Universidade Federal de Santa Catarina

Relatório I:
Implementação da Regressão Linear em Python

Aluno: Jean Carlo Hilger

Junho
2021

Sumário

1 Descrição das Atividades

A implementação da regressão linear, executada com a linguagem python, deu-se de forma similar para os cenários com uma e múltiplas variáveis. Com a biblioteca `numpy`, foi possível utilizar vetorização e por tanto, as duas soluções possuem o mesmo código.

O cálculo da função de custo é executado utilizando á seguinte expressão:

```
J = np.sum((np.dot(theta, X.T) - y) ** 2) / (2 * m)
```

Analogamente, a atualização do vetor `tetha` é concretizada utilizando as expressões:

```
h = np.dot(theta, X.T)
step = alpha * np.sum((h - y) * X.T, axis=1) / m
theta = theta - step
```

onde `np` refere-se à biblioteca `numpy`, `theta` é o vetor de parâmetros da regressão linear (inicializado com valores nulos) e `X` e `y` correspondem às *features* e *labels* do conjunto de dados, respectivamente.

2 Análise dos Resultados

Nesta seção, é feita uma breve análise sobre os resultados obtidos com as implementações. Empregou-se dois *datasets* distintos, um para cada cenário, conforme descrito a seguir.

2.1 Regressão Linear de Uma Variável

Para a regressão linear de uma variável, utilizou-se um *dataset* que descreve o lucro de uma franquia de restaurantes em diferentes cidades. Como única *feature*, consta a população da cidade, em dezenas de milhares de habitantes. O valor alvo (*label*) descreve o lucro do restaurante na cidade, em dezenas de milhares de dólares.

A normalização não surtiu efeito para este caso por dois motivos centrais: o *dataset* possui apenas uma *feature*, esta feature possui valores extremos (máximo e mínimo) muito próximos.

A Figura 1 evidencia a evolução do erro do algoritmo ao longo do processo de treinamento. Executou-se o algoritmo com taxa de aprendizagem (`alpha`) igual à `0.01` e `1500` iterações.

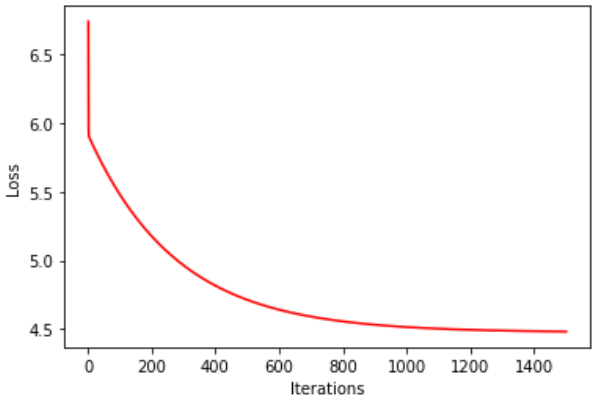

Figura 1: Progressão do erro à cada iteração de treinamento, para regressão linear de uma variável.

2.2 Regressão Linear de Múltiplas Variável

Na regressão linear de múltiplas variáveis fêz-se uso de um *dataset* que descreve preços de casas. As duas *features* são o tamanho (em pés quadrados) e o número de quartos, ao passo que a *label* descreve o valor do imóvel.

Similarmente ao caso anterior, foi produzido um gráfico (Figura 2) que exibe a progressão do erro ao longo do processo de treinamento. O algoritmo do gradiente descendente foi computado com taxa de aprendizagem (`alpha`) igual à `0.5` e `1500` iterações.

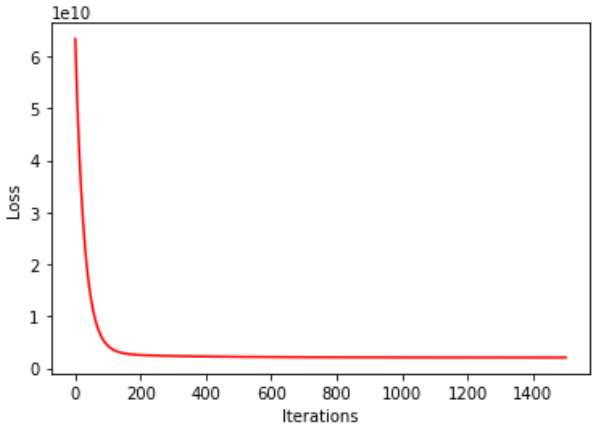

Figura 2: Progressão do erro à cada iteração de treinamento, para regressão linear de múltiplas variáveis.

Neste segundo cenário, a normalização foi de grande valia, visto que as duas features possuem valores muito discrepantes. Para o caso do algoritmo em questão (implementação junto ao arquivo) sem a normalização, erros de *overflow* ocorreram.