**DCA0133 - APRENDIZAGEM DE MÁQUINA E MINERAÇÃO DE DADOS - 2024.1**

**Primeira Lista de Exercícios e Trabalhos**

**Data de apresentação da lista e dos trabalhos: 04/04/2024**

Obs. A lista e os trabalhos podem ser feitos de forma individual ou em grupo de 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 5 padrões. 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 classe e as outras quatro classes nas regiões de não intersecção.

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

Instruções:

1. Escolha 2 dos 3 trabalhos apresentados abaixo;
2. Os trabalhos devem ser, preferencialmente, organizados em um *notebook* no Google Colab e o link do mesmo enviado junto das respostas da lista;

    a. Para compartilhar um *notebook* no Colab vá até “Partilhar” > ”Acesso Geral” > ”Qualquer pessoa com o link” > ”Copiar link”;
    b. Lembre de, no *notebook*, identificar nome do grupo, turma e trabalhos escolhidos;
    c. Exemplos de *notebooks* estão no repositório do curso.
3. Quando necessário, busque apresentar a base de dados escolhida por meio de visualizações de sua preferência (ex. *dataframes*, *scatter plot*, *pair plot*, *box plot*, etc);
    a. O objetivo é visualizar e compreender um pouco as amostras, atributos e correlações antes de aplicar algum algoritmo de aprendizado.
4. O uso de ferramentas tipo GenAI (ex. ChatGPT) é permitido e incentivado. No entanto, é necessário entender os conceitos aplicados para responder possíveis questionamentos do professor e de outros alunos durante a apresentação, já que isto fará parte da avaliação.

- **Trabalho 1:**

Pesquise e apresente um trabalho sobre o algoritmo Naïve-Bayes para a detecção (classificação) de Spam em mensagens de email.

Dicas:

- Base de dados de spam em emails
- Usando a biblioteca sklearn em python, alguns métodos importantes: train_test_split, CountVectorizer e MultinomialNB.

- **Trabalho 2:**

Pesquise e apresente um trabalho sobre Random Forest para inferir o preço de uma casa (regressão) baseado em atributos como área do terreno, número de quartos, número de banheiros, etc.

Dicas:

- Base de dados de preços de casas
- No Kaggle você pode ver o código de outras pessoas (na aba “Code”) resolvendo o mesmo problema com diferentes estratégias.
- Usando a biblioteca sklearn em python, alguns métodos importantes: train_test_split, MinMaxScaler, RandomForestRegressor.

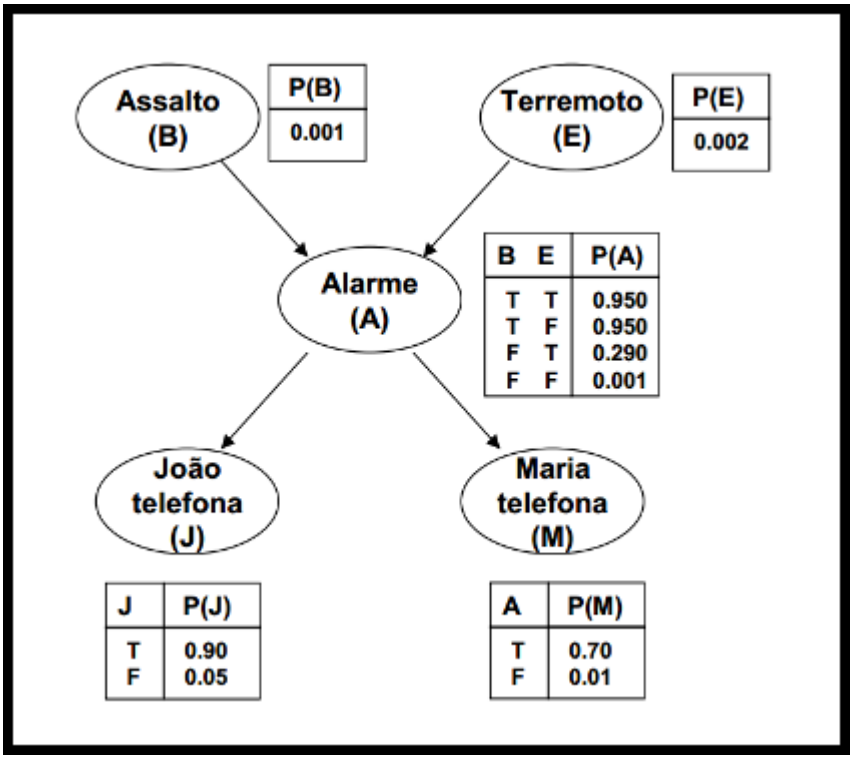
- **Trabalho 3:**

Construa e apresente uma rede bayesiana para um problema simples de livre escolha e use-a para resolver a probabilidade independente de um evento final.

Isto inclui desenhar a rede (da maneira que preferir) e criar as tabelas de probabilidades. O grafo acíclico direcionado deve ter pelo menos quatro nós e dois níveis de dependência:

Na figura 1 há um exemplo de rede bayesiana para um sistema de alarme, similar ao visto em sala. Aqui poderíamos realizar o cálculo da probabilidade de João telefonar P(J) ou a

robabilidade de Maria telefonar P(M).




Sugestões de problemas (incompletos):

- Probabilidade de se obter uma carta de recomendação condicionado ao evento de se obter uma boa nota;
- Probabilidade de chegar atrasado em um compromisso condicionado ao evento de dormir demais condicionado ao evento do alarme funcionar;
- Probabilidade de escorregar condicionado ao evento do chão estar molhado e ao evento de estar usando um calçado sem aderência.