Módulo 3 – Aprendizado de Máquina Supervisionado I

Trabalho 1
House Pricing

1 Descrição do Dataset

Nestre trabalho você irá trabalhar com o dataset *California Housing Prices*, um conjunto de anotações a respeito de imóveis de diversos distritos da California (baseado em um censo de 1990) e os preços medianos de venda naquele distrito. As anotações disponíveis são:

- **Longitude**;
- **Latitude**;
- **Idade mediana dos imóveis do distrito**;
- **Número de cômodos no distrito**;
- **Número de quartos no distrito**;
- **População do distrito**;
- **Número de imóveis familiares no distrito**;
- **Renda mediana do distrito**;
- **Proximidade com o oceano**;
- **Preço mediano dos imóveis do distrito** (valor alvo que queremos prever).

2 Tarefas

Pedimos que você:

1. Inspecione os dados. Quantos exemplos você tem? Como você irá lidar com as features discretas? Há exemplos com features sem anotações? Como você lidaria com isso?
2. Normalize os dados de modo que eles fiquem todos no mesmo intervalo.
3. Como *baseline*, faça uma regressão linear para predizer os preços. Calcule o erro nos conjuntos de treino e validação.
4. Implemente soluções alternativas baseadas em regressão linear através da combinação dos features existentes (multiplicação e/ou divisão) para melhorar os resultados obtidos no baseline. Compare suas soluções nos conjuntos de treino e validação.
5. Implemente soluções alternativas baseadas em regressão polinomial (elevando o grau de features) para melhorar os resultados obtidos no baseline. Plote o erro no conjunto de treino e de validação pelo grau do polinômio.
6. Escreva um relatório de no máximo 3 páginas:
   (a) Descreva o que foi feito, bem como as diferenças entre o seu melhor modelo e o baseline;
   (b) Reporte os resultados do melhor modelo obtido no conjunto de de teste (este último será disponibilizado 1 dia antes do prazo final de submissão) e compare possíveis diferenças nos resultados. Ocorreu overfitting no treinamento?
   (c) Escreva pelo menos 1 parágrafo com as conclusões tiradas na atividade;

3 Arquivos

Os arquivos disponíveis no Moodle são:

- *housePricing_ trainSet.csv*: conjunto de dados para treinamento;
- *housePricing_ valSet.csv*: conjunto de dados para validação;
- *housePricing_ testSet.csv* **(será disponibilizado 1 dia antes do prazo final da submissão)**: conjunto de dados retido pelo professor;

4 Avaliação

O dataset foi previamente dividido aleatoriamente em três conjuntos — treino, validação e teste — e apenas os dois primeiros serão disponibilizados para que você implemente as suas soluções.

Um dia antes do prazo final de submissão, iremos disponibilizar no Moodle o conjunto de teste e iremos avisá-los pelo canal da disciplina no Slack. No relatório, você deve reportar os seus resultados no conjunto de validação e no conjunto de teste.

A avalidação consistirá da análise do relatório e do código submetidos no Moodle. Iremos avaliar se as tarefas pedidas foram realizadas, como o treinamento e validação foi feito, os resultados reportados e as conclusões feitas.

Observações sobre a avaliação:

- O trabalho poderá ser feito individualmente ou em duplas, podendo haver repetição das duplas a cada trabalho;
- O código e o relatório deverão ser submetidos no Moodle por **apenas um integrante da dupla**;
- Não se esqueçam de listar os nomes dos integrantes da dupla no início do relatório;
- As notas do trabalho serão divulgadas em até uma semana após o prazo da submissão;