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) 33 (iv) 40 (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

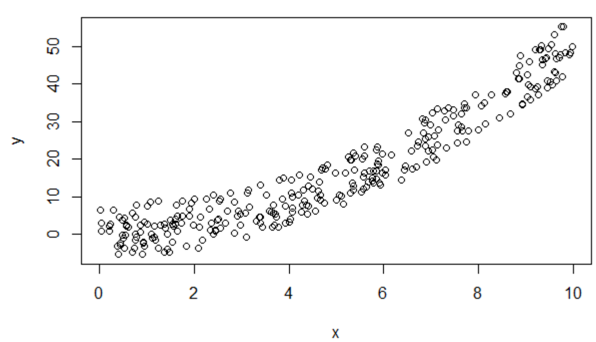

Figura 1: Amostra de um vetor aleatório (X, Y).

Solução:

- $\mathbb{E}(Y|X = 8) \approx 33$ (iii)
- $\mathbb{E}(X|Y = 0) \approx 2$ (i)
- $\sigma(Y|X = x) = \sqrt{\mathbb{V}(Y|X = x)}$ é uma função de x. Ela é: (ii) constante com respeito a x. Veja que é a $\mathbb{E}(Y|X = x)$, a esperança de Y como função de x, quem tem uma forma crescente (ligeiramente parabólica). Considerando que $X = x$ está fixado (condicionado), veja que o tamanho típico dos desvios de Y em torno de sua média condicional $\mathbb{E}(Y|X = x)$ é aproximadamente constante em x, não muda se variarmos x.

$y\|x$	$x=0$	$x=1$	$x=2$
$y=0$	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).

- $\sigma(X|Y=0) = \sqrt{\mathbb{V}(X|Y=0)} \approx 1$ (i). A faixa de variação de X dado que $Y=0$ vai de 0 a 4, aproximadamente, sendo que $\mathbb{E}(X|Y=0) \approx 2$ (segundo item acima). Assim, o desvio em relação a este valor esparado (condicionado em $Y=0$) deve ser com certeza menor que 2. As opções $\sigma(X|Y=0)=1$ ou 0.1 são as únicas disponíveis com um valor menor que 2 e a opção 0.1 não é razoável, muito pequena.

---

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 $(Y|X=1)$.

Solução: Temos $\mathbb{P}(X=x) = \sum_y \mathbb{P}(X=x, Y=y)$. Assim,

$$\mathbb{P}(X=x) = \begin{cases} 0.1+0.1+0+0.05, & \text{para } x=0 \\ 0.05+0+0.2+0.1, & \text{para } x=1 \\ 0.05+0.2+0.1+0.05, & \text{para } x=2 \end{cases}$$

Ou seja,

$$\begin{aligned} \mathbb{P}(X=0) &= 0.25 \\ \mathbb{P}(X=1) &= 0.35 \\ \mathbb{P}(X=2) &= 0.40 \end{aligned}$$

Para a distribuição condicional, temos $\mathbb{P}(Y=y|X=1) \propto \mathbb{P}(X=1, Y=y)$. Assim,

$$\begin{aligned} \mathbb{P}(Y=0|X=1) &\propto 0.05 \\ \mathbb{P}(Y=1|X=1) &\propto 0 \\ \mathbb{P}(Y=2|X=1) &\propto 0.2 \\ \mathbb{P}(Y=3|X=1) &\propto 0.1 \end{aligned}$$

Agora, basta normalizar estes valores para que eles somem 1 e assim encontrar a distribuição condicional. Como a soma é $0.05+0+0.2+0.1=0.35$, teremos

$$\begin{aligned} \mathbb{P}(Y=0|X=1) &= 0.05/0.35 \\ \mathbb{P}(Y=1|X=1) &= 0 \\ \mathbb{P}(Y=2|X=1) &= 0.2/0.35 \\ \mathbb{P}(Y=3|X=1) &= 0.1/0.35 \end{aligned}$$

---

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.

Solução: A soma dos autovalores é $\sum_i \lambda_i = 6+4+1+0.1+0.01 = 11.11$. O primeiro autovalor sozinho é responsável por $6/11.11 \approx 50\%$ da variabilidade total. Os dois primeiros autovalores respondem por $10/11.11 \approx 90\%$ da variabilidade total. Assim, usar os dois primeiros autovetores para reduzir a dimensionalidade é a decisão indicada.

---

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

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

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?

Solução: Basta mostrar que $\mathbf{\Sigma v} = \lambda \mathbf{v}$ onde λ é um valor real e positivo. De fato, multiplicação matricial elementar produz

$$\mathbf{\Sigma v} = (2\rho + 1)\mathbf{v}$$

e portanto $\mathbf{v}$ é autovetor com autovalor igual a $(2\rho + 1)$.

---

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 $\mathbf{x} = \mathbf{v}_1$.

Solução: Esta é uma prova ligeiramente diferente daquela do livro. Costuma existir mais de uma maneira de se provar um resultado matemático. Pelo teorema espectral, podemos fatorar (ou decompor) a matriz $\mathbf{B}$ como $\mathbf{B} = \mathbf{P}\ \mathbf{\Lambda}\ \mathbf{P}'$ onde $\mathbf{P}$ é uma matriz ortogonal (isto é, $\mathbf{P}'\ \mathbf{P} = \mathbf{I} = \mathbf{P}\ \mathbf{P}'$) com suas colunas formadas pelos autovetores de $\mathbf{B}$ e com $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_p)$ onde $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_1 > 0$. Assim, podemos escrever

$$\begin{aligned}
\frac{\mathbf{x}'\mathbf{B}\mathbf{x}}{\mathbf{x}'\mathbf{x}} &= \frac{\mathbf{x}'\mathbf{B}\mathbf{x}}{||\ \mathbf{x}\ ||^2} \\
&= \frac{\mathbf{x}'}{||\mathbf{x}||}\mathbf{B}\frac{\mathbf{x}}{||\mathbf{x}||} \\
&= \mathbf{y}'\ \mathbf{B}\ \mathbf{y} \qquad \text{onde } \mathbf{y} = \mathbf{x}/||\mathbf{x}|| \text{ tem norma 1} \\
&= \mathbf{y}'\ (\mathbf{P}\ \mathbf{\Lambda}\ \mathbf{P}')\ \mathbf{y} \\
&= (\mathbf{P}'\mathbf{y})'\ \mathbf{\Lambda}\ (\mathbf{P}'\mathbf{y}) \\
&= \mathbf{z}'\ \mathbf{\Lambda}\ \mathbf{z} \qquad \text{onde } \mathbf{z} = \mathbf{P}'\mathbf{y} \\
&= \sum_i \lambda_i z_i^2 \qquad \text{já que } \mathbf{\Lambda} \text{ é diagonal} \\
&\leq \sum_i \lambda_1 z_i^2 \qquad \text{já que } \lambda_i \leq \lambda_1 \\
&= \lambda_1 \sum_i z_i^2 \\
&= \lambda_1 ||\mathbf{z}||^2 \\
&= \lambda_1 \qquad \text{pois } ||\mathbf{z}||^2 = \mathbf{z}'\mathbf{z} = \mathbf{y}'\mathbf{P}\mathbf{P}'\mathbf{y} = \mathbf{y}'\mathbf{y} = 1
\end{aligned}$$

.

Além disso, se tomarmos $\mathbf{x} = \mathbf{v}_1$, teremos

$$\frac{\mathbf{x}'\mathbf{B}\mathbf{x}}{\mathbf{x}'\mathbf{x}} = \frac{\mathbf{v}_1'\mathbf{B}\mathbf{v}_1}{\mathbf{v}_1'\mathbf{v}_1} = \frac{\mathbf{v}_1'\mathbf{B}\mathbf{v}_1}{1} = \mathbf{v}_1'\lambda_1\mathbf{v}_1 = \lambda_1||\mathbf{v}_1||^2 = \lambda_1\ .$$

---

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 $\mathbb{V}(Y) = \lambda_1$.

   Solução: Seja $Y = \mathbf{l}'\mathbf{X}$ onde $||\mathbf{l}|| = 1$. Então

$$\mathbb{V}(Y) = \mathbf{l}'\mathbf{\Sigma}\mathbf{l} = \frac{\mathbf{l}'\mathbf{\Sigma}\mathbf{l}}{1} = \frac{\mathbf{l}'\mathbf{\Sigma}\mathbf{l}}{||\mathbf{l}||} .$$

   Pelo resultado anterior, sabemos então que $\mathbb{V}(Y) \leq \lambda_1$ para qualquer escolha para $||\mathbf{l}||$ e que este limite (ou cota) superior é atingido se escolhermos $\mathbf{l} = \mathbf{v}_1$. Assim,

$$\mathbb{V}(Y) = \mathbb{V}(\mathbf{v}_1'\mathbf{X}) = \mathbf{v}_1'\mathbf{\Sigma}\mathbf{v}_1 = \mathbf{v}_1'\lambda_1\mathbf{v}_1 = \lambda_1||\mathbf{v}_1||^2 = \lambda_1 .$$

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. Use a abordagem de componente principal para encontrar: (ii) a matriz de carga de fator, (iii) as comunalidades, (iv) as variâncias específicas.

   Solução: Pelo teorema espectral, a matriz 3×3 de covariância $\mathbf{\Sigma}$ pode ser fatorada como

$$\begin{aligned} \mathbf{\Sigma} &= \left[\begin{array}{c|c|c} 1/2 & 1/2 & 1/\sqrt{2} \\ 1/2 & 1/2 & -1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} & 0 \end{array}\right] \begin{bmatrix} 2.25 & 0 & \\ 0 & 1.96 & 0 \\ 0 & 0 & 0.16 \end{bmatrix} \left[\begin{array}{c|c|c} 1/2 & 1/2 & 1/\sqrt{2} \\ 1/2 & 1/2 & -1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} & 0 \end{array}\right]' \\ &\approx \begin{bmatrix} 1.13 & 0.97 & 0.10 \\ 0.97 & 1.13 & 0.1 \\ 0.10 & 0.10 & 2.1 \end{bmatrix} \qquad \text{aprox com duas casas decimais} \end{aligned}$$

   Usando apenas o primeiro componente principal, vamos obter uma aproximação para a matriz de covariância seguindo as notas de aula:

$$\mathbf{\Sigma} \approx \begin{bmatrix} \sqrt{2.25}/2 \\ \sqrt{2.25}/2 \\ \sqrt{2.25}/\sqrt{2} \end{bmatrix} \times \begin{bmatrix} \sqrt{2.25}/2 & \sqrt{2.25}/2 & \sqrt{2.25}/\sqrt{2} \end{bmatrix} \approx \begin{bmatrix} 0.56 & 0.56 & 1.59 \\ 0.56 & 0.56 & 1.59 \\ 1.59 & 1.59 & 4.50 \end{bmatrix}$$

   Esta é uma aproximação pobre para $\mathbf{\Sigma}$ pois o segundo autovalor não é pequeno. Mas este é apenas um exercício numérico de exame para checar o entendimento das definições do modelo fatorial.

   No modelo fatorial (ver definição no slide 26 das notas de aula), temos:

$$\mathbb{V}(X_i) = \underbrace{||\mathbf{l}_i||^2}_{\text{comunalidade}} + \underbrace{\psi_i}_{\text{variância específica}}$$

onde $\mathbf{l_i}$ é a linha i da matriz de cargas $\mathbf{L}$. Como temos apenas um fator neste exemplo, a matrriz de cargas $\mathbf{L}$ é uma matriz-coluna 3×1: $\mathbf{L} = \sqrt{2.25}\mathbf{v}_1$. Portanto, as comunalidades são simplesmente os quadrados dos três elementos de $\mathbf{L}$. A variância específica é o resto que falta para completar $\mathbb{V}(X_i)$. Assim:

$$\begin{aligned}
1.13 = \mathbb{V}(X_1) &= \overbrace{(\sqrt{2.25}/2)^2}^{\text{com.}} + \overbrace{\psi_1}^{\text{var. esp.}} \implies \psi_1 = 0.57 \\
1.13 = \mathbb{V}(X_1) &= (\sqrt{2.25}/2)^2 + \psi_2 \implies \psi_2 = 0.57 \\
2.10 = \mathbb{V}(X_1) &= (\sqrt{2.25}/\sqrt{2})^2 + \psi_3 \implies \psi_3 = 0.98
\end{aligned}$$

O problema da prova termina aqui. Mas, aproveitando o embalo, veja que a igualdade matricial do teorema espectral, $\mathbf{\Sigma} = \mathbf{P}\ \mathbf{\Lambda}\ \mathbf{P}'$, pode ser escrita de outra forma, usando as colunas-autovetores de $\mathbf{P}$:

$$\mathbf{\Sigma} = \lambda_1 \underbrace{\mathbf{v}_1}_{3\times1} \underbrace{\mathbf{v}_1'}_{1\times3} + \lambda_2 \underbrace{\mathbf{v}_2}_{3\times1} \underbrace{\mathbf{v}_2'}_{1\times3} + \lambda_3 \underbrace{\mathbf{v}_3}_{3\times1} \underbrace{\mathbf{v}_3'}_{1\times3}$$

Na expressão acima, note que aparece o produto $\mathbf{v}\mathbf{v}'$, que é uma matriz 3×3, e não o produto interno $\mathbf{v}'\mathbf{v}$, que é um escalar(um número real, 1×1). Como os autovetores possuem comprimento 1, as matrizes tem uma tamanho controlado, pequeno, Se o autovalor λ_i for pequeno, próximo de zero, podemos jogar fora os termos envolvendo os autovalores pequenos.

---

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 $2x_1 + x_2 \leq \log(18)$.

   Solução: A região ótima R_1 de classificação na população 1 (no sentido de minimizar o erro de classificação errada) é dada pelos pontos $\mathbf{x} = (x_1, x_2) \in \mathbb{R}^+ \times \mathbb{R}^+$ tais que

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{c(1|2)}{c(2|1)} \frac{\pi_2}{\pi_1} = 3 \times \frac{0.1}{0.9} = \frac{1}{3} \, .$$

   Mas

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} = \frac{6\exp(-(3x_1 + 2x_2))}{\exp(-(x_1 + x_2))} = 6\exp(-2x_1 - x_2) \, .$$

   Portanto, a região R_1 é dada pelos pontos $\mathbf{x}$ tais que

$$6\exp(-2x_1 - x_2) \geq \frac{1}{3} \quad \rightarrow \quad -2x_1 - x_2 \geq \frac{1}{18} \quad \rightarrow \quad 2x_1 + x_2 \leq \log(18)$$
