UNIVERSIDADE FEDERAL DE SERGIPE

CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Profs.: Jânio Canuto e Jugurta Montalvão

Aluno: ______________________________ **Data:** ____________

Avaliação de Aprendizagem de Máquina

1) Sejam as probabilidades *a priori* de duas clases, $P(c_1)$ e $P(c_2)$, idênticas. Sabe-se que as duas classes possuem distribuições Gaussianas multivariadas, com médias e covariâncias dadas por:

$$\mu_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \mu_2 = \begin{bmatrix} 0.5 \\ 1.0 \end{bmatrix}, \Sigma_1 = \Sigma_2 = \begin{bmatrix} 0.2 & 0 \\ 0 & 0.2 \end{bmatrix}$$

a) Projete um classificador que minimize a probabilidae de erro.

b) Projete um classificador que minimize o risco médio quando a matriz de perda é

$$\Lambda = \begin{bmatrix} 0 & 0.5 \\ 1.0 & 0 \end{bmatrix}$$

2) Dadas as funções de distribuição de probabilidade para padrões de duas classes abaixo:

$$p(x|\omega_1) = \begin{cases} 1/9 & 1 \leq x_1 \leq 4 \; e \; 1 \leq x_2 \leq 4 \\ 0 & caso \; contrário \end{cases}$$

$$p(x|\omega_2) = \begin{cases} 1 & 2 \leq x_1 \leq 3 \; e \; 2 \leq x_2 \leq 3 \\ 0 & caso \; contrário \end{cases}$$

onde $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$. Assuma que $P(\omega_1) = P(\omega_2) = 0.5$.

a) Esboce o gráfico das regiões de decisão.

b) Proponha uma estratégia de classificação.

3) Um classificador linear, cuja fronteira é dada pela equação $f(x) = 0$, com $f(x) = \omega^T x + \omega_0$, foi usada para classificar $x_1 = \begin{bmatrix} 1.0 \\ -1.0 \\ 0.5 \end{bmatrix}$ e $x_2 = \begin{bmatrix} 0.5 \\ -1.5 \\ 0.5 \end{bmatrix}$, com $\omega = \begin{bmatrix} 0.5 \\ -0.5 \\ 0.5 \end{bmatrix}$ e $\omega_0 = -1/4$. O sinal de $f(x)$ determina a classe atribuída a x (*i.e.*, classe $+1$ ou classe -1).

a) Mostre que x_1 e x_2 estão do mesmo lado do plano da fronteira.

b) Qual a distância (euclidiana) mínima de x_1 (ou x_2) à fronteira de decisão?

4) Sejam as probabilidades *a priori* de duas classes, $P(C_1) = 0.6$ e $P(C_2) = 0.4$. Numa pequena região do espaço amostral, onde as densidades de probabilidade condicionais $p(x|C_1)$ e $p(x|C_2)$ são aproximadamente constantes, sabe-se que $p(x|C_1) = 3p(x|C_2)$. Dentro dessa pequena região, foram encontrados K padrões. Qual a probabilidade da maioria dos padrões encontrados serem realmente da classe mais provável dentro da região analisada se:

a) $K = 1$

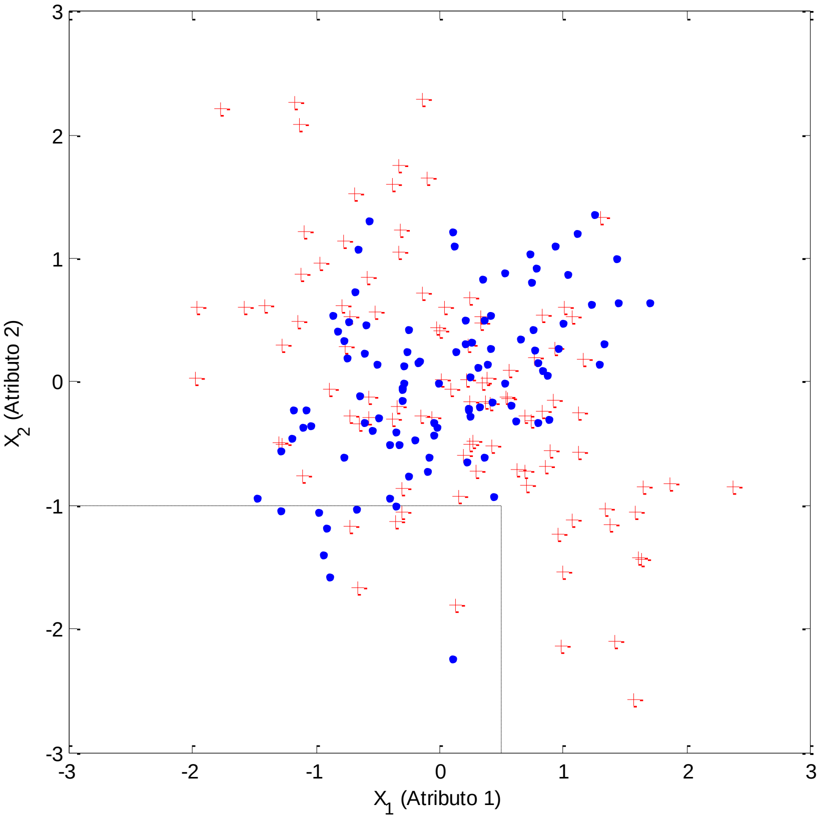
b) $K = 3$

c) $K = 5$

5) O gráfico abaixo representa 100 amostras coletadas da classe 1 (pontos) e 100 amostras da classe 2 (cruzes).

a) Use o método K-NN, com $K = 2$, e estime as densidades condicionais $p(x|classe\ 1)$ e $p(x|classe\ 2)$, no ponto $x = (0.5, -1)$, indicado na figura.

***Obs**: Use régua ou estime visualmente as distâncias, indicando claramente os valores usados na sua solução.*



b) Assumindo que as duas classes são Gaussianas, ambas com médias em (0,0), e covariâncias $\Sigma_1 = \begin{bmatrix} 0.5 & 0.25 \\ 0.25 & 0.5 \end{bmatrix}$ e $\Sigma_2 = \begin{bmatrix} 1.0 & -0.5 \\ -0.5 & 1.0 \end{bmatrix}$, recalcule as densidades condicionais $p(x|classe\ 1)$ e $p(x|classe\ 2)$, no ponto $x = (0.5, -1)$, indicado na figura.

c) Dado que as probabilidades *a priori* são iguais, *i.e.* $P(classe\ 1) = P(classe\ 2)$, como um classificador Bayesiano classificaria o ponto $x = (0.5, -1)$? Qual das densidades você prefere usar nessa decisão: aquela estimada no item (a), ou aquela obtida no item (b)? Explique.

6) Na retropropagação do erro, em uma rede neural artificial, o 'custo' associado à propagação do erro através de um neurônio com função de ativação do tipo tangente hiperbólica é $(1 - y^2)$, onde y representa a própria saída neurônio. Determine, também em função de y, qual seria esse 'custo' se a função de ativação fosse

a) $y(x) = \dfrac{1}{1 + e^{-\beta x}}, \beta > 0$

b) $y(x) = \dfrac{1}{\sqrt{x^2 + \beta}}. \beta > 0$

7) Uma MLP pode ser usada para *aprender* probabilidades $P(C_i|x)$, onde C_i, representa a i-ésima classe, e x é uma observação que se deseja classificar. Num problema onde x é um escalar (1-D) que pode ter sido gerado por uma das 3 classes: C_1, C_2 e C_3, deseja-se treinar uma rede neural com 1 entrada e 3 saídas, de forma que cada observação, x, seja associada a um vetor de saída desejada, $t = \begin{bmatrix}1\\0\\0\end{bmatrix}, \begin{bmatrix}0\\1\\0\end{bmatrix}$ ou $\begin{bmatrix}0\\0\\1\end{bmatrix}$, conforme a classe de origem da observação é C_1, C_2 ou C_3, respectivamente. Além disso, a função de ativação dos neurônios da camada de saída deve ser exponencial $y(x) = e^x$, e todo vetor de saída $y = \begin{bmatrix}y_1\\y_2\\y_3\end{bmatrix}$ deve ser 'corrigido' para um vetor de probabilidades $p = \begin{bmatrix}p_1\\p_2\\p_3\end{bmatrix}$, tal que $p_i = \frac{y_i}{y_1+y_2+y_3}$.

Assuma, por simplicidade, que há apenas um neurônio na camada escondida, com função de ativação do tipo tangente hiperbólica, e que o critério de ajuste da rede é o erro quadrático médio entre t e p. Quais as equações de ajuste, pelo gradiente negativo, para cada um dos pesos e polarizações dessa MLP?

***Observação importante**: A saída dessa rede é p, não y.*

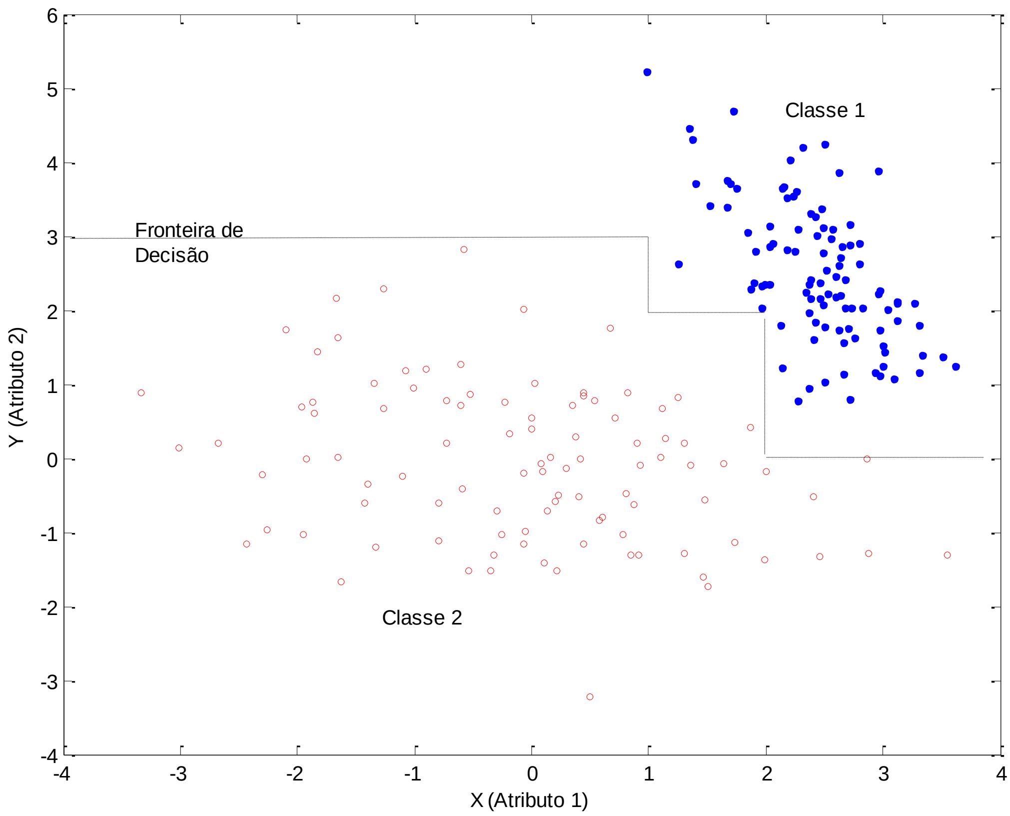
8) Dados foram coletados aleatoriamente e experimentalmente, resultando na seguinte coleção de medidas: $\{-2, -1, 2.5, -1.8, 3.5, 1.5, 2.1, 3.8\}$. Sabe-se que os dados $\{-2, -1, 2.5, -1.8\}$ vieram de uma dada classe de medidas (Classe 1), enquanto que os dados $\{3.5, 1.5, 2.1, 3.8\}$, da outra classe de medidas (Classe 2). Use esses dados como base de treinamento e, através da minimização do erro quadrático médio (pseudo-inversão), obtenha um classificador automático de dados. Explique claramente como funcionará esse classificador para os novos dados que forem coletados.

9) Uma empresa que empresta dinheiro acumuou alguma experiência com clientes antigos, e essa experiência foi sintetizada na tabela abaixo. Agora, a empresa gostaria de treinar um classificador automático de novos clientes com base na experiência acumulada. Assuma que você é o encarregado de criar esse classificador, e que você escolheu usar uma árvore de decisão.

a) Com base no critério da variação de entropia, dado um novo cliente, qual parece ser o melhor atributo para prever se esse cliente pagará o empréstimo no prazo? Qual a variação máxima de entropia decorrente da decisão baseada nesse atributo?
b) Explique claramente como você sugere que sejam tratados os dados não anotados na tabela (indicados com o símbolo "—").

Atributos				*Classificação*
Nível de Instrução	Salário aproximado	Possui casa própria?	Sexo	Pago o empréstimo no prazo?
Universitário	R$ 7.000,00	Sim	M	Sim
Segundo grau	—	Sim	M	Sim
Segundo grau	R$ 2.000,00	Não	F	Sim
Universitário	R$ 7.000,00	Não	M	Não
Universitário	R$ 7.000,00	Sim	M	Sim
Universitário	R$ 2.000,00	Sim	F	Sim
Segundo grau	R$ 7.000,00	Não	M	Não
Universitário	R$ 2.000,00	—	F	Sim
Segundo grau	R$ 2.000,00	Não	M	Não

10) Uma árvore de decisão poderia prover a fronteira de decisão detalhada no gráfico abaixo? Se sim, quantos ramos e folhas possuiria essa árvore? Faça um esboço.



11) Um projetista está ajustando uma GMM, com apenas 2 Gaussianas, através do método EM. O ajuste já foi iniciado e os parâmetros atuais (obtidos no último passo) foram:

$$\mu_1 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}, \Sigma_1 = \begin{bmatrix} 0.5 & 0.25 \\ 0.25 & 0.5 \end{bmatrix}, \mu_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 1.0 & -0.5 \\ -0.5 & 1.0 \end{bmatrix}, \alpha_1 = \alpha_2.$$

onde μ são as médias, Σ as matrizes de covariância e α os pesos da mistura.

Usando os dados fornecidos na tabela abaixo, execute apenas mais um passo do EM e apresente os novos parâmetros após o ajuste.

Atributos	
X_1	X_2
-0.5	-0.5
-1.0	-1.0
1.0	0.0

Lembrete: para uma matriz não-singular 2×2, $R = \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix}$, temos que $R^{-1} = \frac{1}{|R|}\begin{bmatrix} r_{22} & -r_{12} \\ -r_{21} & r_{11} \end{bmatrix}$, onde $|R| = r_{11}r_{22} - r_{12}r_{21}$.

12) Faça o agrupamento dos padrões da tabela abaixo em $K = 3$ *clusters*, usando o K-médio, explicando cada passo.

Atributos	
X_1	X_2
-0.76	-0.85
0.67	0.25
1.47	-1.54
1.14	-0.68
0.85	-0.72
0.45	0.81

13) Use o resultado da questão anterior para inicializar os parâmetros de uma GMM com 3 gaussianas e

a) Explique como foi feita essa inicialização
b) Execute, manualmente, um passo de Esperança (E) do EM.
c) Execute, manualmente, um passo de Maximização (M) do EM.
d) Comente suas soluções para os problemas numéricos e/ou de convergência caso encontre algum neste problema.

14) Mostre que a estimativa de máxima verossimilhança da média aritmética, $\mu = \frac{\sum_{i=1}^{N} x_i}{N}$, pode ser calculada recursivamente como

$$\mu_{i+1} = \mu_i + \frac{1}{i+1}(x_{i+1} - \mu_i)$$

Obs.: Esta fórmula pode ser usada para melhorar a eficiência computacional de algoritmos de ajuste sequencial, pois elimina a necessidade de recalcular o somatório a cada etapa.

15) Utilizando a linguagem de programação de sua preferência, forneça uma implementação de:

a) Um classificador
b) Um agrupador

16) Utilize os métodos desenvolvidos na questão anterior em dois conjuntos de dados de sua escolha. Descreva e comente seus resultados. (*Os métodos funcionaram corretamente? Foi necessário algum ajuste especial? Alguma iéia de como obter resultados melhores?*)

Obs.: Muitos conjuntos de dados podem ser encontrados em http://archive.ics.uci.edu/ml/datasets.html