Mineração de Dados e Descoberta de Conhecimento

Profs. Heitor S. Lopes e Thiago H. Silva (UTFPR, 2023) - Exercício #3

A) Classificação - Árvores de decisão

1. **Objetivo**: explorar a classificação de dados com árvores de decisão usando o software Orange.
2. **Questões**:
   A. O dataset Soybean se refere ao diagnóstico de 19 doenças comuns da soja. Ele tem 35 atributos e 683 instâncias. Faça o pré-processamento necessário para upload no Orange.
   B. Utilizando as ferramentas de visualização de dados, o que é possível preliminarmente inferir preliminarmente sobre os atributos deste dataset?
   C. Selecione a coluna “class” como o alvo da classificação, sendo as demais colunas os atributos previsores. Use validação cruzada estratificada de 5-folds para o treinamento de uma Árvore de Decisão com os parâmetros default. Anote o tamanho da árvore obtida (número total de nós, profundidade e número de nós-folhas) e as medidas de qualidade (acurácia, precision, recall e F1 score). Justifique qual a medida de qualidade adequada para este caso.
   D. Mostre a matriz de confusão gerada pelo treinamento/teste da árvore de decisão. Identifique nesta árvore quais foram as classes que tiveram 100% e 0% de acerto, respectivamente. Justifique este comportamento (em especial para as classes com 0% de acerto).

B) Classificação

3. **Objetivo**: explorar a classificação de dados com árvores de decisão árvores de decisão utilizando validação cruzada, Knn e redes neurais
4. **Questões**:

   2- Considere ainda uma árvore de decisão para classificar se um indivíduo sobreviveu ou não com base no dataset “titanic.csv”. Qual o resultado médio de acurácia e F1-score utilizando a estratégia de validação cruzada 5-fold? Discuta os resultados.

3- Ainda considerando o dataset “titanic.csv”, construa um modelo utilizando k-NN para prever se uma pessoa sobreviveu ou não. Considere diferentes valores de k vizinhos: k={2,3,4,5,6,7}. Use validação cruzada 5-fold na avaliação. Houve variação significativa nos diferentes modelos testados? Algum deles foi melhor do que a estratégia baseada em árvore de decisão?

4- Avalie o resultado de uma rede neural com, pelo menos, duas camadas escondidas e dez neurônios em cada. Use validação cruzada 5-fold na avaliação. Existe tendência de melhora em relação às tentativas anteriores?

C) Regressão

5. **Dados:** use o dataset 'datasetCarros.csv' em todos os exercícios.
6. **Objetivo**: explorar os conceitos de regressão.
7. **Questões**:

1

a. Faça um modelo de regressão linear simples utilizando a variável 'KmRodado' para prever a 'PrecoVenda'.

b. Calcule o R2 para o modelo criado.

2

c. Separe o dataset em teste (5%) e treino (95%). Use o método 'train_test_split' do sklearn; configure o parâmetro random_state=10.

d. Treine um modelo de regressão linear múltipla no dataset de treino utilizando todas as variáveis (exceto 'Nome') para prever a 'PrecoVenda' e exiba os coeficientes do modelo.

e. Avalie o modelo encontrado utilizando o dataset de teste. Calcule o R2 e MSE.