Exercício 1 - Regressão Linear
Abalone dataset

1 Descrição do Problema

Abalone é um gênero de moluscos marinhos, cujas conchas possuem uma estrutura em espiral, caracterizada pela presença de diversos poros respiratórios alinhados próximos a sua extremidade. Devido a variedade de cores que elas podem assumir, as conchas são vistas como artigos de decoração e bijuteria, além da carne dos abalones ser valorizada na gastronomia, em especial nos países asiáticos [1].

A idade de um abalone pode ser determinada cortando-se a concha, colorindo-a e contando o número de anéis por meio de um microscópio — uma tarefa bastante demorada. Porém, outras características, como o comprimento ou diâmetro da concha, são mais fáceis de serem medidas e podem ser utilizadas para estimar a idade do molusco.

Neste exercício, nós desejamos determinar a idade de um abalone analisando tais medidas. O número de anéis internos da concha pode ser um *proxy* para a sua idade (*vide a variável "rings" no arquivo "abalone.names"*), então o nosso método irá predizer a quantidade de anéis, através do *gênero, altura, comprimento, diâmetro* e *peso* do abalone.

2 Tarefas

Neste exercício, pedimos que você:

1. Inspecione os dados. Quantos exemplos você tem? Quais são as features disponíveis?
2. Particione os dados em conjuntos de treinamento e de teste para reportar seus resultados e evitar overfitting.
3. Como uma primeira solução de referência (*baseline*), faça uma regressão linear sobre as features para predizer o número de anéis da concha. 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.
5. Faça um gráfico do valor da função de custo no conjunto de treinamento pelo número de iterações e analise a complexidade do modelo. Quais são as suas conclusões? Quais seriam os seus próximos passos após estas análises?
6. Use diferentes taxas de aprendizado (α) durante a otimização por Descida do Gradiente (DG). Como elas afetam a convergência do treinamento?
7. Se possível, compare soluções baseadas em DG com Equações Normais. Quais são as suas conclusões?
8. Tome cuidado com as variáveis que são **discretas**. Como você as adicionaria ao seu conjunto de features do modelo?

3 Arquivos

Os arquivos disponíveis no Moodle são:

- *abalone.data*: contém os dados que serão utilizados no exercício;
- *abalone.names*: contém uma breve descrição do conjunto de dados;
- *aux_ linearRegression.r*: código R com métodos auxiliares (descida do gradiente, equações normais);

4 Referências

1. *Abalone*. De Wikipedia, em `https://en.wikipedia.org/wiki/Abalone`.
2. *Abalone dataset*. UCI Machine Learning Repository. `https://archive.ics.uci.edu/ml/datasets/abalone`.