Disciplina: Aprendizagem de Máquina
Período: 2021.1
Professor: César Lincoln Cavalcante Mattos

Lista 2 - Regressão logística, métodos estatísticos, KNN e árvores de decisão

Instruções

- Com exceção dos casos explicitamente indicados, os algoritmos e modelos devem ser implementados do início em qualquer linguagem de programação (Python, R, Octave...).
- Pacotes auxiliares (sklearn, matplotlib, etc) podem ser usados somente para facilitar a manipulação dos dados e criar gráficos.
- A entrega da solução pode ser feita via pdf ou Jupyter notebook pelo SIGAA.

Observações

- **Graduação**: No item 1-a abaixo você pode escolher 3 dos modelos para avaliar.
- **Pós-graduação**: Sem mudanças.

Questão 1

Considere o conjunto de dados disponível em **breastcancer.csv**, organizado em 31 colunas, sendo as 30 primeiras colunas os atributos e a última coluna a saída. Os 30 atributos coletados de exames médicos são usados no diagnóstico do câncer de mama, sendo 1 a classe positiva e 0 a classe negativa. Maiores detalhes sobre os dados podem ser conferidos em `https://scikit-learn.org/stable/datasets/toy_dataset.html#breast-cancer-dataset`.

a) Considerando uma validação cruzada em 10 *folds*, avalie modelos de classificação binária nos dados em questão. Para tanto, use as abordagens abaixo:

   - **Regressão logística** (treinado com GD ou SGD);
   - **Análise do discriminante Gaussiano**;
   - **Naive Bayes Gaussiano**;
   - **KNN** (escolha $k = 3$ e distância Euclidiana);
   - **Árvore de decisão** (você pode usar uma implementação já existente com índice de impureza de gini).

b) Para cada modelo criado, reporte valor médio e desvio padrão das métricas de **acurácia**, **revocação**, **precisão** e **F1-score**.