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 seus preços médios de venda. As anotações disponíveis são:

- **Longitude**;
- **Latitude**;
- **Idade do imóvel**;
- **Número de cômodos**;
- **Número de quartos**;
- **População do distrito em que a casa está localizada**;
- **Número de imóveis familiares no distrito**;
- **Renda média do distrito**;
- **Proximidade com o oceano**;
- **Preço médio do imóvel** (valor alvo que queremos prever).

2 Tarefas

Para as tarefas deste trabalho, você podem utilizar a função **lm** para fazer a regressão linear como fizemos em sala. Neste trabalho 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 o valor médio da casa. Calcule o erro no conjunto de teste.
4. Implemente soluções alternativas mais poderosas baseadas em regressão linear (através da combinação dos features existentes) e compare-as com o baseline.

Como atividades extras que valem pontos para melhorar a nota:

1. Explore o comportamento da função de custo no conjunto de treinamento a medida que as iterações progridem e analise a complexidade do modelo. Quais são as suas conclusões? Quais seriam os seus próximos passos após estas análises?
2. Use diferentes taxas de aprendizado (α) durante a otimização por Descida do Gradiente (DG). Como elas afetam a convergência do treinamento?
3. Se possível, compare soluções baseadas em DG com Equações Normais. Quais são as suas conclusões?

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* **(apenas durante a avalidação)**: conjunto de dados retido pelo professor a ser utilizado na avaliação;
- *aux_ linearRegression.r*: código R com funções auxiliares (descida do gradiente, equações normais).

4 Avaliação

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

No dia da avaliação, nos 40 minutos finais da aula, o professor irá disponibilizar o conjunto de teste e validar a solução de cada aluno (ou dupla de alunos). A avalidação consiste da análise dos resultados reportados sobre o conjunto de teste e possíveis perguntas a respeito das decisões tomadas pelo aluno durante a implementação de sua solução.

Observações sobre a avaliação:

- **NÃO** haverá submissão do código ou relatório do trabalho no Moodle;
- O trabalho poderá ser feito em duplas, podendo haver repetição das duplas a cada trabalho;
- Pelo menos um membro da dupla deverá estar presente no momento da avaliação;
- As notas do trabalho serão divulgadas em até uma semana após a avaliação;