Problema 1

Considere o conjunto de dados da tabela, em que y representa a classe de saída, e a_k os atributos discretos.

a_1	a_2	a_3	y
Z	B	2	0
Z	X	1	0
A	B	2	1
A	B	2	1
A	X	2	0
A	X	2	0
Z	B	1	0
Z	X	2	1

Derive a árvore de classificação para este problema utilizando o algoritmo ID3. Apresente todos os cálculos e determine a percentagem de erros de classificação para a árvore obtida.

Problema 2

Considere o seguinte conjunto de dados em $\mathbb{R}^2$

$$\mathcal{T} = \{(-5,-5),(5,5),(-1,1),(1,-1)\}$$

a) Determine os valores próprios e as direcções principais destes dados.

b) Admita que pretende efectuar uma projecção dos dados $\mathcal{T}$ em $\mathbb{R}^1$ de forma a que, a partir desta projecção, seja possível recuperar os dados originais segundo um critério de erro quadrático mínimo (problema de compressão de dados). Como sabe, este resultado pode ser obtido determinando a primeira componente principal dos dados $\mathcal{T}$.

   i) Determine o valor da primeira componente principal relativa a cada uma das observações de $\mathcal{T}$.

   ii) Determine a reconstrução óptima das observações originais a partir das projecções determinadas em i).

   iii) Indique qual a variância do erro de reconstrução.

Problema 3

Admita que tem um conjunto X de padrões (vectores de dimensão n) cuja densidade de probabilidade $p(\mathbf{x})$ pretende estimar.

Indique como poderia utilizar uma rede neuronal para estimar esta densidade de probabilidade.

Problema 4

Em problemas de Aprendizagem Automática é frequente a utilização de métricas euclideanas. O quadrado da distância entre dois pontos é definido por

$$D^2(\mathbf{x}_1, \mathbf{x}_2) = (\mathbf{x}_1 - \mathbf{x}_2)^T(\mathbf{x}_1 - \mathbf{x}_2)$$

No entanto, outras distâncias são frequentemente utilizadas na prática. Uma família que observa as propriedades associadas a métricas e que inclui a distância euclideana como caso particular é aquela em que o quadrado da distância entre dois pontos é definido por

$$D^2(\mathbf{x}_1, \mathbf{x}_2) = (\mathbf{x}_1 - \mathbf{x}_2)^T\mathbf{G}\ (\mathbf{x}_1 - \mathbf{x}_2) \qquad (1)$$

onde $\mathbf{G}$ é uma matriz simétrica definida positiva.

Pode mostrar-se que usar a distância definida com a métrica (1) é equivalente a realizar uma transformação linear dos vectores de entrada $\mathbf{x}$ e calcular a distância euclideana no espaço transformado. A transformação que conduz a este resultado é definida por $\mathbf{y} = \mathbf{A}\mathbf{x}$, com $\mathbf{A}\mathbf{A} = \mathbf{G}$. Note que se $\mathbf{G}$ for simétrica definida positiva, existe sempre uma matriz $\mathbf{A}$, igualmente simétrica definida positiva, tal que $\mathbf{G} = \mathbf{A}\mathbf{A}$ (por este motivo, utiliza-se frequentemente a notação $\mathbf{A} = \mathbf{G}^{1/2}$).

a) Em aplicações práticas, pode ser útil escolher $\mathbf{G}$ de acordo com a estrutura dos dados de entrada. Um caso particular corresponde à chamada distância de Mahalanobis, onde a matriz G é definida pela inversa da matriz de covariância dos dados de entrada, ou seja, $G = \mathbf{R}_{xx}^{-1}$, com

$$\mathbf{R}_{xx} = E[(\mathbf{x} - \overline{\mathbf{x}})(\mathbf{x} - \overline{\mathbf{x}})^T]$$

onde $E[]$ representa a média estatística e $\overline{\mathbf{x}} = E[\mathbf{x}]$.

Determine a expressão da matriz de covariância dos dados após a transformação linear $\mathbf{A}$ no caso particular da distância de Mahalanobis.

b) Condidere um conjunto de treino num espaço bidimensional, constituído por quatro padrões:

$$\mathcal{T} = \{(-10, 1), (-10, -1), (10, 1), (10, -1)\}$$

i) Indique a melhor partição do espaço em dois agrupamentos que seria possível obter com o algoritmo k-médias, admitindo que é utilizada uma distância eudclideana. Justifique a sua resposta sem realizar cálculos.

ii) Embora a distância de Mahalanobis seja frequentemente vantajosa, no caso particular que se está a considerar a sua utilização não permitiria definir qual a partição óptima dos dados. Justifique.

Problema 5

Considere o seguinte conjunto de dados unidimensionais:

$$\mathcal{T} = \{-4, -3, 0, 1\}$$

Admita que se pretende aplicar o algoritmo EM com duas gaussianas a estes dados. Admita que a estimativa da fdp é incializada com

$$p(x) = \pi_0 p_1(x) + \pi_1 p_2(x)$$

sendo:

$$\pi_0 = \pi_1 = 0,5$$

$$p_1(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x+4)^2}{2}\right)$$

$$p_2(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-1)^2}{2}\right)$$

Indique qual a estimativa da fdp que seria obtida após uma iteração do algoritmo EM.