Lista de Exercícios de Mineração de Dados

Lista Individual

Data de Entrega: 13/02/2025

Não utilize funções prontas de algoritmos aprendidos em sala de aula (a não ser quando informado o contrário). Implemente as suas e apresente-as na lista. Faça um relatório explicando como foi resolvido o exercício e envie junto com o código fonte. Envie a lista para o classroom da disciplina.

1) A base de dados Nebulosa (disponibilizada em anexo) está contaminada com ruídos, redundâncias, dados incompletos (substituídos pelo valor -100), inconsistências e *outliers*. Para esta base:

a) Obtenha os resultados da classificação (métrica acurácia) usando a técnica do vizinho mais próximo (NN) e Rocchio. Utilize a distância Euclidiana e a base de dados crua, sem pré-processamento. Use o conjunto de 143 amostras para treino e o de 28 amostras para teste. Remova as amostras com dados incompletos.

b) Realize um pré-processamento sobre os dados de forma a reduzir os ruídos, as redundâncias, inconsistências, *outliers* e a interferência dos dados incompletos. Obtenha os resultados da classificação usando a técnica do vizinho mais próximo (NN) e Rocchio usando a distância Euclidiana e a mesma divisão dos dados.

c) Compare os resultados obtidos em a) e b). Qual deles retornou o melhor resultado? Por quê?

2) Dada a base de dados de análise de sentimentos Tweets_Mg.csv do Kaggle (https://www.kaggle.com/datasets/leandrodoze/tweets-from-mgbr?resource=download), aplique a biblioteca Spacy do Python para obter a representação embedding dos textos que estão no campo “Text” (pode carregar os pesos de pt_core_news_sm). Considere a variável alvo o campo “Classificacao”. Em seguida, faça:

a) Obtenha a acurácia de classificação quando usando o classificador vizinho mais próximo (NN) (utilize a distância Euclidiana). Use as primeiras 8000 amostras para treinar o modelo e as demais para teste. Compare o resultado obtido pela sua implementação do vizinho mais próximo com a do scikit-learn. Foram semelhantes?

b) Aplique o PCA sobre os dados de treino e selecione o número de componentes até eles corresponderem a 90% da informação de variância dos dados. Use a matriz de coeficiente de correlação para isto. Quantos componentes foram selecionados? Calcule a nova acurácia do NN usando as componentes selecionadas (use a versão do NN que você programou). O resultado alterou de forma significativa em relação ao obtido em a)? Qual foi a vantagem observada usando PCA?

3) Para a base de dados Auto MPG (disponibilizada em https://archive.ics.uci.edu/ml/datasets/Auto+MPG) faça:

a) Baixe o arquivo auto-mpg.data, remova as linhas que tem interrogação (?) e remova a última coluna (por quê?). Com as 150 primeiras linhas obtenha um modelo de regressão linear multivariada para predizer o valor da primeira variável (mpg). Avalie o resultado sobre o restante da base de dados, usando a métrica RMSE.
b) Aplique o método do SHAP values (faça os cálculos e explique o desenvolvimento) para obter a contribuição de cada variável da primeira amostra na saída do modelo estimado.
c) Verifique quais são os atributos que estão relacionados com a saída: A partir dos coeficientes obtidos, aplique o teste F de Snedecor sobre cada variável individualmente (conforme nos slides). Indique quais foram os atributos que podem ser desconsiderados. Obtenha sobre o restante da base de dados a métrica RMSE com o modelo sem considerar esses atributos (não precisa estimar um novo modelo, só considere os valores dos coeficientes deles iguais a zero). Compare os resultados obtidos em a) e em b). Considere que os resíduos do modelo possui distribuição aproximadamente normal e que $F_{1,142} = 3,908$.

4) Para a base de dados MAGIC Gamma Telescope (disponível em https://archive.ics.uci.edu/ml/datasets/MAGIC+Gamma+Telescope), ordene ela de forma aleatória, e separe 4020 amostras para serem as amostras de teste. As demais 15 mil amostras, divida-as em três partes de forma estratificada. Em seguida, faça:
a) Realize o 3-fold cross validation usando a implementação do kNN do scikit-learn. Para cada execução do 3–fold, determine o melhor valor de k ao avaliar diferentes valores com o kNN treinado com duas parte e validado por outra. Obtenha o resultado de F1 sobre o terceiro conjunto (considere a classe positiva a classe minoritária). Após as execuções, aplique o melhor modelo encontrado durante o 3–fold sobre as amostras de teste. Retorne as métricas de acurácia, recall, precisão, F1 e a matriz de confusão.
b) Implemente a técnica de seleção de características SFS. Use o valor do k do melhor modelo encontrado em a). Selecione 6 atributos usando duas partes como treinamento e valide os atributos sobre a terceira parte usando a métrica F1. Após determinar os 6 atributos, obtenha as métricas de acurácia, recall, precisão, F1 e a matriz de confusão sobre os dados de teste. Retorne os 6 atributos selecionados.
c) Implemente a técnica de seleção de características SBS e repita o mesmo procedimento da letra b).
d) Compare os resultados obtidos nas letras anteriores os tempos de execução do SFS e SBS.

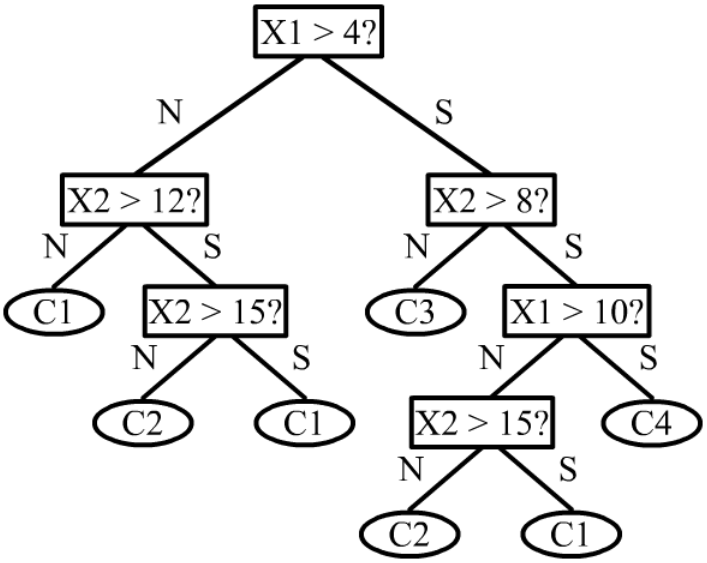
5) Aplique o Naive Bayes sobre a base de dados Monk's Problems (disponível em http://archive.ics.uci.edu/ml/). Obtenha a acurácia, treinando com monks-2.train e testando em monks-2.test. Realize os experimentos:
a) Considerando uma distribuição Gaussiana dos atributos;
b) Discretizando os valores em intervalos de tamanho 1;
c) Discretize os valore da mesma forma que em b) usando a suavização de Laplace.

6) A figura abaixo ilustra uma árvore de decisão com dois atributos (X1 e X2) e 4 classes (C1, C2, C3 e C4).Obtenha:

a) As regras de decisão da árvore de decisão (remova as redundâncias, caso existam);

b) Faça um gráfico para ilustrar como o espaço 2D está sendo dividido entre as 4 classes pela árvore (semelhante ao apresentado em sala de aula);

c) Considere que em uma amostra foi observado somente o valor de X1 = 3,5 e o valor de X2 está faltando. Qual será a classe atribuída para esta amostra se for utilizado o método probabilístico mostrado em sala de aula? Por quê?



7) Para a base Appliances Energy Prediction (disponível em http://archive.ics.uci.edu/ml/):

a) Construa uma árvore de regressão usando o tree.DecisionTreeRegressor com max_depth = 2. Selecione aleatoriamente 75% dos dados para treinamento que serão usados para construir a árvore. Retorne a estrutura da árvore construída. Pode usar o comando tree.plot_tree, tal que mostre o atributo avaliado, o erro quadrado, número de amostras e o valor de cada nó.

b) Use os restantes 25% dos dados para avaliação. Retorne as medidas MAPE e RMSE.

c) Retorne as regras obtidas pela árvore.

8) Para as bases de dados Spiral e Jain (disponíveis em http://cs.joensuu.fi/sipu/datasets/), agrupe os dados em 3 e 2 grupos, respectivamente, usando kmeans e clusterização hierárquica. Avalie o resultado com a métrica de acurácia com o seguinte procedimento: para cada cluster verifique qual foi a classe predominante, amostras pertencentes a outras classes estão no grupo errado. Faça os experimentos com a distância Euclidiana. Gere gráficos com os grupos formados pelo kmeans e clusterização hierárquica. Comente os resultados. Lembre-se de não usar o atributo da classe para agrupar os dados.

9) Use o Apriori algoritmo para encontrar as regras de associação do conjunto de transações no arquivo Market_Basket_Optimisation.txt. Faça o processamento que achar necessário e comente a respeito. Estabeleça os valores de confiança mínima e suporte de forma a encontrar algumas regras de associação. Apresente as regras encontradas.

Questões Teóricas

1) Comente sobre a veracidade das afirmações:

a) "Quanto mais variáveis de entrada forem usadas em um modelo de aprendizado de máquina, melhor será a qualidade do modelo".

b) "Independente da qualidade, quanto mais amostras forem obtidas para uma base de dados, maior a tendência de se obter modelos mais adequados".

c) "Às vezes com simples manipulações na base de dados (limpeza, conversão de valores, etc.) pode-se conseguir melhoras significativas nos resultados, sem fazer nenhuma alteração na técnica de aprendizado de máquina usada".

2) Dada a seguinte tabela de investimentos, os seus retornos para cada situação da economia, e a probabilidade a priori de cada situação econômica, qual é o investimento mais apropriado segundo o critério de Decisão de Bayes, ou seja, aquele que de forma geral retorna mais lucro? Retorne os cálculos.

	Economia Crescente	**Economia Estável**	**Economia Decrescente**
Investimento Conservador	$30.000,00	$5.000,00	$-10.000,00
Investimento Especulativo	$40.000,00	$10.000,00	$-30.000,00
Investimento Cíclico	$-10.000,00	$0,00	$15.000,00
Probabilidade a Priori	0.1	0.5	0.4

3) Explique a relação entre força e correlação entre as árvores do algoritmo Random Forest.