

Disciplina: Aprendizagem de Máquina
Período: 2022.1
Professor: César Lincoln Cavalcante Mattos

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 (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 **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 11 modelos de **regressão polinomial**, com ordens de 1 a 11.

- 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** com fator $\lambda = 0.01$.

Nota: Normalize os dados antes do treinamento e “desnormalize” antes de calcular o RMSE.