INF-0615 – Aprendizado de Máquina Supervisionado I

Exercício 1
House Pricing

1 Descrição do Dataset

Neste exercício 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) com os respectivos preços de venda. As anotações disponíveis são:

- **Longitude**;
- **Latitude**;
- **Idade mediana dos imóveis do distrito**;
- **Número médio de cômodos nos imóveis do distrito**;
- **Número médio de quartos nos imóveis do 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 melhor preparados para o treinamento.
3. Como *baseline*, faça uma regressão linear para predizer os preços dos imóveis. Calcule o erro nos conjuntos de treino e validação.
4. Implemente soluções alternativas baseadas em regressão linear através da combinação das features existentes (multiplicação, divisão, etc.) para melhorar os resultados obtidos no baseline. Compare suas soluções nos conjuntos de treino e validação. Se preferirem, vocês podem utilizar o ChatGPT para propor possíveis combinações não-lineares de features.
5. Implemente soluções alternativas baseadas em regressão polinomial (elevando o grau das features) para melhorar os resultados obtidos no baseline. Plote o erro no conjunto de treino e de validação pelo grau do polinômio.
6. Treine novamente a regressão linear, mas agora utilizando a Descida do Gradiente. Varie a Taxa de Aprendizado (*Learning Rate*) e o número de iterações no treinamento. Reporte os erros no conjunto de validação.
7. Tome os melhores modelos desenvolvidos no trabalho e reporte o erro no conjunto de teste. Os resultados diferem muito do conjunto de validação?

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*: conjunto de dados para teste final;
- *Ex01.R*: código que implementa a solução do problema explorando combinação de features e features polinomiais;
- *Ex01.R*: código que implementa a solução do problema utilizando Descida do Gradiente;