Segunda Prova
Fundamentos Estatísticos para Ciência dos Dados

12/05/2016

1. Esboce aproximadamente o vetor-direção do primeiro componente principal no gráfico a esquerda na Figura 1 ignorando os dois tipos de símbolo para os pontos. No gráfico da direita, esboce aproximadamente o vetor direção do (primeiro) discriminante linear de Fisher considerando as duas classes para os pontos.

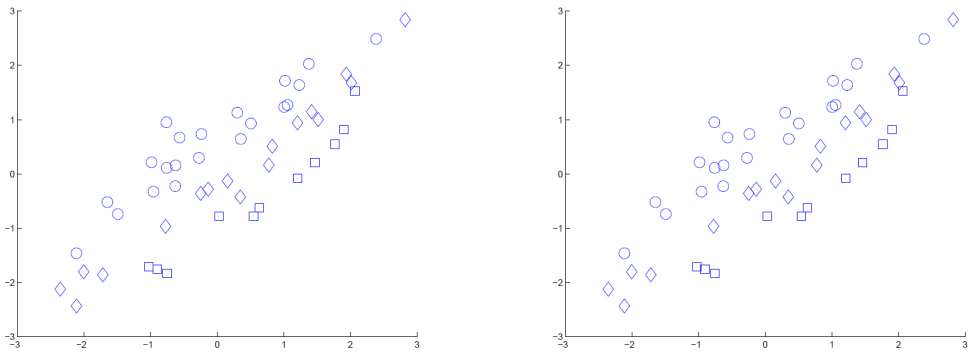

Figura 1: Desenhe o primeiro componente principal (esquerda) e o primeiro discriminante linear (direita).

2. Todo objeto pertence a uma de duas classes ou populações, π_1 ou π_2. Suponha que a classe π_1 é extremamente rara (isto é, $\mathbb{P}(\in \pi_1) \approx 0$).

   - Isto significa que se classificarmos todo novo objeto na população 2 teremos um pequeno número de erros, supondo que os custos dos dois erros sejam iguais. Mostre que, com esta regra de classificação radical temos
   $$\mathbb{P}(\text{erro}) \approx 0$$
   - Podemos tentar colocar as duas classes em pé de igualdade considerando a seguinte quantidade:
   $$\eta = \mathbb{P}(\text{class } 2| \in 1) + \mathbb{P}(\text{class } 1| \in 2)$$
   Mostre que $\eta = 1$ com a regra de classificação radical
   - Mostre que a regra que minimiza η é aquela em que a região R de classificação é dada por
   $$R = \{\mathbf{x} \text{ tais que } f_2(\mathbf{x}) < f_1(\mathbf{x})\}$$

3. Considere um vetor aleatório bi-dimensional $\mathbf{X} = (X_1, X_2)' \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Marque V ou F justificando sua resposta:

   - Se X_1 e X_2 são independentes, a nuvem de pontos $(x_{i1}, x_{i2}), i = 1,\ ldots, n$ com as instâncias observadas e a curva de nível $f(\mathbf{x}) = c$ com a densidade tem necessariamente a forma de um círculo.
   - Se X_1 e X_2 são independentes, a nuvem de pontos $(x_{i1}, x_{i2}), i = 1,\ ldots, n$ com as instâncias observadas e a curva de nível $f(\mathbf{x}) = c$ com a densidade tem de ter a forma elíptica com os eixos principais da elipse alinhados com os eixos das abscissas e ordenadas (SE NECESSÁRIO, considere um círculo como uma forma especial de uma elipse)

- A matriz de covariância $\mathbf{\Sigma}$ é simétrica apenas se X_1 e X_2 são independentes.
- A matriz de covariância $\mathbf{\Sigma}$ é diagonal se X_1 e X_2 são independentes.
- A matriz de covariância $\mathbf{\Sigma}$ é diagonal se X_1 e X_2 possuem correlação positiva.

4. A Figura 2 mostra dados bi-dimensionais extraídos de uma distribuição normal multivariada.

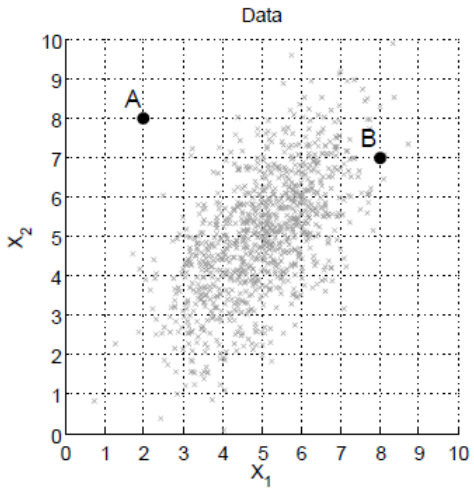

Figura 2: Dados bi-dimensionais extraídos de uma distribuição normal multivariada.

- Qual é o valor esperadode cada variável? Estime a resposta visualmente e arredonde para o número inteiro mais próximo.
- O valor do elemento $(1, 2)$ da matriz de covariância $\mathbf{\Sigma}$ é positivo, negativo ou zero?
- Defina $\mathbf{e}_1$ e $\mathbf{e}_2$ como as direções do primeiro e segundo componentes principais, ambos com comprimento igual a 1. Estas direções definem uma nova base do $\mathbb{R}^2$ onde cada ponto $\mathbf{x}$ é tranformado para $\mathbf{z} = (z_1, z_2)$ com

$$z_1 = (\mathbf{x} - \mu)'\mathbf{e}_1$$

e

$$z_2 = (\mathbf{x} - \mu)'\mathbf{e}_2$$

Esboce E ROTULE $\mathbf{e}_1$ e $\mathbf{e}_2$ na figura 2. Os vetores devempartir do centro da distribuição-nuvem de pontos.

- A covariância $Cov(Z_1, Z_2)$ é negativa, positiva ou aproximadamente zero?
- Como A e B estão mais ou menos a uma mesma distância do centro eles possuem igual chance de serem observados. V ou F? Justifique.
- A probabilidade de observar um ponto tão afastado de μ quanto B é aproximadamente 5%, 50% ou 95%?

5. Seja $\mathbf{X} = (X_1, X_2, X_3)' \sim N_3(\mathbf{0}, \mathbf{\Sigma})$ com

$$\mathbf{\Sigma} = \begin{bmatrix} 4 & 0 & 2 \\ 0 & 2 & 0 \\ 2 & 0 & 9 \end{bmatrix}$$

- Ache a distribuição de $Y = 3X_1 + 2X_2 + X_3$.
- Para que valores de a e b as variáveis $X_1 + aX_3$ e $X_1 + bX_3$ possuem covariância 0 e portanto são independentes.

6. Mostre que a matriz de covariãncia

$$\mathbf{\Sigma} = \begin{bmatrix} 1.0 & 0.63 & 0.45 \\ 0.63 & 1.0 & 0.35 \\ 0.45 & 0.35 & 1.0 \end{bmatrix}$$

do vetor $\mathbf{X} = (X_1, X_2, X_3)' \sim N_3(\mathbf{0}, \mathbf{\Sigma})$ pode ser gerada pelo modelo de análise fatorial com $m = 1$ fator e dado por

$$\begin{aligned} X_1 &= 0.9F_1 + \epsilon_1 \\ X_2 &= 0.7F_1 + \epsilon_2 \\ X_3 &= 0.5F_1 + \epsilon_3 \end{aligned}$$

com $var(F_1) = 1$, $Cov(F_1, \boldsymbol{\epsilon}) = \mathbf{0}$ e

$$\mathbf{\Psi} = Cov(\epsilon) = \begin{bmatrix} 0.19 & 0 & 0 \\ 0 & 0.51 & 0 \\ 0 & 0 & 0.75 \end{bmatrix}$$