№17 Дана выборка. Найти главные направления и дисперсии по главным компонентам.

x_1	4	0	-1	3	4
x_2	2	-3	-2	1	2
x_3	3	2	2	1	-3

Решение:

$$X = \begin{pmatrix} 4 & 2 & 3 \\ 0 & -3 & 2 \\ -1 & -2 & 2 \\ 3 & 1 & 1 \\ 4 & 2 & -3 \end{pmatrix}$$ признаки

объекты

1) Центрируем. Находим среднее и вычитаем из каждой компоненты.

$$\hat{x} = (2, 0, 1), \quad X_c = \begin{pmatrix} 2 & 2 & 2 \\ -2 & -3 & 1 \\ -3 & -2 & 1 \\ 1 & 1 & 0 \\ 2 & 2 & -4 \end{pmatrix}$$

2) $$C = X_c^T X_c = \begin{pmatrix} 2 & -2 & -3 & 1 & 2 \\ 2 & -3 & -2 & 1 & 2 \\ 2 & 1 & 1 & 0 & -4 \end{pmatrix} \begin{pmatrix} 2 & 2 & 2 \\ -2 & -3 & 1 \\ -3 & -2 & 1 \\ 1 & 1 & 0 \\ 2 & 2 & -4 \end{pmatrix} = \begin{pmatrix} 22 & 21 & -9 \\ 21 & 22 & -9 \\ -9 & -9 & 22 \end{pmatrix}$$

$$\frac{1}{N-1} C = \begin{pmatrix} 5{,}5 & 5{,}25 & -2{,}25 \\ 5{,}25 & 5{,}5 & -2{,}25 \\ -2{,}25 & -2{,}25 & 5{,}5 \end{pmatrix}$$ – выборочная матрица ковариации

3) Найдем с.ч. и с.в. C:

$$\det(C - \lambda I) = \begin{vmatrix} 22-\lambda & 21 & -9 \\ 21 & 22-\lambda & -9 \\ -9 & -9 & 22 \end{vmatrix} = -(\lambda-1)(\lambda-16)(\lambda-49) = 0.$$

$$\lambda_1 = 1, \quad v_1 = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}$$

$$\lambda_2 = 16, \quad v_2 = \begin{pmatrix} 1/3 \\ 1/3 \\ 1 \end{pmatrix}$$

$$\lambda_3 = 49, \quad v_3 = \begin{pmatrix} -3/2 \\ -3/2 \\ 1 \end{pmatrix}$$

Собственные векторы нормируем:

$$v_1 = \frac{\sqrt{2}}{2}\begin{pmatrix}-1\\1\\0\end{pmatrix}, \quad v_2 = \frac{\sqrt{11}}{11}\begin{pmatrix}1\\1\\3\end{pmatrix}, \quad v_3 = \frac{\sqrt{22}}{22}\begin{pmatrix}-3\\-3\\2\end{pmatrix}$$

$\boxed{v_1, v_2, v_3 \text{ - главные компоненты}}$

$$\frac{1}{N-1}\lambda_1 = \frac{1}{N-1}\sigma_1^2 = \frac{1}{4} = 0{,}25$$

$$\frac{1}{N-1}\lambda_2 = \frac{1}{N-1}\sigma_2^2 = 4$$

$$\frac{1}{N-1}\lambda_3 = \frac{1}{N-1}\sigma_3^2 = \frac{49}{4} = 12{,}25$$

дисперсии по главным компонентам

Видим, что 3-я компонента наиболее объясняет суммарную дисперсию стандартизованных признаков, а, например 1-я - значительно меньше. В задаче понижения размерности пространства признаков можно было бы попробовать избавиться от 1 компоненты и посмотреть на результат

(№35) Задача восстановления регрессии с квадратичной функцией потерь $L(y', y) = (y' - y)^2$. Док-ть, что если $f^*(x) = \arg\min_c E((Y-c)^2 / X = x)$, то $f^*(x) = E(Y / X = x)$ - регрессионная ф-я. Чему тогда равен средний риск $R(f^*)$?

Решение: ~~функция потерь~~ ~~Δ~~

используется для док-ва:

$$R(f) = E\,L(f(X), Y) = \int_{X \times Y} L(f(x), y)\,p(x, y)\,dx\,dy$$

мат ожидание ф-ии потерь, средний риск или ожидаемая ошибка предсказания

$$R(f) = \int_X \int_Y L(f(x), y)\,p(y/x)\,dy\,p(x)\,dx$$

$$//\ p(x,y)\,dx\,dy = p(x)\,dx\ p(y/x)\,dy$$

1) Для квадратичной ф-ии потерь:

$$R(f) = \int_X \int_Y (f(x)-y)^2 p(y|x)\,dy\, p(x)\,dx = \int_X E((f(x)-Y)^2/x)\,p(x)\,dx$$

$\to min$

в задаче восстановления регрессии $R(f) \to \min_f$, минимизировать можно поточечно:

$$f^*(x) = \arg\min_c E((c-Y)^2/x)$$ фиксир. x

$\frac{\partial}{\partial c} E((c-Y)^2/x) = 0$ — найдем из этого условия

$$\frac{\partial}{\partial c} \int_Y (c-y)^2 p(y|x)\,dy = 2\int_Y (y-c)p(y|x)\,dy =$$

$$= 2\Big(\underbrace{\int_Y y\,p(y|x)\,dy}_{=E(y|x)} - \underbrace{\int_Y c\,p(y|x)\,dy}_{=c\cdot 1}\Big) \ominus$$

$$\ominus\ 2(E(y|x) - c) = 0 \Rightarrow c = E(y|x)$$

т.к. решали $\forall x \in X$, то обобщим для всего множества:

$$f^*(x) = E(Y/X=x) \quad \text{ч.т.д.}$$

Т.о. в случае квадратичной ф-ии потерь наилучшим предсказанием y в ответ на вход x явл. условное среднее (регрессионная ф-я)

$R(f^*)$ — байесовский риск (неустранимая ошибка)

2) Для $R(f^*)$ докажем: $R(f^*) = E_X D_Y(Y/X)$:

$$R(f^*) = \int_X \int_Y (f^*(x)-y)^2 p(y|x)\,dy\,p(x)\,dx =$$

$$= \int_X p(x)\Big[\int_Y (E(y|x)-y)^2 p(y|x)\,dy\Big]dx =$$

$$= E_y((E_y(y|x)-y)^2/x) \overset{\Delta}{=} \int_X p(x) D_y(y|x)\,dx \overset{\Delta}{=}$$

$$\overset{\Delta}{=} E_x(D_y(y|x)) \quad \text{ч.т.д.}$$

№36. Задача восстановления регрессии с ф-ей потерь $L(y', y) = |y' - y|$. Док-ть, что минимум среднего риска доставляет при этом условная медиана

$$f^*(x) = \text{median}(Y \mid X = x)$$

Решение:

$$R(f) = \int_X \underbrace{p(x)}_{\geq 0} \left[\int_Y |f(x) - y| \, p(y|x) dy \right] dx \to \min$$

минимизировать можно поточечно:

$$\int_Y |c - y| \, p(y|\overset{\text{фиксир.}}{x}) dy \to \min, \quad \text{т.к. } \frac{\partial}{\partial c} E[L(c,y)|x] = 0$$

$$\frac{\partial}{\partial c} \int_Y L(c,y) p(y|x) dy = \frac{\partial}{\partial c} \int_Y |c - y| \, p(y|x) dy =$$

$$= \int_{Y \setminus \{c\}} \text{sign}(c-y) p(y|x) dy = \int_{-\infty}^{c} \text{sign}(c-y) p(y|x) dy +$$

$$+ \int_{c}^{+\infty} \overbrace{\text{sign}(c-y)}^{=-1} p(y|x) dy = P(\{c > y\}|x) - P(\{c < y\}|x) =$$

$$= 0 \quad \Rightarrow \quad P(\{c > y\}|x) = P(\{c < y\}|x)$$

Если $p(y|x)$ - непрерывно, то $P(\{c = y\}|x) = 0$

$$\Rightarrow P(\{c > y\}|x) = P(\{c < y\}|x) = \frac{1}{2}$$

$$\Rightarrow c = \text{median}(y|x)$$

т.к. решили $\forall x \in X$, то обобщим на все мн-во:

$$f^*(x) = \text{median}(Y|X = x) \quad \text{чтд}$$

№37. Как должна выглядеть ф-я потерь, чтобы минимизир. среднему риску давала условная мода?

Решение:

$$R(f) \to \min_f \iff E_y\left(L(c,y) \mid \overset{\text{фиксир.}}{x}\right) \to \min_c$$

т.е. $\int\limits_{\mathbb{R}} L(c,y)\, p(y|x)\, dy \to \min\limits_c$

$f^*(x) = \text{mode}(Y|X=x) \Rightarrow p(c|x) \to \max$

при фиксир. x и $f^*(x) = c^*$: $c^* = \text{mode}(y|x) \Rightarrow$

$\Rightarrow p(c^*|x) = \max\limits_y p(y|x)$

$\Rightarrow c^* = \text{mode}(y|x) \Rightarrow -p(c^*|x) = \min\limits_c(-p(c|x))$

$\int\limits_{\mathbb{R}} L(c,y)\, p(y|x)\, dy = -p(c|x)$

дельта-ф-я Дирака удовлетв. усл-ю: $-L(c,y) = \delta(c-y) = \delta(y-c)$

св-во: $\int\limits_{\mathbb{R}} f(x)\,\delta(x-x_0)\,dx = f(x_0)$

$L(c,y) = -\delta(y-c)$, т.к. $\delta(y-c) \geq 0$, $L(c,y) \geq 0$, то

добавим $\alpha > 0$: $\boxed{L(c,y) = \alpha - \delta(y-c)}$

№4.1 Док-ть, что можно положить $v_0 = \hat{x}$. Описать ли во всех возможных v_0 дост-х $\min \sum (\ )^2$ расстояний до искомого многообразия

Решение:

$$W_k = v_0 + L(v_1, v_2, \dots v_k) \text{ — k-мерное лин. многообр.}$$

$$\sum_{i=1}^{N} dist^2(x^{(i)}, W_k) \to \min$$

К $W_0 = v_0 \in \mathbb{R}^d \quad \|v_0\| = 1$

$$W_0 = \underset{v_0 \in \mathbb{R}^d}{\arg\min} \sum_{i=1}^{N} dist^2(x^{(i)}, v_0) = \underset{v_0 \in \mathbb{R}^d}{\arg\min} \sum_{i=1}^{N} \|x^{(i)} - v_0\|^2 = 0$$

$$\frac{d}{dv_0} \bigcirc = 0 : \quad \sum_{i=1}^{N} \frac{d}{dv_0}\left(\|x^{(i)} - v_0\|^2\right) = 0.$$

$$\sum_{i=1}^{N} \left(\underbrace{\frac{d(x^{(i)} - v_0)}{dv_0}}_{-1} \cdot \underbrace{\frac{d(\|x^{(i)} - v_0\|)^2}{d(x^{(i)} - v_0)}}_{2 \cdot (x^{(i)} - v_0)} \right) = 0$$

$$\sum_{i=1}^{N} (-1) \cdot 2 \cdot (x^{(i)} - v_0) = 0 \quad \Big| \left(-\tfrac{1}{2}\right)$$

$$\sum_{i=1}^{N} (x^{(i)} - v_0) = 0 \Rightarrow \sum_{i=1}^{N} x^{(i)} - v_0 \sum_{i=1}^{N} 1 = 0$$

$$\sum_{i=1}^{N} x^{(i)} - N \cdot v_0 = 0 \Rightarrow \sum_{i=1}^{N} x^{(i)} = N \cdot v_0$$

$$\Rightarrow \quad v_0 = \frac{1}{N} \sum_{i=1}^{N} x^{(i)} = \hat{x} \qquad \text{чтд}$$