UNIVERSIDADE ESTADUAL DE MONTES CLAROS - UNIMONTES
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM
COMPUTACIONAL E SISTEMAS - PPGMCS
Machine Learning
Prof: Honovan Rocha

Trabalho Prático 2 - Machine Learning

O objetivo deste trabalho é implementar e avaliar o algoritmo Naive Bayes aplicado a diferentes problemas de classificação. Adicionalmente o algoritmo K-Médias deve ser utilizado para clustering.

1. (Classificação - Base de Dados Iris) :

1.1. Conforme visto nas primeiras aulas, busque e carregue o **Iris Dataset**:

- 4 Features: Sepal Length, Sepal Width, Petal Length, Petal Width
- 3 Classes: Versicolor, Setosa e Virginica

1.2. Com objetivo de visualizar os dados, plote os gráficos a seguir:

- As 6 combinações em 2D de dois dos quatro atributos do Iris dataset.
- As 4 combinações em 3D de três dos quatro atributos do Iris dataset.

1.3. **Implemente** o algoritmo Gaussian Naive Bayes e aplique à base de dados.

- Plote a superfície de decisão obtida pelo algoritmo.

1.4. **Implemente** o algoritmo Regressão Logística (2 versões, com e sem feature engineering) e aplique à base de dados.

- Encontre o melhor valor para o parâmetro α e para o grau do polinômio na versão com feature engineering.
- Plote a superfície de decisão obtida pelo algoritmo.

1.5. Compare a eficácia dos classificadores com relação à acurácia, utilizando a divisão dos dados (Treinamento, Validação e Teste). Execute cada experimento 10 vezes e compare as médias dos resultados.

1.6. Retire os rótulos da base de dados e **Implemente** o algoritmo K-Médias para encontrar agrupamentos nos dados. Utilizando diferentes valores de K, teste diferentes métricas de distância (pelo menos 3) e avalie os clusters obtidos.

- Plote os clusters e também os centróides finais obtidos.
- Compare visualmente os clusters com as classes reais do problema.

2. **(Classificação - Base de Dados Bi-dimensional)** Com o objetivo de fixar o conteúdo, você deverá implementar o algoritmo Gaussian Naive Bayes, e aplicar à base de dados para predição de aprovação de um estudante com base nos resultados de 2 avaliações realizadas por ele. A base de dados(arquivo ex2data1.txt utilizada no TP1) contem dados históricos referentes a avaliações passadas, onde as colunas da base são: Avaliação 1, Avaliação 2 e resultado(aprovado ou reprovado). Compare com os resultados obtidos no TP1. Os resultados são melhores do que a regressão logística implementada no TP1?

   Após implementar o algoritmo, mostre a superfície de decisão gerada por ele e compare os resultados(Acurácia média e tempo computacional para o treinamento e o teste) utilizando validação cruzada.

3. **(Classificação - Base Câncer de Mama)** O objetivo agora é implementar o algoritmo Gaussian Naive Bayes e Regressão Logística para resolução do problema de detecção de pacientes com câncer de mama. Os algoritmos deverão ser comparados ao final, onde você deverá apontar os prós e contras dos algoritmos com relação ao desempenho obtido. Não se esqueça de **normalizar** e **separar** os dados(treinamento, validação e teste) para efetuar uma avaliação correta.

   A base de dados a ser utilizada nesta questão é conhecida como **Wisconsin Diagnostic Breast Cancer - WDBC** e foi produzida pelo Dr. William H. Wolberg, pesquisador do departamento de Cirurgia Geral da Universidade de Wisconsin. Ela está disponível publicamente no UCI Machine Learning Repository (<https://archive.ics.uci.edu/static/public/17/breast+cancer+wisconsin+diagnostic.zip>). Esta base de dados é composta por 569 pacientes (357 saudáveis e 212 com câncer) e 32 campos (dos quais 30 atributos são úteis) no total, sendo que o primeiro pode ser descartado, por se tratar do identificador do paciente e o último campo é o rótulo da classe (0 - Saudável, 1 - com Câncer).

O que deve ser entregue: Deve ser entregue um relatório contendo o código e um relatório contendo a análise de cada algoritmo aplicado aos problemas (Máximo de 2 páginas). Um notebook contendo relatório e códigos também pode ser enviado.