Disciplina: Aprendizagem de Máquina

 ${\bf Período:}\ \ 2023.1$

Professor: César Lincoln Cavalcante Mattos

Lista 2 - Regressão logística e métodos estatísticos

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.

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;
- b) Para cada modelo criado, reporte valor médio e desvio padrão da acurácia global e da acurácia por classe.

Questão 2

Considere o conjunto de dados disponível em **vehicle.csv**, organizado em 19 colunas, sendo as 18 primeiras colunas os atributos e a última coluna a saída. Os 18 atributos caracterizam a silhueta de veículos, extraídos pelo método HIPS (Hierarchical Image Processing System). A tarefa consiste em classificar o veículo em 4 classes (bus, opel, saab, e van). Maiores detalhes sobre os dados podem ser conferidos em https://www.openml.org/search?type=data&sort=runs&id=54.

- a) Considerando uma validação cruzada em 10 folds, avalie modelos de classificação multiclasse nos dados em questão. Para tanto, use as abordagens abaixo:
 - Regressão softmax (treinado com GD ou SGD);
 - Análise do discriminante Gaussiano;
 - Naive Bayes Gaussiano;
- b) Para cada modelo criado, reporte valor médio e desvio padrão da **acurácia global** e da **acurácia por classe**.