow$ гильберт)

$\|\xi\|$ - анализ элемент вектора

Из линейной алгебры мне знаем,

это для $\xi_1, \xi_2 \in L_2$ $\cos(\xi_1, \xi_2) = \frac{(\xi_1, \xi_2)}{\|\xi_1\| \|\xi_2\|}$ по опр

$$\cos(\xi_1, \xi_2) = \frac{E(\xi_1 \cdot \xi_2)}{\sqrt{E(\xi_1^2) E(\xi_2^2)}} \quad (4) \quad \Leftrightarrow f(\xi_1, \xi_2), \text{ если } E(\xi_1) = E(\xi_2) = 0$$

$\Rightarrow$ Опр: $\xi_1, \xi_2 \in L_2$ - ортогона ($\perp$)

$$(\xi_1, \xi_2) = 0 = E(\xi_1 \cdot \xi_2) \quad (5)$$

$\Leftrightarrow \xi_1$ и ξ_2 некоррелированы (при уде $E(\xi_1) = E(\xi_2) = 0$)

В наших пр ва L_2 рассмотрим ел

задачу: $\exists H \in L_2$ замкнутое линейное

подпр-во, $\eta \in L_2$

Опр: Существование $\hat{\eta}$ - наилучшее
 линейное приближение случайной

величины η в H , если

1) $\hat{\eta} \in H$

$$\|\eta - \hat{\eta}\|^2 \quad (6)$$

2) $\forall \xi \in H \quad \|\eta - \hat{\eta}\|^2 = E(|\eta - \hat{\eta}|^2) \leq E(|\eta - \xi|^2) \quad 3$

Введем к линейному гр-10



Лемма о перпендикуляре:

$\hat{z}$ есть ортогональное проецирование z

в $H \Leftrightarrow 1) \hat{z} \in H$

$$2) \forall \xi \in H \Rightarrow (z - \hat{z}, \xi) = E((z - \hat{z})\xi) = 0$$

(7) на самом деле СЛАУ

(7)

Пример: $z, \xi \in L_2$

Рассмотрим пр-во H линейных
векторов следующего вида:

$$c_0 + c_1 \xi = c_0 \cdot 1 + c_1 \cdot \xi \Rightarrow H - \text{одномерное линейное пр-во, порожд. вектор } \xi_1 = 1 \text{ и } \xi_2 = \xi$$

В силу леммы о перпендикуляре

$$\hat{z} = \hat{c}_0 + \hat{c}_1 \xi$$

и (т.к. вектор $\perp$ пр-во $\Rightarrow \perp$ основе)

$$\left. \begin{aligned} (z - \hat{z}, \xi_1) &= 0 \\ (z - \hat{z}, \xi_2) &= 0 \end{aligned} \right\}$$

$$3) \text{ если } \begin{cases} \eta_1 = a_1 \xi_1 + b_1 \\ \eta_2 = a_2 \xi_2 + b_2 \end{cases} \Rightarrow \rho(\eta_1, \eta_2) = \rho(\xi_1, \xi_2) \operatorname{sign}(a_1 a_2) \quad (a_1 a_2 \neq 0)$$

$$4) \text{ если } \rho(\xi_1, \xi_2) = +1 \Rightarrow \exists a > 0 \text{ и } b \in \mathbb{R}_1 : \xi_2 = a \xi_1 + b$$

$$\text{Если } \rho(\xi_1, \xi_2) = -1 \Rightarrow \exists a < 0 \text{ и } b \in \mathbb{R}_1 : \xi_2 = a \xi_1 + b$$

Задача ξ_1 и $\xi_2 \in \mathcal{H}$ и $E(\xi_1) = E(\xi_2) = 0$

Найдем наилучшее линейное приближение

ξ_2 с помощью ξ_1

$$\hat{\xi}_2 = \hat{C} \xi_1, \text{ где } \hat{C} = \rho(\xi_1, \xi_2) \sqrt{\frac{D(\xi_2)}{D(\xi_1)}}$$

$e = \xi_2 - \hat{\xi}_2$ ортогонально к ξ_1

$$\Rightarrow \xi_2 = \hat{\xi}_2 + e \text{ и } \hat{\xi}_2 \perp e \Rightarrow e \text{ и } \hat{\xi}_2 \text{ не коррелирует} \Rightarrow$$

$$\Rightarrow D(\xi_2) = D(\hat{\xi}_2) + D(e) \quad (*)$$

$$D(\hat{\xi}_2) = \rho^2(\xi_1, \xi_2) \frac{D(\xi_2)}{D(\xi_1)} =$$

$$= \rho^2(\xi_1, \xi_2) D(\xi_2)$$

$\Downarrow$

$$\boxed{D(\hat{\xi}_2) = \rho^2(\xi_1, \xi_2) D(\xi_2)} \quad (**)$$

(дисперсия — степень изменчивости в выборке)

ξ_2 ~~то~~ $(*)$ показывает, что $(\rho(\xi_1, \xi_2))^2$
измеряет долю целесообразности
 ξ_2 , которую можно объяснить
линейными внешними ξ_1

$$** \} \Rightarrow D(e) = (1 - |\rho(\xi_1, \xi_2)|^2) D(\xi_2)$$

$1 - |\rho(\xi_1, \xi_2)|^2$ - измеряет ту часть
целесообразности ξ_2 котор
не удается объяснить линейными
внешними ξ_1 , и необходимо привести
дополн. факторы

Все сводят в квадрат коэффициент корреляции
Замечания

1) Если $\rho(\xi_1, \xi_2) > 0 \Rightarrow \xi_1$ и ξ_2 - положительно
(если ξ_1 растет $\rightarrow \xi_2$ растет) коррелированы

2) Если $\rho(\xi_1, \xi_2) < 0 \Rightarrow \xi_1$ и ξ_2 - отрицательно
коррелированы

(ξ_1 растет $\rightarrow$ в среднем ξ_2 убывает)

- мало зависит $0 \leq |\rho(\xi_1, \xi_2)| < 0.5$ ($(0.5)^2 \approx 0.25$)
- средне зависит $0.5 \leq |\rho(\xi_1, \xi_2)| < 0.7$ ($(0.7)^2 \approx 0.5$)
- много зависит $0.7 \leq |\rho(\xi_1, \xi_2)|$

09.09, 2015

Лекция 2

Задача: Есть ли связь между
с.в. Y и X_0 ?

С.в. (X, Y) имеет совместный нормальный
распределение X и Y - независимы $\Leftrightarrow \rho = \rho(X, Y) = 0$

Проверим гипотезу

$H_0: \rho = 0$

против

$H_1: \rho \neq 0$

$(X_1, Y_1), \dots, (X_n, Y_n)$ есть повторная
выборка из совместного нормального
распределения

Для проверки H_0 против H_1 будем
использовать выборочный коэффициент корреляции.

$$R = \frac{S_{xy}}{S_x S_y}$$

$$\text{где } S_{xy} = \frac{1}{N} \sum_{j=1}^N (x_j - \bar{x})(y_j - \bar{y})$$

$$S_x^2 = \frac{1}{N} \sum_{j=1}^N (x_j - \bar{x})^2$$

$$S_y^2 = \frac{1}{N} \sum_{j=1}^N (y_j - \bar{y})^2$$

Рассмотрим новую с.в.:

$$T = \frac{R}{\sqrt{1-R^2}} \sqrt{N-2}$$

$R \in [-1, 1] \Rightarrow T \in \mathbb{R}$
зависит от того, какой

$\Rightarrow T$

[T:] Если верно $H_0 (p=0) \Rightarrow$ с. в. T
имеет распределение Стьюдента
с $N-2$ степенями свободы

Описание процедуры проверки H_0 vs H_1

- 1) Для заданного малого α (0.1, 0.05, 0.01...)
 (уровень значимости)
по таблицам распредел. Стьюдента
находим критическую константу $t(\alpha) > 0$;
 $P(|T| > t(\alpha) | H_0) = \alpha$
- 2) Если реально полученное значение
 $T_{\text{исп. стат.}}$ статистики T таково, что
 $|T_n| > t(\alpha) \Rightarrow$ отвергаем H_0
- 3) В противном случае, говорим, что
 H_0 не противоречит экспериментальным
данным

Замечание! В пакетах прикладных программ
по статист. кон. группой метод
проверки гипотез (т.к. был разработан)
(прикладной метод)

T полученное значение T_n стат. T .

По таблицам распредел. Стьюдента

вычисляем $P(|T| > |T_n| | H_0) = p$

p -достигнувший уровень значимости
(p -value)

Далее сравниваем это p с
предложенным уровнем значимости

Если $p < \alpha \Rightarrow$ отвергаем H_0

$p \geq \alpha \Rightarrow H_0$ не против экспер данных

Стандартный метод проверки гипотез:
либо 1 либо 2 крит.

Пример: Изучается годовое потребление (Y)
(~~до~~ после налогов)
в зависимости от располагаемого дохода
(X).

1958-1967: (в ценах 1958г.)

$$R^2 = 0.994$$

$$N = 10, N - 2 = 8$$

Отсюда мы видим, что потребление
 Y на 99,4% объяснено линейным
внешним X .

$$T_{\text{н}} = 37.8 \Rightarrow p\text{-value} = 0.0000$$

$\Rightarrow H_0$ отвергается на всех разумных

уровнях значимости $\Rightarrow$ обнаружен сильный
положительный связь 10

? Это значит, если X и Y имеют не
зависимый нормальный распределение?

При больших N , статистика
при больших N имеет
асимптот. нормал. распредел.

Предположим $\rho \neq 0$, мы хотим
получить его оценку.

Для этого строится доверительный
интервал.

Р. Фишер предложил такую статистику

$$Z = \frac{1}{2} \ln \frac{1+\rho}{1-\rho}$$

Она имеет асимпт. норм. распредел.
со средним

$$\frac{1}{2} \ln \frac{1+\rho}{1-\rho} + \frac{\rho}{2(N-1)} = a$$

и дисперсией $\frac{1}{N-3}$

В этом случае для a по Z можно
построить доверительный интервал
заданной надежности

$a(\rho)$ является строго монотонно
на $(-\infty; \infty)$

§2. Множественный коэффициент корреляции.

(В простом случае Y и 1 фактор, все множественный, в множ-м случае)

Пытаемся объяснить поведение Y с помощью нек-го набора факторов $X_1, X_2, \dots, X_m$
 $m \geq 2$

$\hat{Y} = \alpha + \beta_1 X_1 + \dots + \beta_m X_m$ - наймем для приближения Y .

Опр.: Множеств. коэф. корреляции между величиной Y и набором $X_1, \dots, X_m$ назыв. число

$$R_{Y, X_1, \dots, X_m} = R(Y, \hat{Y})$$

По аналогии с корреляцией, получаем:

$$D(\hat{Y}) = R_{Y, X_1, \dots, X_m}^2 D(Y) \quad (1)$$

$$D(e) = (1 - R_{Y, X_1, \dots, X_m}^2) D(Y) \quad (2)$$

где $e = Y - \hat{Y}$ - ошибка аппрокс.

(1) $R_{Y, X_1, \dots, X_m}^2$ показывает, какую долю дисперсии

Y можно объяснить линейными влияниями выбранных факторов

(2) $(1 - R_{Y, X_1, \dots, X_m}^2)$ - то, что обусловлено неучтенными факторами

§3 Частный коэффициент корреляции

Изучаем зависимость Y от факторов

$X_1, \dots, X_m$

Введем некоторый фактор $X_k \in X_1, \dots, X_m$

S - набор всех ост факторов.

Y^c - наилучшее линейное приближ
 Y через S .

X_k^c - наилучш лнй прибли X_k через S

Введем

$$\tilde{Y} = Y - Y^c, \quad \tilde{X}_k = X_k - X_k^c$$

Опр: Частичный коэф корреляции

Y и X_k , когда устранено влияние
всех остальных факторов, введенных

$$R_{YX_k \cdot S} = \rho(\tilde{Y}, \tilde{X}_k)$$

Св-ва частного коэф коррел:

1) $|R_{YX_k \cdot S}| \leq 1$

2) $1 - R_{Y, X_1, \dots, X_k}^2 = (1 - R_{Y, X_1, \dots, X_{k-1}}^2) (1 - R_{YX_k \cdot X_1, \dots, X_{k-1}}^2)$

$R_{YX_k \cdot X_1, \dots, X_{k-1}}^2$ покаж какую долю необъясн

застершим удалось объяснить введением
еще одного фактора

Частный коэф корр измеряет
чистое влияние этого
фактора на величину Y

Когда факторов много прост к.к.
может давать неверную картину
о том, как фактор влияет на с.в.

Пример:

Изучается качество производ ткани
 $Y = X_0$ в зависимости от числа
профилактических проверок станков X_1
и числа обрывов нити X_2 .

Провели обследов 37 предприятий
и получили след результаты:
(оценки коэф корр.)

$$R_{01} = 0.105$$

$$R_{02} = -0.024$$

$$R_{12} = -0.996$$

Формально корр между X_1 и X_2 не велика,
на качество работы, что не похоже на правду
практически.

Вычисляем $R_{01.2}$ и $R_{02.1}$

$$R_{01.2} = 0.307$$

$$R_{02.1} = -0.906$$

коэф. корр: есть
ли связь между
велич и направл она
велика?

Факторы действ в против сторону
и маскируют друг друга

16.09.2015

ЛЕКЦИЯ 3

Глава 3 Многомерное нормальное распределение§1 Определение и числовые характеристики

Опр: Случайный вектор $\xi = (\xi_1, \dots, \xi_n)^T$ имеет многомерное нормальное распределение, если его плотность имеет вид

$$\begin{aligned} p_{\xi}(x) &= p_{\xi}(x_1, \dots, x_n) = \\ &= \left(\frac{1}{2\pi} \right)^{\frac{n}{2}} \det(A) \exp\left(-\frac{1}{2} (A(x-m), (x-m))\right) = \\ &= \left(\frac{1}{2\pi} \right)^{\frac{n}{2}} \det(A) \exp\left(-\frac{1}{2} \sum_{p,q=1}^n a_{pq} (x_p - m_p) (x_q - m_q)\right) \end{aligned}$$

$$m = \begin{pmatrix} m_1 \\ \vdots \\ m_n \end{pmatrix} \quad A = (a_{pq}) - \text{симметричная положит. определ. матрица}$$

$$\text{Обозначим } \Sigma = (b_{pq}) = A^{-1}$$

Обозначение: $\xi \sim N(m, \Sigma)$ - ξ имеет многомерное нормальное распределение с параметрами m и Σ .

Тогда $\varphi_{\xi}(t) = \varphi_{\xi}(t_1, \dots, t_n)$ - характеристическая функция случайного вектора ξ , т.е.

$$\varphi_{\xi}(t) := E[e^{i(t, \xi)}]$$

15

(по характ. функции можно составить моменты,

Можно показать, что

$$\varphi_{\xi}(t) = \exp \left\{ i(t, m) - \frac{1}{2}(\Sigma t, t) \right\}$$

[Т:] Если $\xi \sim N(m, \Sigma) \Rightarrow E(\xi_p) = m_p$,
 $\text{cov}(\xi_p, \xi_q) = \sigma_{pq}$ (m - вектор средних)

Опр: ξ имеет стандартное
нормальное центр. распред. в $\mathbb{R}^n$,
если $m_p = 0 \forall p$, $\Sigma = I_n$, т.е.
 $D(\xi_p) = 1$, $\text{cov}(\xi_p, \xi_q) = 0$, $p \neq q$

§2 Распределение, связанное с нормальным

Опр: Если $\xi \sim N(0, I_n)$, ^{т.е. центр. норм. распред.} то с.в.

$$\chi_n^2 := \xi_1^2 + \dots + \xi_n^2$$

имеет по определению

χ^2 -распределение с n степенями свободы

Опр: $\xi_1, \xi_2, \dots, \xi_n, \xi_{n+1}$ - независимы и имеют
станд. норм. распред. Тогда

Т.2: $\frac{\xi_{n+1}}{\sqrt{\frac{1}{n}(\xi_1^2 + \dots + \xi_n^2)}}$ имеет по опред.

распределение Стьюдента (t -распред.)

Опр: $\xi_1, \dots, \xi_n, \xi_{n+1}, \dots, \xi_{n+m}$ - независимы
и имеют станд. норм. распред. $\Rightarrow$

с.в. $F = \left(\frac{1}{n}(\xi_1^2 + \dots + \xi_n^2) / \left(\frac{1}{m}(\xi_{n+1}^2 + \dots + \xi_{n+m}^2) \right) \right)$

имеет то определенное распределение
Средокора - Фишера с (n, m) степенями
свободы (F распределение)

Замечание: T^2 имеет распредел
Фишера с 1 и n степенями свободы

§3 Св-ва многомерного нормального
распределения (* Крамер - книга)
Боксу - ва

Л1 Если $C: \mathbb{R}^n \rightarrow \mathbb{R}^d$ - линейное отображ
(d - какое-то какое) и $\xi \sim N(m, \Sigma)$ в $\mathbb{R}^n$, то
не факт, что $d < n$
след. в. $\eta = C\xi \sim N(Cm, C\Sigma C^T)$

Следствие: ξ - имеет стандартное
нормальное распредел в $\mathbb{R}^n$, т.е. $\xi \sim N(0, I_n)$,
 C - ортогонал проекция $\mathbb{R}^n$ на $\mathbb{R}^d$, $d < n$

Тогда $\eta = C\xi$ имеет станд норм
распредел в $\mathbb{R}^d$

$\xi = \begin{pmatrix} \xi' \\ \xi'' \end{pmatrix}$ векторы Тогда $m = \begin{pmatrix} m' \\ m'' \end{pmatrix}$ $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$

Т3: Если $\xi \sim N(m, \Sigma)$, то $\xi' \sim N(m', \Sigma_{11})$

2) ξ' и ξ'' независимы $\Leftrightarrow \Sigma_{12} = \Sigma_{21}^T = 0$
(для данного распредел ~~не~~ независимость
и некоррелир. совпадают)

[T₄]: Если $\xi \sim N(m, \Sigma)$, то условное
распределение ξ'' при условии: $\xi' = x'$, есть
многомерный нормальный распредел с вектором
средних $m'' + \Sigma_{21} \Sigma_{11}^{-1} (x' - m')$
и матрицей ковариаций

$$\Sigma_{22 \cdot 1} = \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}$$

при усл, что это фиксир

[T₅]: Если $\xi \sim N(m, \Sigma)$, то с.в. $Q = (\xi - m)^T \Sigma^{-1} (\xi - m)$
имеет χ^2 распределение с n степеней
свободы

§4 Оценки параметров

m, Σ - параметры многомерного
нормального распределения

[1] $X_1, X_2, \dots, X_N$ есть повторная выборка
из многомерного нормального распределения,
то есть 1) $X_1, \dots, X_N$ - независимы,

2) X_j имеет многомерный
нормальный распредел с пар m и Σ
($X_j \sim N(m, \Sigma)$)

$$X_j = \begin{pmatrix} X_{1j} \\ X_{2j} \\ \vdots \\ X_{mj} \end{pmatrix} \quad j = \overline{1, N}$$

$m_k = E(\xi_k) \Rightarrow X_{k1}, \dots, X_{kN}$ - повторная
выборка из одномерного
нормального распределения

$$\Rightarrow \hat{m}_k = \bar{X}_k := \frac{1}{N} \sum_{j=1}^N X_{kj} \leftarrow$$

- исчисляемая и состоятельная величина

$$\sigma_{pq} = \text{cov}(\xi_p, \xi_q) := E[(\xi_p - m_p)(\xi_q - m_q)]$$

$$S_{pq} := \frac{1}{N} \sum_{j=1}^N (X_{pj} - \bar{X}_p)(X_{qj} - \bar{X}_q) \leftarrow$$

- выборочная ковариация

$S = (S_{pq})$ - состоятельная величина для Σ , но исчисляемая

Замечание: Для компьютерной реализации лучше всего брать не $(N-1)$ $\Rightarrow$

$\Rightarrow \hat{S} = \left(\frac{N}{N-1} S_{pq} \right)$ - исчисляемая величина для Σ (т.к. мы используем дополнительный параметр)

[T:] С. вектор $\bar{X} = (\bar{X}_1, \dots, \bar{X}_n)^T$ и матрица S (или $\hat{S}$) независимы

§5 Проверка гипотез

н 1. Одномерные результаты

[T:] Если $X_1, \dots, X_n$ есть повторная выборка из одномерного нормального распределения с пар (a, σ^2) и

$$\bar{X} = \frac{1}{N} \sum_{j=1}^N X_j, \quad S_1^2 = \frac{1}{N-1} \sum_{j=1}^N (X_j - \bar{X})^2, \text{ то}$$

1) $\bar{X} \sim N(a, \sigma^2/N)$

2) $(N-1) S_1^2 / \sigma^2$ - имеет χ^2 распредел с $n-1$ степен свободы

3) $\bar{X}$ и S_1^2 независимы

4) слуг. вел $T = \frac{(\bar{X} - a)}{S_1} \sqrt{N}$

(T откл $\bar{X} - a$, a не $\in S_1 \Rightarrow T \sqrt{N}$, a не $\in T \sqrt{N}$)

имеет распределение Стьюдента, с $N-1$ степен своб.

Проверяемая гипотеза:

$H_0: a = a_0$

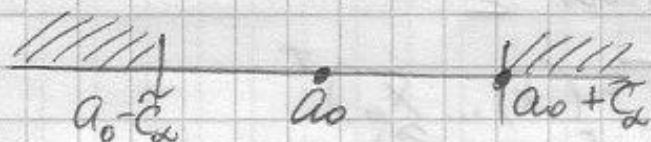
vs

$H_1: a \neq a_0$

В смысле формулы выше Теорема:

если верна H_0 , то с.в. $T = \frac{(\bar{X} - a_0)}{S_1} \sqrt{N}$

~~то~~ имеет распределение Стьюдента с $N-1$ ст.св.



Для заданного малого α находим по таблицам $t_{N-1}(\alpha/2) > 0$:

$$P(|T| > t_{N-1}(\alpha/2) | H_0) = \alpha$$

1) Если реально полученное значение T_n статистик $T: |T_n| > t_{N-1}(\alpha/2)$, то отвергаем H_0

2) В противном случае, говорим, что H_0 не противоречит эксперим. данным

Литература:

Глава 4. Основы задачи MCA§1 Описание задачи задачи

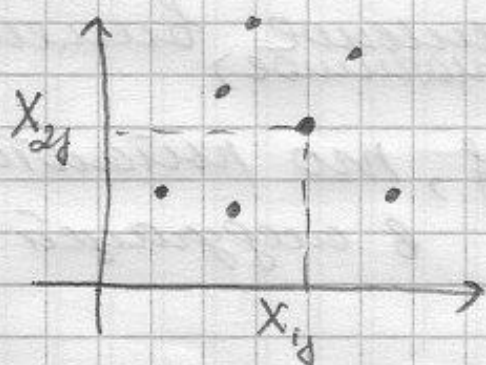
Тут мы имеем с.в. $\vec{X} = (x_1, x_2, \dots, x_d)^T$
 с вообще говоря неизвестными
 средними $m = (m_1, \dots, m_d)^T$ и матрицей
 ковариаций $\Sigma = (\sigma_{pq})^{\#}$

Имеем N независимых измерений
 объектов из изучаемой совокупности

$\vec{X}_1, \vec{X}_2, \dots, \vec{X}_N \Rightarrow X = (X_{kj})$ - матрица наблюдений

$k = \overline{1, d}$ $j = \overline{1, N}$. Обычно рассматривают
 эти измерения как точки в $\mathbb{R}^d$

$d = 2$: (двумерные измерения)



Удобно перейти к центрированным

измерениями. $\bar{X}_k := \frac{1}{N} \sum_{j=1}^N X_{kj}$

Получим $x_{kj} = X_{kj} - \bar{X}_k$,

$x = (x_{kj})$

удобно переходить к таким т.к. для
них ортогональность и коррелируемость -
одно и то же

§2 Основы задачи МСА.

П1: Корреляционный анализ

Изучается наличие и, если
присутствует, сила связи между
с.в. Для этого используется коэффц.
корреляции.

П2: Регрессионный анализ

Рассматривается объясняемая переменная
 y (назовем ее отклик) и несколько
(вероятно, 1) объясняющих факторов

$x_1, \dots, x_m$

Если обнаружено сильное внешнее
(значимое) влияние
факторов $x_1, \dots, x_m$ на y , то пытаются
найти вид их связи в следующей
форме:

$$Y = g(x_1, \dots, x_m) + \varepsilon, \text{ что-то еще, что выведет}$$

$g(x_1, \dots, x_m)$ - внешние факторы

ε - внешние неутраченные факторы

ПЗ: Методы снижения размерности

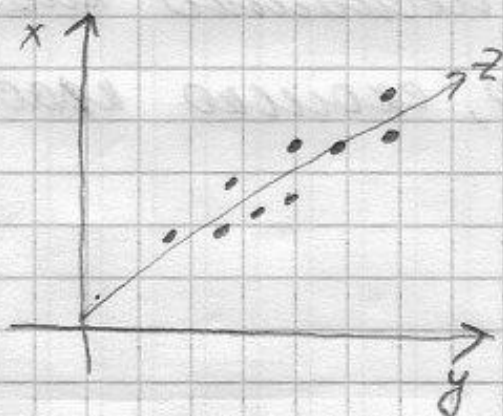
Важно размерность d -велика

Пытаются найти небольшое количество факторов

(не все тех, которые были, но и новых, выратт через старые),
~~факторы~~ которые достаточно хорошо
 представляют из себя в
 данных исходной совокупности.

методы факторного анализа,
 главные компоненты и т.д.,
 ≈ визуальный метод (мехол.)

Пример: 2-мерное пр-во, хотим 1



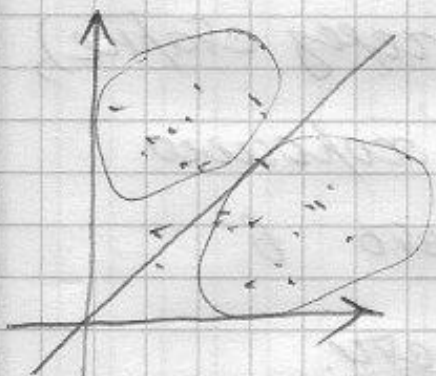
По x и y все
 не столько
 меняется
 как по z (2-мерное,
 не 3d) $\Rightarrow$
 достаточно
 оставить
 только z

174: Дискриминантный анализ

Предположим, что наши данные неоднородны. Например, они собраны из двух совокупностей с разными средними

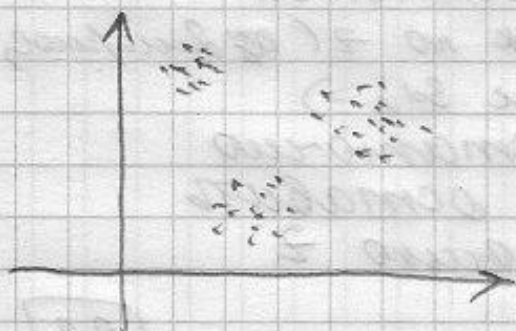
Основная задача

Найти процедуру (правило), позволяющую разделить все наблюдения по признаку принадлежности к одной из совокупностей



175: Кластерный анализ

Визуально или вводя, что наши данные как-то группируются в несколько классов заранее не известно, сколько классов.



Задача: Предположить некоторое
правильно обобщенная теория в 1 класс

Глава 5: Регрессионный анализ

§1 Постановка задачи

Имеем одну объясняемую
переменную Y и несколько объясняющих
 $X_1, \dots, X_d$.

Рассмотрим сл. представление

$$Y = f(x_1, \dots, x_d) + \varepsilon, \text{ где } f \in \mathcal{F} (\text{класс ф-ций})$$

Необходимо найти такую функцию

$g \in \mathcal{F}$, которая в опред смысле
наилучшим образом приближает
 Y с помощью факторов $x_1, \dots, x_d$

Определим наилуч. обр. приближает ~??:

Если расстояние между с.в.

измеряется в среднем квадрате,

то наилучшее приближение задается

по правилу:

$$g(x_1, \dots, x_d) = E(Y | x_1 = x_1, \dots, x_d = x_d) \quad (1)$$

тогда $g(x_1, \dots, x_d)$ — функция регрессии
с.в. Y на факторы $x_1, \dots, x_d$.

но вычислить $E(\text{май. от})$
относительно (распред. м.б.
нестривнально), поэтому:

Основная задача: по эксперим.
данным оценить функцию
регрессии.

§2 Классическая линейная модель регрессии.

Проводится N одновременных
измерений величины Y и факторов
 $X_1, \dots, X_d$. При этом предполагается,
что $Y_j = g(X_{j1}, \dots, X_{jd}) + \varepsilon_j$ (2)

Основные предположения:

- 1) Модель линейная по параметрам
то есть $Y_j = \alpha + \beta_1 X_{j1} + \dots + \beta_d X_{jd} + \varepsilon_j$ (3)
- 2) Факторы X_{jk} измерены точно
то есть это не случайные
величины
- 3) $E(\varepsilon_j) = 0$, $\forall j$, т.е. нет систематических
ошибок
- 4) Дисперсия ε_j : $D(\varepsilon_j) = \sigma^2$, $\forall j$
(дисперсия одинакова для всех j)
— условие гомосkedастичности

$$5) \text{cov}(\varepsilon_j, \varepsilon_k) = 0, j \neq k -$$

ошибки некоррелированы

6) ε_j - имеет нормаль распредел

Готово использовать следующую матричную обозначения:

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_N \end{pmatrix}, \vec{X}_k = \begin{pmatrix} X_{1k} \\ X_{2k} \\ \vdots \\ X_{Nk} \end{pmatrix}, \vec{X}_0 = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

$$\Theta = \begin{pmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_d \end{pmatrix} = \begin{pmatrix} \Theta_0 \\ \Theta_1 \\ \vdots \\ \Theta_d \end{pmatrix}$$

$$X = (X_{jk}) \quad j = \overline{1, N} \quad k = \overline{0, d}$$

$$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_N \end{pmatrix}$$

$$\Rightarrow Y = X\Theta + \varepsilon = \Theta_0 \vec{X}_0 + \Theta_1 \vec{X}_1 + \dots + \Theta_d \vec{X}_d + \varepsilon \quad (4)$$

т.е. $\Theta_0, \dots, \Theta_d$ и σ^2 - параметры модели

Из основных орг следует:

Y имеет многомерное нормальное распредел со средним $X\Theta$

и матрицей ковариаций $\Sigma_Y = \Sigma_\varepsilon = \sigma^2 I_N$

Основные задачи:

- 1) Оценка параметров
- 2) Проверка гипотез о параметрах
- 3) Выбор значимых факторов [27]

- 4) Проверка адекватности модели
 5) Проверка основных гарантий

* Будем использовать метод наименьших квадратов. Оценка в $\Delta_2 \Rightarrow$ Будем в Δ_2

$$Y = X\Theta + \varepsilon = \Theta_0 \vec{x}_0 + \dots + \Theta_m \vec{x}_m + \varepsilon(y)$$

 $X = (x_{jk})$ - матрица план эксперимента

Лекция 5

30.09.2015

§2 Оценка параметров. Метод наименьших квадратов (МНК)

$\Theta_0, \Theta_1, \dots, \Theta_m$ и $\sigma^2 = D(\varepsilon_j)$ - параметры модели

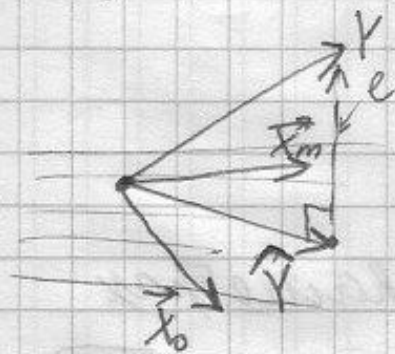
Для оценки параметров модели будем решать сл. экстр. задачу:

$$Q(\Theta) = \|Y - X\Theta\|^2 = \sum_{j=1}^n [Y_j - \Theta_0 - \Theta_1 x_{j1} - \dots - \Theta_m x_{jm}]^2 \rightarrow \min_{\Theta} \quad (5)$$

Имеем невырожденную ситуацию, если векторы $\vec{x}_0, \dots, \vec{x}_m$ линейно независимы (эквив. матрица X имеет ранг $m+1$)

Замечание: Фактически есть задача

- о наименьшей линейной оценке Y в линейном пр-ве, порожденном $\vec{x}_0, \dots, \vec{x}_m$



Задача на минимум решается с помощью необходимых условий экстремума:

$$\frac{\partial \Theta}{\partial \theta_k} = 0, \quad k = \overline{0, m} \quad (6)$$

После некоторых преобразований, получаем (5) подставляем в (6))

$$X^T \cdot X \cdot \Theta = X^T Y \quad (7)$$

(7) - система нормальных ур-ий

В невырожденном случае матрица $X^T X$ - невырожденная $\Rightarrow$

$$\Rightarrow \hat{\Theta} = (X^T X)^{-1} X^T Y \quad (8) -$$

- оценка пар Θ по МНК

$\hat{Y} = X \hat{\Theta}$ - вектор предсказанных значений

$e = Y - \hat{Y}$ - вектор остатков

$$\sigma^2 = D(\varepsilon_j) = E(\varepsilon_j^2) \quad (\text{среднее у } \varepsilon_j = 0)$$

Если бы ε_j были известны, то:

$$\frac{1}{N} \sum_{j=1}^N \varepsilon_j^2 \xrightarrow{P} E(\varepsilon_j^2) = \sigma^2$$

Но ε_j^2 не наблюдаемые величины

Замечание: остатки ε_j наследуют многие св-ва ε_j

Предлагается след оценка для σ^2 :

$$S^2 = \hat{\sigma}^2 := \frac{1}{N-(m+1)} \sum_{j=1}^N e_j^2 \quad (9)$$

← e_j — это ортогональные векторы $(x_0, \dots, x_m)$

Опр: Оценка $\hat{\beta}$ некоторого параметра β называется линейной (по Y), если $\exists$ такая неслучайная матрица A , что

$$\hat{\beta} = AY$$

Предложение 1: Оценка $\hat{\theta}$ параметра θ по МНК есть линейная оценка

Док-во: (8) $\Rightarrow A = (X^T X)^{-1} X^T$ т.т.г.

Предложение 2: $\hat{\theta}$ — несмещенная оценка для θ

Док-во: $\hat{\theta} = (X^T X)^{-1} X^T Y \stackrel{(4)}{=} (X^T X)^{-1} X^T (X\theta + \varepsilon) =$

$$= \theta + (X^T X)^{-1} X^T \varepsilon \quad (10)$$

$$\Rightarrow E(\hat{\theta}) = \theta + (X^T X)^{-1} X^T \overset{= \theta}{E(\varepsilon)} = \theta$$

т.т.г.

Предложение 3: $\exists$ выполнены условия 1-5) основы оценок. Тогда матрица ковариаций с.в. вектора θ имеет вид:

$$\Sigma_{\hat{\theta}} = \sigma^2 (X^T X)^{-1}$$

Д-во: $\hat{\theta} = (X^T X)^{-1} X^T Y = AY$

$$\Sigma_{\theta} = A \Sigma_Y A^T = (X^T X)^{-1} X^T G^2 X (X^T X)^{-1} = G^2 (X^T X)^{-1}$$

матрица
ковариаций
т.м.г

Предложение 4:

3 выполнимых усл 1)-5) основ. сбр,
 $N \rightarrow \infty$, при чем для элементов матрицы

$(X^T X)^{-1}$ стремится к нулю (а значит $D \rightarrow 0$)

Тогда $\hat{\theta}_N \xrightarrow{P} \theta, N \rightarrow \infty$

Д-во: Для доказательства достаточно
 применить неравенство Чебышева.

т.м.г

Лемма: Линейная оценка $\hat{\theta} = AY$
 является несмещенной оценкой (с)
 $AX = I$ (11)

Д-во
 возьмем Y , (4), посыл E , и будет $AX\hat{\theta} = \theta$.

т.м.г

Теорема: (Гаусса - Маркова):

Впр: Если выполнены условия 1)-5)
 основных ограничений, то оценка $\hat{\theta}$ по МНК
 является оптимальной в среднем квадрат
 в классе всех линейных и несмещ оценок,
 то есть $E[\|\hat{\theta} - \theta\|^2] \leq E[\|\tilde{\theta} - \theta\|^2]$

Док-во:

$$\hat{\theta} = (X^T X)^{-1} X^T Y = AY$$

$\hat{\theta} = BY$ - любая линейная оценка

$$B = A + C$$

В силу (11)

$$BX = AX = \pm 1 \Rightarrow CX = (XC)^T = 0$$

$$\Sigma_{\tilde{\Theta}} = G^2 \cdot B \cdot B^T = G^2 (A+C)(A+C)^T = G^2 (A+C)(A^T+C^T) =$$

$$= G^2 (AA^T + CA^T + AC^T + CC^T) \equiv$$

$$AC^T = (C \cdot A^T)^T \equiv$$

$$(XX^T)^T X^T C^T = (XX^T)^T (CX)^T = 0$$

$$\equiv 0$$

$$\equiv G^2 (AA^T + CC^T) = \Sigma_{\tilde{\Theta}} + G^2 CC^T \Rightarrow$$

$\Rightarrow \Sigma_{\tilde{\Theta}} - \Sigma_{\tilde{\Theta}} = G^2 CC^T$ - неотриц. опред. матрица,
а у нас на диаголь все число $\neq$ неотриц. $\Rightarrow$

$\Rightarrow$ а т.к. на диагональ различны
дисперсий $\tilde{\Theta}$ и $\tilde{\Theta}$ векторов $\Rightarrow$

$\Rightarrow$ дисперсия компон $\tilde{\Theta}$ не меньше
дисперсия соотв. комп $\tilde{\Theta}$

Теорема
Закрарне

Предложение 5:

Если выполнены условия ор 1)-6)
основных ограничений, то оценка S^2
вида (9) для G^2 является несмещ.
и состоятельной.

§3 Построение доверительных интервалов для параметров

Теорема $\hat{\theta}$ есть оценка θ по МНК
Тогда:

1) $\hat{\theta}$ имеет многомерный нормальный распределение со средним θ и ковариационной матрицей $\sigma^2 \cdot (X^T X)^{-1}$

2) с.в. $(\hat{\theta} - \theta)^T \cdot (X^T X) \cdot (\hat{\theta} - \theta) / \sigma^2$
- имеет χ^2 распределение с $m+1$ степенями свободы

3) с.в. $\sum_{j=1}^N (y_j - \hat{y}_j)^2 / \sigma^2 = \sum_{j=1}^N e_j^2 / \sigma^2$

имеет χ^2 распределение с $n - (m+1)$ степенями свободы

4) с.в. S^2 и $\hat{\theta}$ (или $\hat{y}$) независимы.
(доказ. т.к. e и $\hat{y}$ ортогональны, а т.к. распределение нормальное и независимы $\Rightarrow$ и функции от них независимы. $\hat{\theta}$ и $\hat{y}$ линейные функции e друг от друга линейными преобразованиями)

Вводящие b_{kk} - k -ый диагональный элемент из матрицы $(X^T X)^{-1}$

$\Rightarrow \hat{\theta}_k$ имеет одномерное нормальное 33

распределение со средним θ_k и дисперсией

$\sigma^2 \cdot b_{kk} \Rightarrow \frac{\hat{\theta}_k - \theta_k}{\sigma \sqrt{b_{kk}}}$ - имеет стандартную нормальную распределение

* $(N - (m+1)) S^2 / \sigma^2$ - имеет χ^2 распределение с $(N - (m+1))$ степеней свободы

$\Rightarrow$ по определению имеем, что:

$$\underline{T_k} = \frac{\hat{\theta}_k - \theta_k}{\sigma \sqrt{b_{kk}}} / \frac{S}{\sigma} = \frac{\hat{\theta}_k - \theta_k}{S \sqrt{b_{kk}}} = \frac{\hat{\theta}_k - \theta_k}{S_k}$$

- имеет распределение Стюдента с $(N - (m+1))$ степеней свободы

$\underline{S_k}$ - по определению стандартная ошибка при оценке θ_k .

Для заданного $f(x)$ по таблицам распределения находим константу

$$t_{N-(m+1)}(f) > 0: P(|T_k| < t_{N-(m+1)}) = f \Rightarrow$$

$$\Rightarrow \underline{\hat{\theta}_k - t_{N-(m+1)}(f) \cdot S_k} < \theta_k < \hat{\theta}_k + t_{N-(m+1)}(f) \cdot S_k \quad (12)$$

современный интервал
уровня f для параметра θ_k .

§5 Проверка значимости внешнего фактора

$$y = X\theta + \varepsilon$$

$$y_i = \theta_0 + \theta_1 x_{i1} + \dots + \theta_m x_{im} + \varepsilon_i \quad (1)$$

Фиксируем некоторое k :

$0 \leq k \leq m$ и проверим гипотезу

$$H_0: \theta_k = 0$$

vs

$$H_1: \theta_k \neq 0$$

— проверка о значимости внешнего фактора

Ранее было показано, что при верной H_0 статистика $T_k = \frac{\hat{Q}_k}{S_k} \quad (2)$

имеет распределение Стьюдента с $(n - (m+1))$ ст. свободы

Для заданного уровня значимости

$\alpha(0.1, 0.05, 0.01)$ находим по таблицам

критическую $t_{n-(m+1)}^*(\alpha) > 0$:

$$P(|T_k| > t_{n-(m+1)}^*(\alpha) | H_0) = \alpha$$

1) Если реально ненулевое значение статистики T_k не меньше $t_{n-(m+1)}^*(\alpha)$, то отвергаем H_0 , т.е. обнаружено значимое влияние этого фактора.

В противном случае гипотеза H_0 не противоречит экспериментальным данным, т.е. не обнаружено значимое влияние этого фактора

Такой метод решения задачи задан назыв T-критерий Стюдента

Замечания:

1. Если некоторый фактор призна незначимым, его удаляют из модели и ее оценивают заново

2. Этот критерий позволяет проверить значимость только 1 фактора, а не нескольких одновременно, т.к. задача решается, когда и других факторы вместе влияют на результат

Может быть ситуация, когда 1 фактор перебивает другой либо они тесно связаны.

§6 Проверка линейных гипотез

П1 Общая линейная гипотеза

Имеем линейную модель регрессии:

$$Y = X \cdot \theta + \varepsilon \quad (1) \text{ и вып. осн. пр. 1)-6)}$$

Рассмотрим проверку гипотезы

$$H_0: A\theta = a$$

$$\text{vs } H_1: A\theta \neq a \quad (2)$$

где $A - p \times (m+1)$ - несмещенная матрица
 $a - p$ - несмещенный вектор

$1 \leq p \leq m+1$ и $\text{rank}(A) = p$

Описание процедуры проверки H_0 :

1) Оцениваем модель (1) без учета ограничений и находим сумму квадратов остатков ESS_{ur}
без ограничений

2) Оцениваем модель (1) с учетом ограничений (2) и вновь находим сумму квадратов остатков ESS_r
(с оп.)

3) При верной H_0 , с.в. ESS_{ur}/σ^2 и $(ESS_r - ESS_{ur})/\sigma^2$ независимы и имеют χ^2 распределение: первая с $(N - (m+1))$ и p степеней свободы

4) Нужно учесть степени св. при оценке дисперсии σ^2 .
От этого зависит «кол-во»

При верной H_0 с.в.

$F = \frac{(ESS_r - ESS_{ur})/p}{ESS_{ur}/(N - (m+1))}$ имеет распределение

Следствие Фишера с $(p, N - (m+1))$ ст. своб.

5) При заданном α найти $F(\alpha) > 0$!

$$P(F > F_{\alpha}) = 1$$

6) Если равно наблюдаемое значение статистики $F > F_{\alpha} \Rightarrow$ отвергаем H_0

(отвергаем)

7) В противном случае H_0 не противоречит экспериментальным данным.

38

17.2 Проверка значимости внешнего фактора

Проверяется

$$H_0: D_k = 0$$

против

$$H_1: D_k \neq 0$$

- частный случай задачи (2)

F-критерий для проб знач.

~~Важно~~ замечание 1

F-критерий оценивает чистое внешнее фактора, когда устранено внешнее всех остальных.

T-критерий проверяет значимость внешнего фактора в присутствии всех остальных.

П3: Проверка внешних групповых факторов

$\exists 1 \leq p \leq m$ и выберем $1 \leq i_1 < i_2 < \dots < i_p \leq m$

Проверяется гипотеза

$$H_0: \theta_{i_1} = \theta_{i_2} = \dots = \theta_{i_p} = 0$$

vs

(4)

$$H_1: \exists k: \theta_{i_k} \neq 0$$

Вновь можно считать описанную выше процедуру.

Если не обнаружено совместного значимого внешнего несбалансированных коэффициентов, то их можно одновременно исключить из модели

П4: Проверка адекватности модели

Будем считать модель адекватной, если предикт набор факторов ^{совместно} оказывает значимое влияние на y .

Формально проверяем

$$H_0: \theta_1 = \theta_2 = \dots = \theta_m = 0$$

(5)

$$H_1: \exists k: \theta_k \neq 0$$

39

Если H_0 отверг $\Rightarrow$ модель адекватна (по нашему подходу)

В противном случае выбрать набор факторов

не означает существование
и модель не адекватна $\Rightarrow$
 $\rightarrow$ ищем др. модель

Хотелось бы оценить ~~какую~~
значимость внешних факторов
по 1 числу.

можно показать, что справедливо
суммирование:

$$\sum_{j=1}^N [y_j - \bar{y}]^2 = \sum_{j=1}^N [y_j - \hat{y}_j]^2 + \sum_{j=1}^N [\hat{y}_j - \bar{y}]^2$$

$$TSS = ESS + RSS \quad (\text{total sum of squares})$$

TSS - полная сумма квадратов

ESS - сумма квадратов остатков
(error sum of squares)

RSS - уменьшение $\bar{y}$ за счет внешних
факторов - объясненная сумма
квадратов, объясненная с помощью
регрессии (Regression sum of squares)

В нек. источниках $ESS \leftrightarrow RSS$
(explained) $\leftrightarrow$ Residual ~~коэффициент~~
(объяснен) (остатки)

Определение: коэффициентом детерминации
называется число $R^2 := \frac{RSS}{TSS}$ (6)

Если R^2 близко к 1 $\Rightarrow$ модель хорошая
квадрата

R^2 - оценка $\sqrt{\text{коэффициент}}$ ~~коэф.~~
корреляции

Эта оценка емкая, ~~есть~~

R^2 обычно растет, когда учитывается

Можно показать, что

$$F = \frac{N - (m+1)}{m} \cdot \frac{R^2}{1 - R^2}, \text{ т.е. } F\text{-критерий}$$

однозначно заменяется чрез R^2

* задача: можно с помощью этих процедур проверить одинаковые ли уравнения (даже)

* понятие факторов переменных и оценка их влияния по этим методам

§7 Сравнение двух регрессий

Исследуется зависимость некоторой величины y от нескольких факторов.

Получены 2 набора измерений

$$y_j = \theta_1' x_{j1} + \theta_2' x_{j2} + \dots + \theta_m' x_{jm} + \varepsilon_j \quad (1) \quad j = \overline{1, N_1}$$

$$y_j = \theta_1'' x_{j1} + \theta_2'' x_{j2} + \dots + \theta_m'' x_{jm} + \varepsilon_j \quad (2) \quad j = \overline{N_1+1, N_1+N_2}$$

Предположим, что для (1) и (2) выполнены

основные ограничения и измеры в (1) и (2) независимы и ошибки имеют одинаковые дисперсии:

Проверяется гипотеза:

$$H_0: \theta_k' = \theta_k'' \quad \forall k$$

vs

$$H_1: \theta_k' \neq \theta_k'' \quad \text{для некоторого } k$$

41

Рассмотрим теперь новую модель

$$y_j = \theta_1' \tilde{x}_{j1} + \theta_m' \tilde{x}_{jm} + \theta_1'' \tilde{x}_{j,m+1} + \theta_2'' \tilde{x}_{j,m+2} + \dots + \theta_m'' \tilde{x}_{j,2m} + \varepsilon_j$$

$j = 1, N_1 + N_2$
(4) - новая классификация
(4)

Здесь

$$\tilde{x}_{jk} = \begin{cases} x_{jk}, & 1 \leq j \leq N_1, \\ 0, & N_1 + 1 \leq j \leq N_1 + N_2 \end{cases} \quad \text{для } 1 \leq k \leq m$$

$$\tilde{x}_{jk} = \begin{cases} 0, & 1 \leq j \leq N_1, \\ x_{j, k-m}, & N_1 + 1 \leq j \leq N_1 + N_2 \end{cases}$$

1) Оцениваем (4) без учета ограничений (отдельная оценка 2 составляющих (4)) и находим сумму остатков

$$ESS_{ur} = ESS_1 + ESS_2$$

2) Оцениваем (4) с учетом о.р., т.е. по всем $N_1 + N_2$ измерениям и находим ESS_R

3) Вычисляем

$$F = \frac{(ESS_R - ESS_{ur})/m}{ESS_{ur}/(N_1 + N_2 - 2m)}$$

F при верной H_0 имеет распределение Фишера с $(m, N_1 + N_2 - 2m)$ степенями свободы

Далее ~~для~~ критерий строится по стандартной схеме

Такой критерий называется тестом Чоу (Chow)

и $N_1 + 1$ - break point

Ср

Пример:

Сборные данные о размере заработной платы мужчин и женщин в Германии

$N_1 = N_2 = 75$ - количество мужчин и женщин

Изучается зависимость заработной платы в зависимости от пола и образования

W - величина заработной платы

X_1 - пол

X_2 - уровень образования

Список уравнение для мужчин

$$W^{(1)} = -3.37 + 0.479X_1 + 3.943X_2$$

$$ESS_1 = 5672.328$$

Список уравнение у женщин

$$W^{(2)} = -0.20 + 0.414X_1 + 2.305X_2$$

$$ESS_2 = 1788.344$$

Тогда можно есть и различия заработной платы для мужчин при прочих равных параметрах, применив тест F -оу

$$N_1 = N_2 = 75, N_1 + N_2 = 150, m = 3, 2m = 6$$

Получили следующие наблюдаемые значения F

$$F_n = 3.342$$

$$\text{число степеней свободы} = (3 + 44) = (3N_1 + N_2 - 2m)$$

$\Rightarrow p\text{-value} = 0,000\dots$

$\Rightarrow$ гипотеза H_0 (что уравнение однородно) отвергается, т.к. обнаружено значимое отличие в формировании заработной платы.

Но этот тест позволяет сказать только то, что ур-е гетеросkedастично. Для проверки отсутствия гетеросkedастичности

§8 Фактивные переменные

Ранее предполагалось, что факторы

Y - непрерывные, т.к. $\varepsilon_j \sim$ норм. распредел. и поэтому $\varepsilon \in (-\infty; +\infty)$. Но можно

рассмотреть и Y дискретные и даже категориальные (например пол или ок).

Примеры:

① Изучается зависимость зп Y от некоторых факторов для мужчин и женщин

② Спрос на мороженое в разное время года.

①: Вводим бинарную переменную

$$d = \begin{cases} 1, & \text{если выбрано мороженое} \\ 0, & \text{если не выбрано} \end{cases}$$

и рассматриваем уравнение

$$Y_j = \theta_1 X_{j1} + \theta_2 X_{j2} + \theta_3 d_j + \varepsilon_j, \quad d_j = \begin{matrix} \text{фактивная} \\ \text{(бинарная переменная)} \\ \text{dummy} \end{matrix}$$

②: Подробней учесть то же самое:

$$d = \begin{cases} 0, & \text{если зиме} \\ 1, & \text{весна} \\ 2, & \text{лето} \\ 3, & \text{осень} \end{cases}$$

В этом случае, переход с одного уровня на другой всегда будет осуществляться равномерно, т.е. перед d стоит один и тот же коэффич.

Предположим: учесть факторы (сезон) отдельно, лучше ввести несколько дихотомич. перемен.

$$d_1 = \begin{cases} 1, & \text{зиме} \\ 0, & \text{не зиме} \end{cases}$$

$$d_2 = \begin{cases} 1, & \text{весна} \\ 0, & \text{не весна} \end{cases}$$

$$d_3 = \begin{cases} 1, & \text{лето} \\ 0, & \text{не лето} \end{cases}$$

$$d_4 = \begin{cases} \text{осень} \\ \text{не осень} \end{cases}$$

Но эти факторы линейно зависимы (их сумма равна 1) и метод наим. квадр. не работает, потому что нужно какое-то d_i (одно) $i=1,4$.

③ Теперь рассмотрим задачу: размер зарплаты мужчин и женщин в Голландии

45

$$d = \begin{cases} 1, & \text{мужчина} \\ 0, & \text{женщина} \end{cases}$$

и рассмотрим модель

$$Y_j = \alpha + \beta_1 X_{j1} + \beta_2 X_{j2} + \beta_3 d_j + \epsilon_j$$

$\downarrow$ стаж $\downarrow$ удов. $\downarrow$ опыт $\downarrow$ пол регрессия

Оцениваем по методу наименьших квадратов эту модель $\Rightarrow$

→ 1) Константа (α) не значимая (средний уровень)

2) Все остальные факторы значимые:
в том числе β различия по полу

3) Получили среднюю уравнение:

$$y = -0.036 + 0.44x_1 + 3.24x_2 - 3.55d$$

→ переход из одной категории в другую приводит к разнице в зарплате

на 3.55. (высоз/час)

x_2 - дискр баллы

x_1 - возраст

46

4) $R^2 = 0.52 \Rightarrow$ объяснение факторы (x_1, x_2, d)
на 52% объясняют уровень зарплаты

5) Проверка адекватности

p-value = 0.000000... $\Rightarrow$ не проявляется,
что плохо

Литература (регрессионный анализ):

• Дрейтер и Шмит: "Прикладной регрессионный анализ"

Глава Дисперсионный анализ

* Методы Дисперсионный анализ

§1 Однофакторный дисперсионный анализ

Здесь имеем ~~есть~~ набор измерений:

$$Y_{kj} : \mu_k + \varepsilon_{kj} \quad (1) \quad k = \overline{1, m} \quad j = \overline{1, n}$$

k - уровень фактора

j - номер измерения

Предположения:

- 1) μ_k - неизвестные вещественные числа
- 2) $\varepsilon_{kj} \sim N(0, \sigma^2) \quad \forall k, j$
- 3) ε_{kj} - независимы

Препишем нашу модель в новом виде:

$$\mu = \frac{1}{m} \sum_{k=1}^m \mu_k \quad \text{и положим}$$

$$\alpha_k = \mu_k - \mu \Rightarrow \sum_k \alpha_k = 0 \Rightarrow$$

$$\Rightarrow Y_{jk} = \mu + \alpha_k + \varepsilon_{jk} \quad (2)$$

47

Основная задача: Есть ли различие в поведении Y на разных уровнях

Сформально проверим гипотезу H_0 :

$$H_0: \alpha_1 = \dots = \alpha_m = 0$$

vs

(3)

$$H_1: \exists k: \alpha_k \neq 0$$

Пример: Σ -количество выдыхаемого азота для четырех режимов питания $m=4$ (разные диеты)

Алгоритм проверки гипотез (3)

Применим к модели (2) Метод наим. квадр:

$$Q := \sum_k \sum_j [Y_{kj} - \mu - \alpha_k]^2 \rightarrow \min_{\mu, \alpha_k} \quad (4)$$

α_k - фактически фиктивный параметр

Получаем

$$\hat{\mu} = \bar{Y}_{..} = \sum_k \sum_j Y_{kj} / n \quad (5), \text{ где } n = mn$$

• означает усреднение по всем индексам

$$\hat{\alpha}_k = \bar{Y}_{k.} - \bar{Y}_{..} = \frac{1}{N} \sum_j Y_{kj} - \bar{Y}_{..}$$

Вычислим следующие величины

$$SS_T = \sum_k \sum_j (Y_{kj} - \bar{Y}_{..})^2 \quad \text{— полная сумма квадратов}$$

$SS_B = \sum_k (\bar{Y}_k - \bar{Y}_{..})^2$ - разброс между группами

$SS_R = \sum_k \sum_d (Y_{kd} - \bar{Y}_k)^2$ - разброс внутри групп

[T:] Если верна гипотеза H_0 (различий нет), то

1) Сл. величины SS_B и SS_R независимы и имеют χ^2 распределение с $(m-1)$ и

$(n-m)$ степен. св.

2) Сл. величина

$$F = \frac{SS_B / (m-1)}{SS_R / (n-m)}$$

имеет распределение Фишера

с $m-1, n-m$ степенями свободы

Данные критерий для проверки H_0 vs H_1 строится по стандартной схеме

Так, чтобы критическая область ~~была~~ соответствовала большому значению F . (т.е. Найти F_α , что при верной H_0 высказывает за него с вероятн-ю α)

Замечание: Если модель модели, где

ϵ_k - ссл. велич, независ от ошибок ϵ_{kj} и

$$\epsilon_k \sim N(0, \sigma_k^2)$$

[49]

В такой постановке имеется основная задача: ввести ли игровой фактор явно в общую дисперсионную модель

Формально :

$$H_0: \sigma_x^2 = 0$$

vs

$$H_1: \sigma_x^2 \neq 0 (\sigma_x^2 > 0)$$

Замерание: в однократном анализе 1 пар-фактор 2 парам. измерений 28.10.2015

Пример

Исследуется норма прибыли на обыкновенную акцию для фирм в трех отраслях промышленности: сталелитейная (Y_1), электроника (Y_2), энергетика (Y_3)

Три группы ($m=3$) по 3 измерениям в каждой группе ($N=9$)

Всего ($mN=n=27$) измерений

Мы хотим посчитать есть ли различия в нормах прибыли в разных отраслях промышленности. Алгоритм был описан в прошлой лекции.

Вычисляем среднее по группам:

$$\bar{Y}_1 = 9.78$$

$$\bar{Y}_2 = 18.93$$

$$\bar{Y}_3 = 13.1$$

Общее среднее

$$\bar{Y} = 13.93$$

51

Находим суммы квадратов отклонений в каждой группе и складываем их $\Rightarrow$

$\Rightarrow 436.62 \rightarrow$ сумма квадратов

Находим сумму квадратов отклонений по группам от общего среднего (и умножаем на то, сколько групп было) $\rightarrow 128.73$

Из общей теории статистики F при нулевой гипотезе (различий нет) (из стат: если нужно выявить, проверить нульно, что его нет) ~~соответствует~~ существует распределение Фишера с $(m-1, n-m) = (2, 24)$ степенями свободы.

Вычисляем критическое значение статистики: $F_H = \frac{128.73/2}{436.62/24} = 3.11$

[52] По таблицам распредел. Фишера с заданным количеством степеней свободы находим наименьший достигнутый уровень значимости.

$$p_{\text{value}} := P(F > F_H | H_0) = 0.063 < 0.1 = L$$

$\Rightarrow$ Получить это при верной H_0 маловероятно

$\Rightarrow H_0$ отвергается (~~не~~ обнаружено значимое отличие на уровне $L = 0.1$ по уже $\alpha = 0.05$ не происходит)

§2 Двухфакторный дисперсионный анализ.

Имеем следующую модель измерений:

$$Y_{kj} = \mu_{kj} + \varepsilon_{kj} \quad (1) \quad k = \overline{1, M} \quad j = \overline{1, N} \quad n = MN$$

На исследуемую величину влияют два фактора, k - уровень первого фактора, j - уровень второго фактора.

Замечание:

На этом этапе рассматривается ситуация, когда для каждого ^(совместия) набора факторов только одно измерение.

Предположения:

- 1) μ_{kj} - некоторая константа
- 2) ε_{kj} - независимы
- 3) $\varepsilon_{kj} \sim N(0, \sigma^2)$

Введем следующие обозначения:

$$\mu_{..} = \frac{1}{n} \sum_{k,j} \mu_{kj} \quad (2)$$

$$\mu_{.j} = \frac{1}{m} \sum_k \mu_{kj}$$

53

$$\tau_{j.} = \mu_{.j} - \mu_{..} \quad (3) - \text{эффект столбца}$$

$$\mu_{k.} = \frac{1}{N} \sum_j \mu_{kj}$$

$\Rightarrow$

$$\rho_k = \mu_{k.} - \mu_{..} - \text{эффект строки} \quad (4)$$

$$\Rightarrow M_{kj} = \mu_{..} + \rho_k + \tau_j \quad (5) -$$

- модель без учета взаимодействия факторов.

Перепишем модель:

$$(1) \Rightarrow Y_{kj} = \mu_{..} + \rho_k + \tau_j + \varepsilon_{kj} \quad (6)$$

[54]

случайные ошибки

Решим (6) по методу наименьших квадратов

$$\hat{\mu}_{..} = \bar{Y}_{..} = \frac{1}{n} \sum_{k,j} Y_{kj}$$

$$\hat{\rho}_k = \frac{1}{N} \sum_j Y_{kj} - \bar{Y}_{..} = \bar{Y}_{k.} - \bar{Y}_{..}$$

$$\hat{\tau}_j = \frac{1}{m} \sum_k Y_{kj} - \bar{Y}_{..} = \bar{Y}_{.j} - \bar{Y}_{..} \Rightarrow$$

$$\stackrel{(6)}{\Rightarrow} Y_{kj} = \bar{Y}_{..} + (\bar{Y}_{k.} - \bar{Y}_{..}) + (\bar{Y}_{.j} - \bar{Y}_{..}) + \varepsilon_{kj} \quad (7)$$

ε_{kj} - остатки

Задача: Есть ли внешние факторов

(есть ли эффект строки или столбца)

Проверим H_0 есть ли разница средних по строкам. Формально:

$$H_0: \rho_1 = \rho_2 = \dots = \rho_m = 0$$

vs

$$H_1: \exists k \in [1, m] : \rho_k \neq 0$$

Описание процедуры:

1. Оцениваем модель (6) без учета H_0 и находим сумму квадратов остатков: $S_1^2 = \sum_{k,j} e_{kj}^2$ (8)

2. Оцениваем модель (6) с учетом H_0 :

$$\text{Тогда } \hat{\mu}_{..} = \bar{Y}_{..}, \quad \hat{\gamma}_j = \bar{Y}_{.j} - \bar{Y}_{..}$$

$$Y_{kj} = \bar{Y}_{..} + (\bar{Y}_{.j} - \bar{Y}_{..}) + \tilde{e}_{kj} \quad (9)$$

и находим сумму квадратов остатков

$$S_0^2 := \sum_{k,j} \tilde{e}_{kj}^2$$

155

3. Теорема: При верной H_0

1) с.в. S_1^2 и $S_2^2 = S_0^2 - S_1^2$ — независимы

2) с.в. $S_1^2/6^2$ и $S_2^2/6^2$ — имеют χ^2 распредел с $(N-1)(m-1)$ и $(m-1)$ степеней своб.

3) с.в. $F = \frac{S_2^2/(m-1)}{S_1^2/((N-1)(m-1))}$ имеет распредел Фишера с $(m-1), (N-1)(m-1)$ степеней свободы

такая же процедура проверки применяется по стандартной схеме.

Пример: Изучается процесс выноса по срокам выноса в трех федеральных округах ($m=3$) в зависимости от размера баки.

разр. по размеру банка	Ново- Горк.	Аппонте	Сам-Фран- цеско	Среднее по строке
<10	4.65	4.55	4.44	4.543
10-50	4.76	4.40	4.99	4.717
50-100	4.60	4.92	5.00	4.840
100-500	4.83	4.70	4.97	4.833
≥500	4.93	4.46	4.93	4.793
средн. по столб.				
общее средн.				4.746

56 Проверим, есть ли зависимость ставок от размера банка (зависимость по строкам)

$$H_0: p_1 = p_2 = \dots = p_5 = 0 \quad m=5 \quad N=3 \quad n=5 \times 3 = 15$$

$$H_1: \exists i: p_i \neq 0$$

Вычисляем модель без учета H_0 , получаем

$$S_1^2 = 0.3215$$

Вычисляем модель с учетом H_0 и находим

$$S_R^2 = 0.1779$$

Наблюдаемое значение статистики F:

$$F_H = \frac{S_R^2 / (m-1)}{S_1^2 / (N-1)(m-1)} = 1.1$$

Из общей теории ст. F имеет распределение

Фишера с (4, 8). По табл. достижим. знач.

p-value = 0.42 $\Rightarrow$ не обнаружено значимого различия в ставках в банках (p-value не такое маленькое, такое знач. мы получили случайно)

Глава Дискриминантный анализ

Примеры: 1) классификация клиентов банка

2) классификация студентов по уровню знаний

3) тестирование при приеме на работу

4) изучение эффекта влияния некоторого лекарства

Две основные задачи:

57

1) интерпретация. Основной вопрос: можно ли по измеренным характеристикам различить изучаемые совокупности.

2) классификация. Основная задача: Найти среди них несколько функций от измеренных характеристик, которые позволят разделить изучаемые группы.

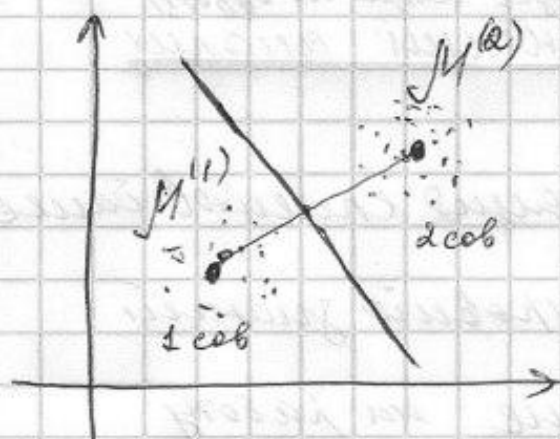
§1 Постановка задачи

Каждый изучаемый объект характеризуется парой чисел x_1 и x_2 (берем двумерный простотки) $(x_1, x_2) = x$

Всюду далее предполагается, что x имеет двумерное нормальное распределение со средним $\mu = (\mu_1, \mu_2)^T$ и некоторой матрицей ковариаций Σ .

Теперь имеем 2 совокупности, которые различаются средними $(\mu^{(1)}, \mu^{(2)}) = (\mu_1^{(1)}, \mu_2^{(1)}, \mu_1^{(2)}, \mu_2^{(2)})$ но имеют одну и ту же матрицу ковариаций Σ .

Изображение 2 совокупностей



Задача: отнести вновь поступивший объект с хар (x_1, x_2) к одной из этих совокупностей.

Наиболее просто построить линейную F и сравнить новую T с этой прямой.

Если $M^{(1)}, M^{(2)}$ - известны и Σ один на обеих совокуп, то задача легко решается

В реальной задаче $M^{(1)}, M^{(2)}$ неизвестны и Σ из разных совокупностей

В этом случае первый этап явл этап обучения, когда мы имеем 2 набора измерений, про каждый из которых мы знаем из какой совокупности

можно найти алгоритм с миним. вероятностью ошибок.

Два основных подхода:

- 1) метод главных компонент
- 2) метод канонических корреляций

09.12.2015

Дискриминантный анализ. Продолжение.

§ Задача классификации в случае известных параметров

У нас есть несколько генеральных совокупностей объектов $W_1, \dots, W_k$, которые описываются p признаков, характеризующих $(x_1, \dots, x_p) = x$ - p -мерный вектор

Задача: разработать правило, чтобы отнести к какому W_i отнести новый вектор

Пример 1:

Приминал комиссия некоторого учебного заведения хочет отнести очередного студента абитуриента к одной из 2 категорий.

W_1 : успешно закончит

W_2 : будет отчислен

$x_1, \dots, x_p$ - результаты сдачи вступительных экзаменов

Пример 2: Врач на основе p симптомов должен определить наличие у больного одной из вероятных к болезней

Далее мы предполагаем, что у нас две совокупности W_1 и W_2 и p -мерный вектор $x = (x_1, x_2, \dots, x_p)$ имеют многомерное нормальное распределение с векторами средних $M = (m_1, \dots, m_p)$

и матрицей ковариаций $\Sigma = (\sigma_{ij})$

Будем считать, что матрица Σ одинакова у обеих совокупностей и они отличаются только средними.

$$m^{(1)} = (m_{11}, \dots, m_{1p})$$

$$m^{(2)} = (m_{21}, \dots, m_{2p})$$

Предполагается, что Σ и m - известны

Для разнесения этих совокупностей будем принимать линейную дискриминантную функцию Z :

$$Z = L_1 x_1 + \dots + L_p x_p \quad (1)$$

(Будем считать, что опорный элемент L_0 - включено в x)

Тогда правило классификации будет иметь вид:

$$\text{если } Z \geq c \Rightarrow W_1$$

$$\text{если } Z < c \Rightarrow W_2$$

т.е. Z имеет одномерное нормальное распределение со средним

$$\boxed{60} \quad \bar{Z} = \sum_{i=1}^p L_i m_i(2) \quad \text{дисперсией } \sigma_Z^2 = \sum_{i,j=1}^p \sigma_{ij} L_i L_j \quad (3)$$

т.е. (2) и (3) известны

Вопрос: чем дальше среднее, тем классифицируем

Расстояние Махаланобиса: $\Delta^2 = \frac{(I_1 - I_2)^2}{\sigma^2}$ (4)

где I_1 и I_2 - средние для Z в W_1 и W_2 .

Надо найти константу C .

Основная задача:

- 1) коэффициенты $L_1, \dots, L_p$
- 2) константа C

Итуговое решение:

1) $L_1, \dots, L_p$ находим из условия, чтоб Δ^2 была максимальной L : $\Delta^2 \rightarrow \max$ (5)

2) C - есть середина отрезка с концами I_1 и I_2

Далее для решения этой задачи используется Байесовская постановка:

На основе прошлой информации мы знаем, что очередной наблюдатель ответит принадлежащий W_1 с вер-ю q_1 и W_2 с вер-ю q_2 , $q_1 + q_2 = 1$

сл. вектор X имеет распределение

$p(x|W_1)$ в случае W_1 , $p(x|W_2)$ в случае

второй совокупности.

[61]

Если для вектора X известны конкретные значения x , то по формуле Байеса

получаем апостериорные вероятности:

$$p(w_2 | x) = \frac{q_2 P(x | w_2)}{\sum_{c=1}^2 q_c P(x | w_c)} \quad (6)$$

Классификация в байесовской установке проводится по правилу:

1) Если $p(w_1 | x) \geq p(w_2 | x) \Rightarrow w_1$

2) Если $p(w_2 | x) > p(w_1 | x) \Rightarrow w_2$

$P(2|1)$ - вероятность выбрать 2, когда
верна 1,

$P(1|2)$ - вероятность выбрать 1, когда
верна 2.

Рассмотрим величину $q_1 P(1|1) + q_2 P(1|2)$ (7)

(7) - ожидаемая вероятность ошибочной
классификации

Результат: можно доказать, что описанное
выше правило байесовской классификации
приводит к минимальной ожидаемой
вероятности ошибочной классификации.

Если применить к нашей модели
эту формулу, то получаем:

1) Принимаем w_1 , если $\sum_{i=1}^P x_i \geq \frac{S_1 + S_2}{2} + \ln\left(\frac{q_2}{q_1}\right)$ (8)

(62)

2) При W_2 , если $\sum_{i=1}^p \alpha_i x_i < \frac{y_1 + y_2}{2} + \ln \left(\frac{q_2}{q_1} \right)$ (9)

Если $q_1 = q_2 \Rightarrow C = \frac{y_1 + y_2}{2}$

И некоторое обобщение такой постановки задачи: применим неверный решение с равной вероятностью.

$C(112)$ - потери, связ с предметами

W_1 , когда имели 2

$C(211)$ - аналогично.

Аналогичное рассуждение приводит нас к обобщенной байесовской классиф.

1) Применим W_1 , если $\sum_{i=1}^p \alpha_i x_i \geq \frac{y_1 + y_2}{2} + \ln \frac{q_2 C(112)}{q_1 C(211)}$ (10)

2) Применим W_2 , если $\sum_{i=1}^p \alpha_i x_i < \frac{y_1 + y_2}{2} + \ln \frac{q_2 C(112)}{q_1 C(211)}$ (11)

Обобщенная байесовская процедура минимизирует ожидаемую стоимость ошибки классификации.

$q_1 C(211) p(211) + q_2 C(112) p(112)$ (12)

В случае нормального распределения имеем:

$$p(2|1) = \Phi\left(\frac{k - \frac{1}{2}\Delta^2}{\Delta}\right) \quad (13)$$

Функция распредел стандарт норм
закона

$$p(1|2) = \Phi\left(\frac{-k - \frac{1}{2}\Delta^2}{\Delta}\right) \quad (14)$$

$$k = \ln \frac{q_2 C(1|2)}{q_1 C(2|1)} \quad (15)$$

Мы исходим из того что средние и Σ
известны, если нет, то считаем,
далее процедура та же

• Срок 23.03, в 12.00 17-8

16.12.2015

Продолжение

§ Задача дискриминации для
двух совокупностей в
случае неск параметров

Имеем вектор наблюдений

$$X = (x_1, \dots, x_p)^T$$

Задача : Отнести эти наблюд к 1
из 2 совокупностей W_1 и W_2 , кот
имеют многомерные норм

распределения $N(m_1, \Sigma)$ или $N(m_2, \Sigma)$

m_1, m_2, Σ : теперь неизвестны $\Rightarrow$

их надо оценить. $\Rightarrow$ надо провести

наблюдения, где мы знаем совокупности (W_1, W_2)

$\Rightarrow$ обучающая выборка

У нас есть 2 разн. выборки;

$$X_{11}, \dots, X_{1n_1} \sim N(\mu_1, \Sigma) \quad (X_{1j} - \text{вектор размера } p)$$

$$X_{21}, \dots, X_{2n_2} \sim N(\mu_2, \Sigma)$$

Но этими наблюдениями строим оценки

$$\hat{\mu}_1 = \frac{1}{n_1} \sum_j X_{1j}$$

$$\hat{\mu}_2 = \frac{1}{n_2} \sum_j X_{2j}$$

Оцениваем средние

$$\hat{\Sigma} = \frac{1}{n_1 + n_2 - 2} [(n_1 - 1) S_1 + (n_2 - 1) S_2] = S = (S_{jk}) \in \mathbb{R}^{[p \times p]}$$

$$S_1 = \frac{1}{n_1 - 1} \sum_{j=1}^{n_1} (x_{1j} - \hat{\mu}_1)(x_{1j} - \hat{\mu}_1)^T \quad (\text{Аналогично } S_2)$$

Ранее были построены алгоритмы оценивания
Байесовской процедуры для дискриминации
двух нормальных совокупностей с извест-
ными параметрами.

65

Алгоритмы решения задачи
дискриминации (при извест-
ных параметрах)

1) Находим коэффициенты $\alpha_1, \dots, \alpha_p$
как решения следующей системы:

$$\begin{cases} \alpha_1 S_{11} + \alpha_2 S_{12} + \dots + \alpha_p S_{1p} = \hat{m}_{11} - \hat{m}_{21} \\ \alpha_1 S_{p1} + \alpha_2 S_{p2} + \dots + \alpha_p S_{pp} = \hat{m}_{1p} - \hat{m}_{2p} \end{cases}$$

т.к. у нас не m , а $\hat{m} \Rightarrow$ находим
отсюда не α , а $\hat{\alpha}$ (их оценки)

2) Находим оценку для дискриминантной
функции $\hat{Z}$ (тоже только оценка)

$$\hat{Z} = \hat{\alpha}_1 X_1 + \dots + \hat{\alpha}_p X_p$$

3) Находим оценки для S_1 и S_2 по правому

$$\begin{aligned} \text{а) } \hat{Z}_{1j} &= \hat{\alpha}_1 X_{1j}^{(1)} + \dots + \hat{\alpha}_p X_{pj}^{(1)} \\ \hat{Z}_{2j} &= \dots \end{aligned}$$

$$\begin{aligned} \text{б) } \hat{S}_1 &= \frac{1}{n_1} \sum_j \hat{Z}_{1j} \\ \hat{S}_2 &= \frac{1}{n_2} \sum_j \hat{Z}_{2j} \end{aligned}$$

4) Определяем S_2^2 по правому

$$S_2^2 = \sum_{i=1}^p \sum_{j=1}^p \hat{\alpha}_i S_{ij} \hat{\alpha}_j$$

5) Искомый $X = (x_1, \dots, x_p)^T$, если

$$Z = \sum_{i=1}^p \hat{\alpha}_i x_i \geq \frac{\hat{S}_1 + \hat{S}_2}{2} + \ln \frac{q_2 c(1/2)}{q_1 c(2/1)}$$

Глава ... Кластерный анализ

- Литература : 1) Маудель U.D. "Кластерный анализ"
2) Дорай Б., Огера П. "Кластерный анализ"
(по всей лекции)
3) Хорошая обзорная книга "Факторный анализ,
кластерный анализ"

§1 Описание данных

Из некоторой измеримой совокупности Π
выбрано n объектов $I = (I_1, \dots, I_n)$ у
каждого объекта измерено p характеристик
 $C = (C_1, \dots, C_p)^T$

C_k - количественные характеристики

x_{ij} - результат измерения i -ой x -ки у
объекта I_j

$$X = (x_{ij}), \quad i = \overline{1, p}, \quad j = \overline{1, n}$$

$x_j = (x_{1j}, \dots, x_{pj})^T$ - измер. x -ки из объекта I_j

$$X = (x_1, x_2, \dots, x_n)$$

$$x_j \sim \mathbb{R}^p$$

§2 Задача кластерного анализа

I ~~m~~ $m < n$ (если $m = n \Rightarrow$ каждый объект - класс)

Требуется разбить I на m классов $\Pi_1, \dots, \Pi_m$!

так чтобы каждый объект относился
и только одному кластеру.

При этом внутри кластера
объекты в некотором смысле похожи,
а в разных - не сходные.

Замечание: в реальной задаче как
правильно, не известно число кластеров
и если кластер строится последовательно

§ 3 Метрики

§ 4 Дистанции между кластерами

§ 5 Последовательное построение кластеров

- 1) Сначала все объекты рассматриваются
как отдельные кластеры
- 2) Выбираем 2 порога: $0 < \tau < S$
- 3) Если раст между всеми кластерами
больше S , процедура останавливается
- 4) Если расстояние между какими-то
кластерами меньше $S \Rightarrow$ находим
два ближайших и объединяем
их в 1 кластер, если раст между
в новом кластере не больше τ .

5) Пересчитываем новые расстояния между кластерами.

6) Процедура продолжается до тех пор, пока $\ast$ внутри одного кластера не более ϵ , а между какими-то кластерами не более S . $\rightarrow \ast$