UFABC	INF-317B - Tópicos em Inteligência Artificial: Machine Learning
Professores: André Kazuo Takahata e Ricardo Suyama	
Quadrimestre: 2020.1	
Data: 19/03/2020	
Entrega pelo TIDIA4: 02/04/2020	

Exercícios Individuais para Envio pelo TIDIA4

Instruções

1) Para as questões abaixo considere como *d4d3d2d1* os 4 últimos dígitos menos significativos de seu RA. Por exemplo, se o RA=2120201567 então *d4*=1, *d3*=5 *d2*=6 e *d1*=7. Caso não possua RA então utilize 4 números gerados aleatoriamente (utilize os mesmos números no decorrer da lista).
2) Fazer as questões **à mão**, com exceção daquelas marcadas com "**(Computador)**", e enviar a resolução escaneada. Caso contrário elas **não serão consideradas**.
3) Os exercícios são **individuais**. Questões com plágio ou exercícios em duplas **não serão aceitos.**

Questões

1) Considere que uma loja está fazendo uma promoção com o uso da roleta mostrada na Figura 1. A roleta é dividida em 8 setores circulares iguais e a cada setor é associada a um desconto. Na promoção, ao realizar uma compra, o cliente gira a seta e ganha o desconto correspondente ao setor que a seta aponta quando cessa o movimento. Caso a seta aponte para o setor marcado com a interrogação, o cliente ganha *d1*% de desconto.
   a. Seja X uma variável aleatória que representa o desconto que um cliente irá obter em uma compra de R$100,00 e θ o ângulo entre a seta e a vertical conforme mostrado na Figura 1. Dê a descrição de X em função de θ.
   b. Considere que a seta para em qualquer ângulo θ com a mesma probabilidade. Dê a PMF de X.

c. Quais são os valores de E[X] e var(X)?

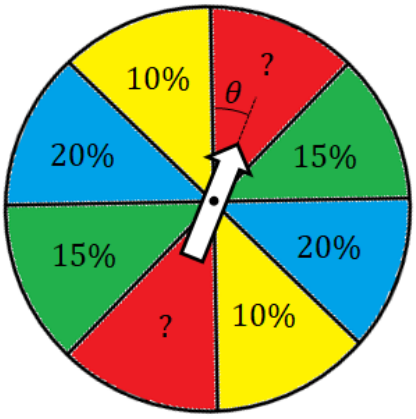

Figura 1: Roleta utilizada na Questão 1.

2) Considere o problema de classificação unidimensional com duas classes em que:

$$p(x|\omega_1) \backsim \mathcal{N}(d1,1)$$
$$p(x|\omega_2) \backsim \mathcal{N}(d1+2,1)$$
$$\Pr(\omega_1) = \Pr(\omega_2) = 1/2$$

a. **(Computador)** Trace os gráficos de $p(x|\omega_i)$ e $p(x) \cdot p(x|\omega_i)$ para ambas classes

b. Obtenha a regra de decisão ótima analiticamente e com o uso do gráfico obtido no item a)

c. **(Computador)** Simule a classificação de $N = 1000$ objetos. Qual é o erro de classificação obtido?

3) Considere o problema de classificação unidimensional com duas classes em que:

$$p(x|\omega_1) \backsim \mathcal{N}(d2,1)$$
$$p(x|\omega_2) \backsim \mathcal{N}(d2,4)$$
$$\Pr(\omega_1) = \Pr(\omega_2) = 1/2$$

a. **(Computador)** Trace os gráficos de $p(x|\omega_i)$ e $p(x) \cdot p(x|\omega_i)$ para ambas classes

b. Obtenha a regra de decisão ótima analiticamente e com o uso do gráfico obtido no item a)

c. **(Computador)** Simule a classificação de $N = 1000$ objetos. Qual é o erro de classificação obtido?

4) Considere o problema de classificação unidimensional com duas classes em que que x é uma grandeza obtida por meio de um exame e ω_1 e ω_2 representam, respectivamente, pacientes saudáveis e doentes tal que:

$$p(x|\omega_1) \backsim \mathcal{N}(d2,1)$$
$$p(x|\omega_2) \backsim \mathcal{N}(d2+3,1)$$
$$\Pr(\omega_1) = 0{,}99$$
$$\Pr(\omega_2) = 0{,}01.$$

a. Obtenha a regra de decisão ótima considerando que os falsos positivos e os falsos negativos possuem o mesmo erro.

b. **(Computador)** Simule a classificação de $N = 10000$ pacientes. Preencha a tabela de contingência abaixo e dê os valores de verdadeiros positivos (TP), verdadeiros negativos (TN), falsos positivos (FP) e falsos negativos (FN). Também dê a acurácia, a sensibilidade e a especificidade do teste.

	Classe verdadeira	
	ω_1(-)	ω_2(+)
Resultado +		
Resultado -		

c. Desenhe um diagrama de Venn ilustrando as informações obtidas no Item b).

d. Caso um paciente receba resultado positivo, qual é a probabilidade de ele estar realmente doente?

e. Caso um paciente receba um resultado negativo, qual é a probabilidade de ele estar realmente saudável?

f. Refaça os itens a) a d) admitindo que a perda associada ao falso negativo é 100 vezes pior que a perda associada ao falso positivo.

5) Considere o problema de classificação bidimensional em que o vetor de características é tal que $\mathbf{x} = [x_1, x_2]^{\mathrm{T}}$ e as classes são representadas por ω_1 e ω_2.

   a. Obtenha a fronteira de decisão ótima para

   $$p(\mathbf{x}|\omega_1) \backsim \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{I})$$
   $$p(\mathbf{x}|\omega_2) \backsim \mathcal{N}(\boldsymbol{\mu}_2, \mathbf{I})$$
   $$\mathrm{P}(\omega_1) = \mathrm{P}(\omega_2) = 1/2$$

   em que $\boldsymbol{\mu}_1 = [d1, d2]^{\mathrm{T}}$ e $\boldsymbol{\mu}_2 = [d3, d4]^{\mathrm{T}}$.

   b. Repita o item a. modificando as probabilidades *a priori* para $\mathrm{P}(\omega_1) = 1/3$ e $\mathrm{P}(\omega_2) = 2/3$.

   c. Repita o item a. agora alterando a matriz de covariância de modo que as variâncias de x_1 e x_2 sejam diferentes e haja correlação entre x_1 e x_2. Qual é o valor da correlação no caso? A matriz de covariância deve ser igual para ambas classes.

   d. **(Computador)** Para os Itens a. a d., simule a classificação de $N = 1000$ objetos. Faça um gráfico mostrando os dados gerados e a fronteira de decisão para cada caso. Qual é o erro de classificação obtido?

   e. **(Computador)** Utilize o classificador obtido em a. para os dados gerados nos Itens b. e c. O que ocorre com o erro de classificação?

6) Considere o problema de classificação bidimensional em que o vetor de características é tal que $\mathbf{x} = [x_1, x_2]^{\mathrm{T}}$ e as classes são representadas por ω_1 e ω_2.

   a. Obtenha a fronteira de decisão ótima para

   $$p(\mathbf{x}|\omega_1) \backsim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$$
   $$p(\mathbf{x}|\omega_2) \backsim \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$$
   $$\mathrm{P}(\omega_1) = \mathrm{P}(\omega_2) = 1/2$$

   em que

   $$\boldsymbol{\mu}_1 = [0{,}0]^{\mathrm{T}}, \boldsymbol{\mu}_2 = [0{,}4]^{\mathrm{T}},$$
   $$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 20 + d1 & 0 \\ 0 & 1 \end{bmatrix} \text{ e } \boldsymbol{\Sigma}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Para isso obtenha os pontos $[x_1, x_2]^{\mathrm{T}}$ que satisfaçam a

$$\mathbf{x}^{\mathrm{T}}\mathbf{W}_1\mathbf{x} + \mathbf{w}_1^{\mathrm{T}}\mathbf{x} + w_{1,0} = \mathbf{x}^{\mathrm{T}}\mathbf{W}_2\mathbf{x} + \mathbf{w}_2^{\mathrm{T}}\mathbf{x} + w_{2,0}$$

em que

$$\mathbf{W}_i = -\frac{1}{2}\mathbf{\Sigma}_i^{-1}, \mathbf{w}_i = \mathbf{\Sigma}_i^{-1}\boldsymbol{\mu}_i$$

$$w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^{\mathrm{T}}\mathbf{\Sigma}_i^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\ln|\mathbf{\Sigma}_i| + \ln P(\omega_i)$$

para $i = 1,2$.

b. **(Computador)** Simule a classificação de $N = 2000$ objetos. Faça um gráfico mostrando os dados gerados e a fronteira de decisão. Qual é o erro de classificação obtido?

c. **(Computador)** Para o mesmo conjunto de dados utilizado no Item b., obtenha o erro de classificação para o classificador de mínima distância. Compare o desempenho obtido com o do classificador utilizado no Item b.

7) **(Computador)** Escolha o classificador mais adequado entre análise de discriminante linear (LDA) ou quadrática (QDA) para cada situação apresentada nos arquivos dataset1.txt a dataset6.txt e faça testes com uso de validação cruzada no classificador escolhido. Não é necessário calcular a fronteira de decisão explicitamente (mas pode ser feito caso ache necessário).

8) **(Computador)** Considere o arquivo dataset2.txt e o uso do LDA para classificação. Faça a validação cruzada com a subamostragem aleatória com 10 repetições considerando:

   a. Caso 1: 5 vetores para treinamento

   b. Caso 2: 20 vetores para treinamento

Faça um gráfico para cada caso sobrepondo as fronteiras de decisão obtidas em cada repetição da subamostragem aleatória. Qual é a mudança de comportamento observada?

Dica: O comando `mvnrnd` do OCTAVE/MATLAB gera valores aleatórios segundo uma distribuição normal multivariada com vetor de média e matriz de covariância arbitrários.