Universidade Federal do Ceará
Centro de Tecnologia
Departamento de Engenharia de Teleinformática
Engenharia de Teleinformática

Lista 01

Aluno Kenneth Brenner dos Anjos Benício – 519189

Professor Guilherme de Alencar Barreto
Disciplina Reconhecimento de Padrões- TIP8311

Fortaleza, 2021

1) SOLUÇÃO:

O produto escalar não pode ser considerado uma medida de similaridade ou dissimilaridade válida, pois seu operador embora seja simétrico pode atingir sem restrições valores negativos. Ademais, quando o operador é aplicado sobre um mesmo vetor é possível obter um valor não nulo. Dessa forma, não há sentido em classificar o produto escalar sem restrições como uma medida de similaridade ou dissimilaridade.

A distância euclidiana é uma medida de dissimilaridade. As razões para tal são: O valor numérico para a distância será sempre positivo para quaisquer dois vetores, o valor da distância será nulo caso seja considerada a distância de um vetor para si mesmo e o valor numérico para a distância será simétrico para quaisquer dois vetores. Por fim, a distância euclidiana é uma medida que indica maior proximidade quanto menor for seu valor numérico, definindo assim uma formulação valida para uma medida de dissimilaridade.

2) SOLUÇÃO:

A expressão mais geral para a distância ponderada (ou distância de Mahalanobis) é dada pela seguinte equação

$$d_p(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})^{\mathrm{T}}\mathbf{Q}(\mathbf{x}-\mathbf{y})}, \{\mathbf{x},\mathbf{y}\} \in \mathbb{R}^P, \mathbf{Q} \in \mathbb{R}^{P\times P}. \tag{1}$$

É interessante analisar com um pouco mais de detalhes a estrutura da matriz $\mathbf{Q}$, pois será a responsável por realizar a ligação entre a distância ponderada e a distância euclidiana. Para que a expressão acima seja válida é necessário que $\mathbf{Q}$ seja uma matriz positiva semidefinida para que não existam valores negativos dentro da raiz quadrada. Por simplicidade iremos adotar a partir de agora o valor $P = 2$. Desse modo, ao abrirmos a expressão para esse caso obtemos o seguinte resultado

$$\mathbf{Q} = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix}, \tag{2}$$

$$d_p(\mathbf{x},\mathbf{y}) = \sqrt{\begin{bmatrix} x_1-y_1 & x_2-y_2 \end{bmatrix} \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix} \begin{bmatrix} x_1-y_1 \\ x_2-y_2 \end{bmatrix}}, \tag{3}$$

$$d_p(\mathbf{x},\mathbf{y}) = \sqrt{a_1(x_1-y_1)^2 + (a_2+a_3)(x_1-y_1)(x_2-y_2) + a_4(x_2-y_2)^2}, \tag{4}$$

para o caso no qual $\mathbf{Q}$ seja uma matriz identidade teremos que $a_1 = 1$, $a_2 = 0$, $a_3 = 0$ e $a_4 = 1$ de forma que assim obteremos a seguinte expressão

$$d_p(\mathbf{x},\mathbf{y}) = \sqrt{1(x_1-y_1)^2 + (0+0)(x_1-y_1)(x_2-y_2) + 1(x_2-y_2)^2}, \tag{5}$$

$$d_p(\mathbf{x},\mathbf{y}) = \sqrt{(x_1-y_1)^2 + (x_2-y_2)^2}, \tag{6}$$

$$d_p(\mathbf{x},\mathbf{y}) = d_e(\mathbf{x},\mathbf{y}). \tag{7}$$

Dessa forma, vemos que a distância ponderada pode ser reduzida à distância euclidiana para o caso no qual a matriz $\mathbf{Q}$ seja a identidade.

3) SOLUÇÃO:

- $g(\mathbf{x}, \mathbf{y}) = \exp\{-\gamma||\mathbf{x} - \mathbf{y}||^2\}$
  (a) $g(\mathbf{x}, \mathbf{y}) \geq 0$: É uma função exponencial, logo seus valores serão sempre iguais a zero ou positivos.
  (b) $g(\mathbf{x}, \mathbf{x}) = 0$: Uma vez que essa função associa dois vetores por meio de uma subtração simples dentro da exponencial o termo que multiplicara γ será nulo para quaisquer vetores desde que ambos sejam exatamente iguais.
  (c) $g(\mathbf{x}, \mathbf{y}) = g(\mathbf{y}, \mathbf{x})$: Uma vez que os vetores são associados por uma subtração ao quadrado podemos afirmar que a operação será sim simétrica.
- $g(\mathbf{x}, \mathbf{y}) = \frac{1}{1+||\mathbf{x}-\mathbf{y}||^2}$
  (a) $g(\mathbf{x}, \mathbf{y}) \geq 0$: O termo capaz de alterar o sinal da função em analise é a subtração entre os vetores presente no denominador. Entretanto, tal termo esta sendo elevado ao quadrado de modo que não é possível que essa subtração assuma valores negativos em algum momento. Portanto, a função terá sempre valores iguais a zero ou positivos.
  (b) $g(\mathbf{x}, \mathbf{x}) = 0$: De forma semelhante ao caso anterior, temos dois vetores associados por uma subtração simples de tal modo que o resultado da operação será nulo para dois vetores exatamente iguais. Entretanto, o termo nulo presente no denominador está sendo adicionado de uma constante não nula. Dessa forma, para vetores exatamente iguais teremos $g(\mathbf{x}, \mathbf{x}) = 1$.
  (c) $g(\mathbf{x}, \mathbf{y}) = g(\mathbf{y}, \mathbf{x})$: De forma semelhante ao caso anterior, o índice ao quadrado que associa os dois vetores é o que garante a simetria da função.

Portanto, temos que a primeira função será uma medida de dissimilaridade válida enquanto a segunda será uma medida de dissimilaridade não valida.

4) SOLUÇÃO:

O objetivo é obter uma estimação para a classe do vetor $\mathbf{x}$ considerando os três diferentes vetores centróides apresentados abaixo

$$\mathbf{x} = \begin{bmatrix} 80.07 & 48.07 & 58.40 & 32.01 \end{bmatrix}^{\mathrm{T}}, \tag{8}$$

$$\mathbf{m}_s = \begin{bmatrix} 51.69 & 12.82 & 43.54 & 38.86 \end{bmatrix}^{\mathrm{T}}, \tag{9}$$

$$\mathbf{m}_c = \begin{bmatrix} 71.51 & 20.75 & 64.11 & 50.77 \end{bmatrix}^{\mathrm{T}}, \tag{10}$$

$$\mathbf{m}_h = \begin{bmatrix} 47.64 & 17.40 & 35.46 & 30.24 \end{bmatrix}^{\mathrm{T}}. \tag{11}$$

Para escolhermos a qual classe $\mathbf{x}$ pertence é necessário antes de tudo obter os valores de distância do referido vetor para os centróides disponíveis. Para isso, será utilizado a definição de distância euclidiana por mera conveniência. Desse modo, temos

$$d_e(\mathbf{x}, \mathbf{m}_s) = \sqrt{(80.07 - 51.69)^2 + (48.07 - 12.82)^2 + (58.40 - 43.54)^2 + (32.01 - 38.86)^2} = 48.12,$$

$$d_e(\mathbf{x}, \mathbf{m}_c) = \sqrt{(80.07 - 71.51)^2 + (48.07 - 20.75)^2 + (58.40 - 64.11)^2 + (32.01 - 50.77)^2} = 34.70,$$

$$d_e(\mathbf{x}, \mathbf{m}_h) = \sqrt{(80.07 - 47.64)^2 + (48.07 - 17.40)^2 + (58.40 - 35.46)^2 + (32.01 - 30.24)^2} = 50.22.$$

Dessa forma, é provável que o vetor $\mathbf{x}$ pertença à classe definida pelo centróide $\mathbf{m}_s$, pois esse é o centróide de menor distância ao vetor $\mathbf{x}$.

5) SOLUÇÃO:

Antes de prosseguir é necessário normalizar os dados para que seja possível aplicar o classificador de máxima correlação. Para realizar a normalização dos vetores é necessário antes de mais nada dividir cada elemento do vetor original por sua norma. A norma de um vetor $\mathbf{x} \in \mathbb{R}^P$ pode ser obtida sem maiores complicações por meio da equação

$$||\mathbf{x}|| = \sqrt{\sum_{i=1}^{P} x_i^2}. \tag{12}$$

Sendo assim, temos

$$\overline{\mathbf{x}} = \frac{\mathbf{x}}{||\mathbf{x}||} = \frac{\mathbf{x}}{114.70} = \begin{bmatrix}0.70 & 0.42 & 0.51 & 0.28\end{bmatrix}^{\mathrm{T}}, \tag{13}$$

$$\overline{\mathbf{m}}_s = \frac{\mathbf{m}_s}{||\mathbf{m}_s||} = \frac{\mathbf{m}_s}{79.01} = \begin{bmatrix}0.65 & 0.16 & 0.55 & 0.49\end{bmatrix}^{\mathrm{T}}, \tag{14}$$

$$\overline{\mathbf{m}}_c = \frac{\mathbf{m}_c}{||\mathbf{m}_c||} = \frac{\mathbf{m}_c}{110.60} = \begin{bmatrix}0.65 & 0.19 & 0.58 & 0.46\end{bmatrix}^{\mathrm{T}}, \tag{15}$$

$$\overline{\mathbf{m}}_h = \frac{\mathbf{m}_h}{||\mathbf{m}_h||} = \frac{\mathbf{m}_h}{68.88} = \begin{bmatrix}0.69 & 0.25 & 0.51 & 0.44\end{bmatrix}^{\mathrm{T}}. \tag{16}$$

Agora nos resta apenas obter o produto escalar do vetor $\overline{\mathbf{x}}$ com os vetores centróides normalizados. O produto escalar entre dois vetores $\{\mathbf{x}, \mathbf{y}\} \in \mathbb{R}^P$ é definido pela seguinte expressão

$$\mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^{P} x_i y_i. \tag{17}$$

Portanto, temos

$$\overline{\mathbf{x}} \cdot \overline{\mathbf{m}}_s = 0.94, \tag{18}$$

$$\overline{\mathbf{x}} \cdot \overline{\mathbf{m}}_c = 0.95, \tag{19}$$

$$\overline{\mathbf{x}} \cdot \overline{\mathbf{m}}_h = 0.97. \tag{20}$$

Visto que o processo é baseado em uma medida de similaridade temos que, segundo o classificador de máxima correlação, o vetor $\mathbf{x}$ pertencerá à classe definida pelo centróide $\mathbf{m}_h$. É interessante notar como os resultados de classificação podem diferir mesmo quando se utiliza o mesmo conjunto de dados, mudando apenas o tipo do classificador.