Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Ciência da Computação

DCC642 Introdução à IA
2º Semestre de 2023
Prof. Luiz Chaimowicz

Trabalho Prático 2 - Aprendizado de Máquina

Data de entrega: 04/12

O objetivo desse trabalho é estudar técnicas de aprendizado supervisionado e não-supervisionado, mais especificamente os algoritmos *k-Nearest Neighbors (kNN)* e *k-Means*.

O trabalho é dividido em duas partes. Nas duas iremos utilizar o conjunto de dados *NBA Rookie Stats*, que contém estatísticas de jogadores novatos da NBA. Ele consiste de 1.340 observações, com as seguintes variáveis (atributos):

- `GP`: Total de jogos disputados
- `MIN`: Minutos jogados
- `PTS`: Pontos marcados por jogo
- `FGM`: Arremessos convertidos
- `FGA`: Tentativas de arremessos
- `FG%`: Percentual de conversão de arremessos
- `3PMade`: Arremessos de 3 pontos convertidos
- `3PA`: Tentativas de arremessos de 3 pontos
- `3P%`: Percentual de conversão de arremessos de 3 pontos
- `FTM`: Lances livres convertidos
- `FTA`: Tentativas de lances livres
- `FT%`: Percentual de conversão de lances livres
- `OREB`: Rebotes ofensivos
- `DREB`: Rebotes defensivos
- `REB`: Total de rebotes
- `AST`: Assistências
- `STL`: Roubos de bola
- `BLK`: Bloqueios
- `TOV`: Perdas de bola
- `TARGET_5Yrs`: Se a carreira do jogador durou pelo menos 5 anos na liga, tem valor 1. Caso contrário, o valor é 0.

A variável `TARGET_5Yrs` é a variável resposta de interesse no uso deste conjunto de dados.

Parte 1 - Aprendizado Supervisionado

Na primeira parte do trabalho você deverá implementar o algoritmo *k-Nearest Neighbors* para classificar se a carreira de um jogador vai durar pelo menos 5 anos na liga ou não, utilizando os atributos. Você deve separar o conjunto de dados em 2 partes: 80% dos dados (1.072 observações) serão usados no conjunto de treino e 20% (268 observações) para o teste (utilize os arquivos `nba_treino.csv` e `nba_teste.csv`). Teste o seu algoritmo com 4 valores de k diferentes (2, 10, 50 e outro valor a sua escolha) e, para cada um deles, classifique os dados do conjunto de **teste**, mostre a matriz de confusão obtida e compute os valores de **acurácia, precisão, revocação (*recall*) e F1**. Analise e discuta os resultados obtidos.

Parte 2 - Aprendizado Não-Supervisionado

Na segunda parte você deverá implementar o algoritmo *k-Means* para agrupar os jogadores em diferentes conjuntos (*clusters*) usando apenas os atributos das amostras (sem utilizar a coluna de resposta `TARGET_5Yrs`). Teste o seu algoritmo com 2 valores de k diferentes (2 e 3), imprimindo o valor obtido para os centróides de cada agrupamento. Além disso, verifique se existe alguma relação entre os agrupamentos obtidos nos dois testes e os rótulos originais dos dados. Por exemplo, com k igual a 2 o algoritmo conseguiu agrupar bem os jogadores nas duas classes de resposta originais? O que acontece para k igual a 3? Analise e discuta os resultados obtidos.

Pontos Extras:

Existem várias bibliotecas prontas que implementam o *kNN*, *k-Means* e vários outros algoritmos de aprendizado de máquina. Um exemplo é a biblioteca *scikit-learn* do Python. Portanto, como atividade extra, você poderá implementar o trabalho usando essas bibliotecas e comparar os resultados com os obtidos em sua implementação. Nesse caso, além da implementação, você deverá discutir a comparação em sua documentação.

O que deve ser entregue?

Faça um .zip ou similar contendo:

- Códigos fonte dos algoritmos desenvolvidos;
- Um arquivo readme.txt com informações sobre como compilar e executar seus programas;
- Documentação contendo uma **descrição sucinta** dos algoritmos, a **apresentação e discussão dos resultados obtidos** e as referências utilizadas.

Observações:

- O trabalho é individual.
- Você deverá implementar os algoritmos. Não é permitido o uso de bibliotecas prontas (exceto para a parte de pontos extras).
- Se você se basear em algum código pronto na sua implementação, indique a fonte na documentação.

Bom trabalho!