Segunda Prova
Fundamentos Estatísticos para Ciência dos Dados

18/05/2018

Parte sem consulta

1. O gráfico na Figura 1 exibe uma amostra do vetor aletório (X, Y) com certa densidade $f(x, y)$. Com base neste gráfico, identifique a opção correta:

   - $\mathbb{E}(Y|X = 8) \approx$??: (i) 8 (ii) 18 (iii) 28 (iv) 38 (v) 48
   - $\mathbb{E}(X|Y = 0) \approx$??: (i) 2 (ii) 4 (iii) 8 (iv) 10 (v) 0
   - $\sigma(Y|X = x) = \sqrt{\mathbb{V}(Y|X = x)}$ é uma função de x. Ela é:
     (i) crescente em x;
     (ii) constante com respeito a x;
     (iii) decrescente em x;
     (iv) parabólica em x.
   - $\sigma(X|Y = 0) = \sqrt{\mathbb{V}(X|Y = 0)}$ é aproximadamente igual a: (i) 1 (ii) 2 (iii) 4 (iv) 8 (v) 0.1

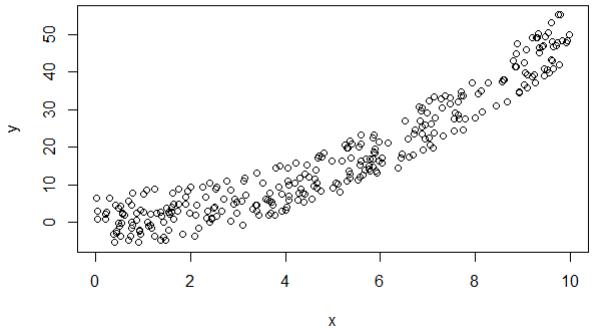

Figura 1: Amostra de um vetor aleatório (X, Y).

2. A Tabela 1 mostra a distribuição conjunta do vetor aletório discreto (X, Y). Obtenha a distribuição marginal da variável X e a distribuição condicional da variável $(Y|X = 1)$.

3. Numa análise de componentes principais com $k = 6$ variáveis, os autovalores foram obtidos: $\lambda_1 = 6, \lambda_2 = 4, \lambda_3 = 1, \lambda_4 = 0.1, \lambda_5 = 0.1$ e $\lambda_6 = 0.01$ Quantos componentes devem ser usados? Justifique sua resposta calculando a proporção acumulada da variância total exlicada pelos primeiros k autovetores.

4. Seja $\rho \in (0, 1)$. Mostre que a matriz de covariância

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix}$$

$y\|x$	$x=0$	$x=1$	$x=2$
$y=0$	0.1	0.05	0.05
$y=1$	0.1	0	0.2
$y=2$	0	0.2	0.1
$y=3$	0.05	0.1	0.05

Tabela 1: Distribuição de probabilidade discreta do vetor (X, Y).

do vetor aleatório $\mathbf{X} = (X_1, X_2, X_3)^t$ possui um autovetor igual a $\mathbf{v} = (1, 1, 1)/\sqrt{3}$. Qual o autovalor associado com este autovetor?

5. Resultado (2-51) de Johnson and Wichern: Seja $\mathbf{B}$ uma matriz definida positiva $p \times p$ com autovalores $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_p > 0$ e associados autovetores $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p$ de comprimento (ou norma) 1. Então

$$\max_{\mathbf{x} \neq \mathbf{0}} \frac{\mathbf{x}'\mathbf{B}\mathbf{x}}{\mathbf{x}'\mathbf{x}} = \lambda_1$$

e este máximo é atingido quando $\mathbf{x} = \mathbf{v}_1$.

6. Resultado 8.1 da página 432 de Johnson and Wichern: : Seja $\mathbf{\Sigma}$ a matriz de covariância do vetor aleatório $\mathbf{X} = (X_1, \ldots, X_p)'$ com autovalores $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_p > 0$ e associados autovetores $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p$ de comprimento (ou norma) 1. Então a combinação linear $Y = l_1 X_1 + \ldots + l_p X_p = \mathbf{l}'\mathbf{X}$ com comprimento $||\mathbf{l}|| = 1$ e que maximiza $\mathbb{V}(Y)$ é obtida ao tomarmos $\mathbf{l}$ igual ao primeiro autovetor. Neste caso, $Y = \mathbf{v}_1'\mathbf{X}$ e a variância desta variável atinge $\mathbb{V}(Y) = \lambda_1$.

Parte com consulta

1. Vamos analisar um conjunto de dados que possui apenas 3 variáveis X_1, X_2, X_3 usando o modelo de análise fatorial ortogonal. Os autovalores e autovetores da matriz de covariância da amostra de três variáveis são os seguintes:

$$\lambda_1 = 2.25, \qquad \lambda = 1.96, \qquad \lambda_3 = 0.16$$

$$\mathbf{v}_1 = \begin{bmatrix} 1/2 \\ 1/2 \\ 1/\sqrt{2} \end{bmatrix} \qquad \mathbf{v}_2 = \begin{bmatrix} 1/2 \\ 1/2 \\ -1/\sqrt{2} \end{bmatrix} \qquad \mathbf{v}_3 = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \\ 0 \end{bmatrix}$$

(i) Reconstrua a matriz de covariância das variáveis com base nesses autovalores e autovetores. Suponha que haja apenas um único fator no modelo fatorial. Usando a abordagem de componente principal para encontrar: (ii) a matriz de carga de fator, (iii) as comunalidades, (iv) as variâncias específicas.

2. Suponha que o vetor aleatório contínuo e positivo $\mathbf{X} = (X_1, X_2)$ possui a densidade $f_1(\mathbf{x}) = 6\exp(-(3x_1 + 2x_2))$ quando o indivíduo pertence à população 1. Quando ele pertence à população 2, temos $f_2(\mathbf{x}) = \exp(-(x_1 + x_2))$. O custo $c(1|2)$ do erro de classificar erradamente no grupo 1 um indivíduo do grupo 2 é 3 vezes maior que o custo contrário $c(2|1)$ de colocar no grupo 2 alguém do grupo 1. Se o grupo 1 constitui 90% da população total, mostre que a região ótima R_1 de classificação no grupo 1 é dada pelo semi-plano $2x_1 + x_2 \leq \log(18)$.