Aprendizagem de Máquinas e Mineração de Dados – 2018.2 - DCA 0133

Primeira Lista de Exercícios
Data de apresentação da lista: 11/09/2018

1. Um fabricante de computador usa chips de três fornecedores diferentes, aqui denominados A, B e C, sendo 25% do fornecedor A, 35% do fornecedor B e 40% do fornecedor C. Os chips dos fornecedores A, B e C apresentam taxas de defeitos dadas pelas seguintes probabilidades respectivamente 0,005, 0,008 e 0,001. Se um chip é selecionado aleatoriamente e visto ser defeituoso, determine qual foi o mais provável fabricante do chip.

2. Considerando os dados de treinamento abaixo, aplique o classificador Naive-Bayes, para atribuir a classe (rotulo) para o registro 11:

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	Não
8	Não	Solteiro	Médio	Sim
9	Não	Casado	Baixo	Não
10	Não	Solteiro	Médio	Sim
11	Não	Divorciado	Médio	?

3. Considere o problema de decisão caracterizado por uma sequência de eventos que podem ser apresentados por um gráfico conhecido como árvore de decisão. O problema em questão consiste das escolhas e das decisões por parte de uma empresa de petróleo. Uma determinada empresa petrolífera obteve a concessão para explorar uma certa região. Os estudos anteriores (testes preliminares) estimam a probabilidade de existir petróleo nessa região em 0,25. A companhia pode optar por um novo teste, que custa US$ 500.000,00, sendo que, se realmente existe petróleo, esse teste dirá com uma probabilidade de 0,85 que existe, e se realmente não existe, dirá com probabilidade 0,75 que não existe. Considerando que o custo de perfuração será de US$ 3.000.000,00 e que, se for encontrado petróleo, a companhia receberá US$ 150.000.000,00. Considere, portanto os seguintes eventos e os seus complementos: (i) Evento T (a companhia faz o teste); (ii) Evento F (o teste é favorável à existência de petróleo; (iii) Evento P (a companhia perfura o poço); (iv) Evento E (existe petróleo).

   a-) Construa a árvore indicando os nós de decisões e os nós ao acaso (probabilidades). Simbolize os dois de forma diferente.
   b-) Determine o lucro =receita - despesas em cada percurso da árvore.

c-)Usando o critério do melhor valor esperado, determine o valor esperado em cada nó de decisão.
d-) Qual o valor esperado do lucro da companhia se forem tomadas as melhores decisões?

4. Seja X uma variável aleatória com distribuição normal com média μ e variância σ^2. Determine os estimadores de máxima verossimelhança dos parâmetros μ e σ^2. Generalize o problema considerando **X** um vetor variável aleatória com distribuição normal multivariada dada por $f(\mathbf{x} \mid C_i) = \frac{1}{(2\pi)^{p/2} \det(\mathbf{C})^{1/2}} \exp[-\frac{1}{2}(\mathbf{x}-\mu_i)^t \mathbf{C}^{-1}(\mathbf{x}-\mu_i)]$ com vetor média μ e C matriz de covariância . Determine os estimadores de máxima verossimelhança dos parâmetros vetor média e μ e da matriz de covariância C.

5. Considere o problema de estimativa da matriz de covariância de uma dada distribuição de vetores aleatórios dados por $\mathbf{x_1} = \begin{bmatrix} 8 \\ 2 \\ 7 \end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix} 5 \\ 6 \\ 10 \end{bmatrix}, \mathbf{x}_3 = \begin{bmatrix} 3 \\ 2 \\ 4 \end{bmatrix}, \mathbf{x}_4 = \begin{bmatrix} 3 \\ 6 \\ 5 \end{bmatrix}$. Estime o vetor média e a matriz de covariância amostral $\mathbf{S} = \{s_{ij}\}$ assumindo que a matriz é simétrica. A média amostral é dada por $\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i$ e a variância amostral é dada por $s_{ij} = \frac{\sum_{l=1}^{N}(x_{il} - \hat{\mu}_i)(x_{jl} - \hat{\mu}_j)}{N-1}$ ou . Determine a distribuição de probabilidade assumida ser Gaussiana.

6. Considere o problema de separação de padrões com distribuições de probabilidades multivariadas. Considere duas distribuições de probabilidades condicionais gaussianas cada uma associadas a uma classe de padrões. As distribuições são dadas por $f(\mathbf{x} \mid C_i) = \frac{1}{(2\pi)^{p/2} \det(\mathbf{C})^{1/2}} \exp[-\frac{1}{2}(\mathbf{x}-\mu_i)^t \mathbf{C}^{-1}(\mathbf{x}-\mu_i)]$ para i=1,2, sendo **C** a matriz de covariância, assumida igual para as duas distribuições, com inversa admissível e μ_i os vetores médias para cada classe. Para os vetores médias são dados por $\mu_1 = [-1,5,0]^t, \mu_2 = [+1,5,0]^t$ e a matriz de covariância é dada por $\mathbf{C} = 0,25\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. Esboce a distribuição dos dados no plano x1, x2. Pesquise e mostre que o classificador Gaussiano de máxima verossimilhança gera uma superfície linear (hiperplano) para separar os padrões associados a cada classe Ci para as condições deste problema. Sob estas condições projeto um classificador

(discriminador) gaussiano. Explique como treinar o classificador e como avaliar o seu desempenho depois de treinado.

7. Considere um sistema que monitora a operação de uma máquina mecânica. A máquina deve produzir em série uma determinada peça mecânica com espessura média de 0,06 pol. A fim de verificar se a máquina no decorrer do tempo de operação ainda está em boas condições, o sistema realiza um processo de amostragem selecionado 20 peças, cujas as espessuras médias são de 0,07 pol, com desvio padrão de 0,008 pol. Considerando que o sistema usa o teste t –Student verifique a hipótese (decisão) do sistema de que a máquina ainda está em boas condições, utilizando um nível de significância (a) de 0,05 (b) de 0,01.

8. Um determinado processo de cauterização química é utilizado para remover cobre de placas de circuito impresso. Sejam X1 e X2 variáveis aleatórias representando as amostragens do processo quando se usam duas concentrações diferentes. Considerando que 8 o número das amostras para cada uma das concentrações. As variâncias amostragens para X1 é 3,89 e 4,02 para para X2. Considerando o nível de significância igual a 0,05, teste a hipótese de que as duas concentrações possuem a mesma variabilidade contra a hipótese alternativa.

Trabalho: Experimentos Computacionais

1 – Utilizando os métodos de geração de números aleatórios, desenvolva um software que possibilite gerar as seguintes distribuições de probabilidades.

a-) Exponencial
b-)Normal (Gaussiana)
c-) Qui-quadrado (pesquise o processo de geração)
d-) Distribuição de Poisson

Apresente os resultados deste software sob forma de gráficos, tabelas e compare com os resultados analíticos.

2-) Utilizando método de Monte Carlo calcule as seguintes integrais definidas. Compare os resultados com valores tabelados e ou soluções analíticas se existirem.

a-) $I = \frac{1}{2\pi}\int_0^2 e^{-x^2/2}dx$

b-) $I = \int_0^1 \int_0^1 e^{-(x^2+y^2)}dxdy$

3-) Implemente o método Naives-Bayes e aplique o mesmo para o problema de detecção de spam.