Disciplina: Aprendizagem de Máquina
Período: 2025.2 e 2026.1
Professor: César Lincoln Cavalcante Mattos
Monitor: Carlos Gabriel Oliveira Freitas

Lista 1 - Regressão linear, polinomial e regularização

Instruções

- Com exceção dos casos explicitamente indicados, os algoritmos e modelos devem ser implementados do início em qualquer linguagem de programação.
- Pacotes auxiliares (sklearn, matplotlib, etc) podem ser usados somente para facilitar a manipulação dos dados e criar gráficos.
- Para a avaliação do trabalho, recomenda-se o envio de arquivo Jupyter notebook com os códigos executados e os resultados visíveis nas células.

Questão 1

Considere o conjunto de dados disponível em **artificial1d.csv** organizado em duas colunas, x e y. Seja um modelo de regressão linear para $\hat{y} = f(x)$.

a) Apresente os parâmetros do modelo e o MSE (erro quadrático médio) obtidos pelo algoritmo **OLS (mínimos quadrados ordinários)**. Plote a reta resultante sobre os dados.

b) Apresente os parâmetros do modelo, o MSE e a curva de aprendizagem obtidos pelo algoritmo **GD (gradiente descendente)**. Plote a reta resultante sobre os dados.

c) Apresente os parâmetros do modelo, o MSE e a curva de aprendizagem obtidos pelo algoritmo **SGD (gradiente descendente estocástico)**. Plote a reta resultante sobre os dados.

Questão 2

Considere o conjunto de dados disponível em **california.csv**, organizado em 9 colunas, sendo as 8 primeiras colunas os atributos e a última coluna a saída. Os 8 atributos são usados na predição da mediana de preços de casas em distritos da Califórnia na década de 1990. Maiores detalhes sobre os dados podem ser conferidos em `https://scikit-learn.org/stable/datasets/real_world.html#california-housing-dataset`.

a) Aleatoriamente, divida o conjunto de dados em treino (80%) e teste (20%).

b) Treine 13 modelos de **regressão polinomial**, com ordens de 1 a 13. Você pode usar o algoritmo OLS.

c) Reporte o RMSE (raiz quadrada do erro quadrático médio) no treinamento e no teste para cada modelo. Faça um gráfico para o treino e um gráfico para o teste.

d) Repita os 2 itens anteriores incluindo um termo de **regularização L2** (por exemplo, com fator $\lambda = 0.01$).

Nota: Normalize os dados (a saída com StandardScaler e as entradas com MinMax) antes do treinamento/teste (antes de criar os regressores polinomiais) e "desnormalize" a saída antes de calcular o RMSE.

Questão 3

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 um modelo de Regressão Logística (treinado com GD ou SGD) nos dados em questão.

b) Reporte valor médio e desvio padrão da **acurácia global** e da **acurácia por classe**.

Questão 4

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 um modelo de Regressão Softmax (treinado com GD ou SGD) nos dados em questão.

b) Reporte valor médio e desvio padrão da **acurácia global** e da **acurácia por classe**.