DCA0133 - APRENDIZAGEM DE MÁQUINA E MINERAÇÃO DE DADOS - 2024.2

Primeira Lista de Exercícios e Trabalhos

Data de apresentação da lista e dos trabalhos: 17/10/2024

Obs. A lista e os trabalhos podem ser feitos de forma individual ou em grupo de até dois alunos. No dia da entrega da lista os alunos devem apresentar os problemas da lista e os trabalhos.

Lista de Problemas

1. Uma psicóloga faz uma pequena enquete sobre "felicidade" com base no seguinte vetor de atributos **x**=(rico,casado, sem problema de saúde). Na enquete ela pede para marcar 1 ou 0, correspondendo as respostas sim ou não para cada atributo e se a pessoa se considera feliz ou não. A tabela abaixo mostra o resultado da enquete. Usando o método Naive-Bayes como seria classificado em termos de felicidade uma pessoa não rica, casada e saudável.

Exemplo:Naive Bayes

Pessoa	Rico	Casado	Saudável	Feliz	
1	1	1	1	1	
2	0	0	1	1	
3	1	1	0	1	
4	1	0	1	1	
5	0	0	0	0	
6	1	0	0	0	
7	0	0	1	0	
8	0	1	0	0	
9	0	0	0	0	
10	0	1	1	?	

7

2-) Um determinado banco deve decidir se um cliente deve ou não receber um empréstimo bancário em função da sua condição de bom ou mau pagador. Considerando os dados de treinamento abaixo, aplique o classificador, no caso uma árvore de decisão, para atribuir a classe (rótulo) para o registro 12 :

Registro	Tem casa própria	Estado Civil	Rendimentos	Bom Pagador
1	Sim	Solteiro	Alto	Não
2	Não	Casado	Médio	Não
3	Não	Solteiro	Baixo	Não
4	Sim	Casado	Alto	Não
5	Não	Divorciado	Médio	Sim
6	Não	Casado	Baixo	Não
7	Sim	Divorciado	Alto	Sim
8	Não	Solteiro	Médio	Sim
9	Não	Casado	Baixo	Não
10	Não	Solteiro	Médio	Sim
11	Sim	Divorciado	Médio	Não
12	Não	Divorciado	Alto	?

3-) Considere o problema de separação de padrões constituído por duas classes $\omega 1$ e $\omega 2$. Assumindo que as distribuições associadas a cada classe são gaussianas com probabilidades a priori dadas por ($P(\omega 1) = P(\omega 2) = 1/2$) . As distribuições gaussianas para cada classe apresentam os seguintes parâmetros (vetor média e matriz de covariância) dados por:

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}; \quad \boldsymbol{\Sigma}_1 = \begin{pmatrix} 1/2 & 0 \\ 0 & 2 \end{pmatrix} \text{ and } \boldsymbol{\mu}_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix}; \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$

As inversas das matrizes de covariância são dados por:

$$\boldsymbol{\Sigma}_1^{-1} = \begin{pmatrix} 2 & 0 \\ 0 & 1/2 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma}_2^{-1} = \begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix}.$$

As funções discriminantes $g_1(\mathbf{x})$ e $g_2(\mathbf{x})$ definem a superfície de separação ou decisão entre os padrões ou classes associadas as distribuições gaussianas. A superfície de separação é obtida fazendo $g_1(\mathbf{x})=g_2(\mathbf{x})$.

Para as condições deste problema as funções descriminantes $gi(\mathbf{x})$, i=1,2 podem ser calculadas pela equação abaixo.

$gi(\mathbf{x}) = -1/2((\mathbf{x} - \boldsymbol{\mu}i)^t \boldsymbol{\Sigma}^{-1} i\ (\mathbf{x} - \boldsymbol{\mu}i) - 1/2 \ln |\boldsymbol{\Sigma}i| \qquad i=1,2$

onde

$\boldsymbol{\mu}$: Vetor Média

$\boldsymbol{\Sigma}$: Matriz de Covariância

$\boldsymbol{\Sigma}^{-1}$: Inversa da Matriz de Covariância

$|\boldsymbol{\Sigma}i|$: **Determinante da matriz de covariância**

a-) Mostre que a superfície de decisão definida por g1(**x**)-g2(**x**)**=0**

$$x_2 = 3.514 - 1.125x_1 + 0.1875x_1^2.$$

b-) Trace o gráfico da superfície de decisão
c-) Indique a que classe pertence os padrões $\mathbf{x_1}=[4,5]^t$ e $\mathbf{x_2}=[-3,4]^t$

4-) Considere o problema de classificação de padrões bidimensionais constituído neste caso de 2 padrões ou duas classes. A distribuição dos padrões tem como base um quadrado centrado na origem interceptando os eixos nos pontos +1 e -1 de cada eixo. Os pontos +1 e -1 de cada eixo são centros de quatro semicírculos que se interceptam no interior do quadrado originando uma figura na forma das asas de uma borboleta. A classe 1 corresponde aos dados em cada asa e a classe 2 os que estão fora das asas. Após gerar aleatoriamente dados que venham formar estas distribuições de dados, selecione um conjunto de treinamento e um conjunto de validação. Treine uma Random Forest para classificar os padrões associados a cada uma das classes. Verifique o desempenho do classificador usando o conjunto de validação e calculando a matriz de confusão.

5-) Uma rede de crença (ou rede bayesiana), modela a relação entre as variáveis: oil (price of oil), inf (inflation) , eh (economy health), bp (British Petroleum Stock price), rt (retailer stock price). Cada variável tem dois estados (l:low) e (h:high), exceto a variável bp que tem adicionalmente o estado (n: normal). A rede de crença modela as variáveis de acordo com a tabela abaixo.

P(eh=l)=0.2	
P (bp=l|oil=l)=0.9	P(bp=n|oil=l)=0.1
P(bp=l|oil=h)=0.1	P(bp=n|oil=h)=0.4
P(oil=l|eh=l)=0.9	P(oil=l|eh=h)=0.05
P(rt=l|inf=l,eh=l)=0.9	P(rt=l|inf=l,eh=h)=0.1
P(rt=l|inf=h,eh=l)=0.1	P(rt=l|inf=h,eh=h)=0.01
P(inf=l|oil=l, eh=l)=0.9	P(inf=l|oil=l, eh=h)=0.1
P(inf=l|oil=h, eh=l)=0.1	P(inf=l|oil=h, eh=h)=0.01

a-) Apresente a rede de crença para este problema
b-) Dado que a bp=n e rt=h, qual é a probabilidade de que a inflação seja alta?

Trabalhos:

Escolha dois dos três trabalhos abaixo:

Trabalho1: Apresente um trabalho sobre aplicação do algoritmo Naïve-Bayes para:

(i) a detecção de Spam em mensagens de email
ou
(ii) para classificar páginas de texto com base em um tema de interesse (esporte, política, ...etc.) presente nas palavras que aparecem nas páginas.
ou
(iii) outra aplicação a ser escolhida

Trabalho 2: Pesquise e apresente um trabalho sobre aplicações da técnica Random Forest para:

(i) Reconhecimento biométrico de faces
ou
(ii) Detectar uma determinada doença com base em exames médicos
ou
(iii) Outra aplicação de livre escolha
.

Trabalho 3: Apresente um trabalho sobre aplicações de redes bayesianas para:

(i) Diagnóstico médico

ou

(ii) Diagnóstico de falha em um carro

ou

(iii) Outra aplicação de livre escolha