CEFET-MG
CENTRO FEDEREAL DE EDUCAÇÃO TECNOLÓGICA DE
MINAS GERAIS

Trabalho Prático 2

Regressão Linear.

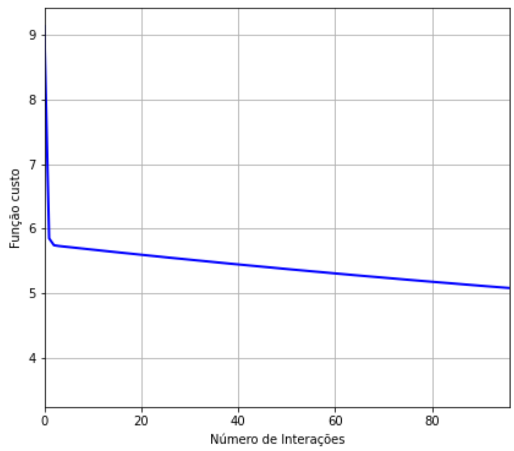
Autores: Samuel Santos

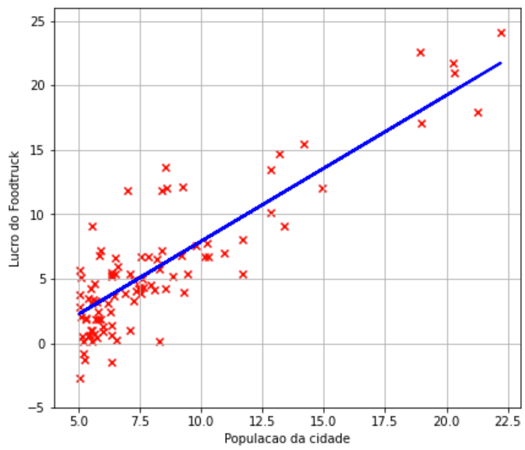
Data Entrega: 20 de junho de 2021

1. Introdução:

Regressão linear é o processo de traçar uma reta através dos dados em um diagrama de dispersão. A reta resume esses dados, o que é útil para fazermos previsões. O trabalho prático desenvolvida tem como objetivo o estudo do comportamento da regressão linear com uma ou múltiplas variáveis e a análise do uso do gradiente e da equação normal para o estudo desse comportamento.

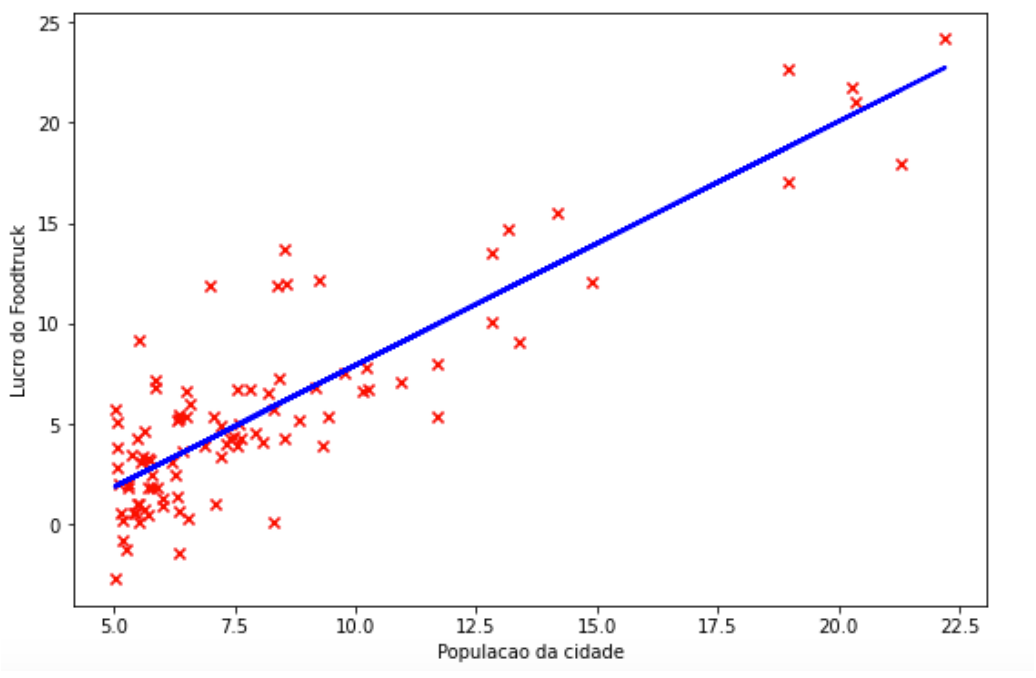
2. Resultado e Discussões:





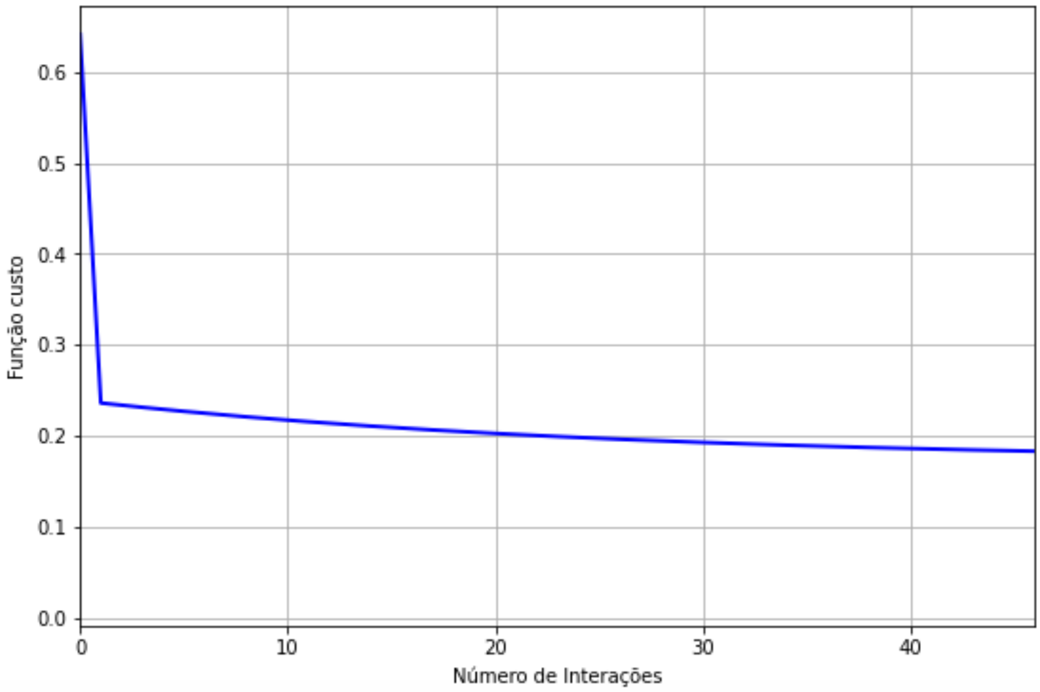
As imagens acima mostram a função custo a esquerda e a regressão linear encontra usando o calculo do gradiente. A função hipótese foi adquirida com 1000 interações e um valor alfa para o gradiente de 0,001.

Para encontrar a regressão linear usando o gradiente é importante que as dimensões dos dados estejam em uma ordem de grandeza muito próximas ou sejam iguais. No exemplo acima, os valores reais são da ordem de grandeza 10^1 variando pouco os seus valores em um intervalo de números inteiros. Para validar se a regressão linear encontrada foi satisfatória foi utilizado outro método utilizando uma função da biblioteca sklearn para analise.



Podemos observar que o uso da função LinearRegression resultou em uma regressão bem semelhante e próxima utilizando o gradiente.

Regressão linear com múltiplas variáveis:



O gráfico acima mostra a função custo encontrada utilizando uma função hipótese para múltiplas variáveis. Vale destacar, que os resultados obtidos entre as funções custo foram parecidos o que mostram que tanto para uma ou múltiplas variáveis é possível traçar um ajuste linear. Sendo necessário, em alguns casos de múltiplas variáveis, a utilização de uma correção polinomial.

Para encontrar a função custo acima foi utilizado o gradiente com o mesmo número de interações (1000) e um alpha pequeno 0.001, vale destacar que a ordem de grandeza entre os valores reais do segundo conjunto de dados era distante, então foi necessário o uso da normalização, subtraindo de cada valor a média entre eles dividido pelo desvio padrão.

Pergunta: Veja que agora não é possível traçar o ajuste linear como no exercício anterior. Por quê?

Pelo fato da existência de mais de duas dimensões os dados não se correlacionam em um plano cartesiano bidimensional, uma maneira de projetarmos uma regressão linear é usando uma correção polinomial, que pode ser feita adicionando uma característica ou seja, elevando um dos termos ao quadrado ou, nesse problema em questão, promover uma combinação linear entre dois valores na hipótese, do tipo x1*x2. Vale ressaltar, que a escolha desses dados assim normalizados se torna arbitrária e os valores de saída podem se diferir consideravelmente e outro fato, não menos importante, é que no problema em questão foi utilizado x1 (tamanho) e x3 (preço) como multiplicadores de tetas, a escolha de outros como x2(quantidade de quartos) podem ocasionar uma diferente resposta.

Comparativo entre Equação Normal e Gradiente:

```
Valores encontrados com a equação normal:

[[ 0.44158621 -0.24514452  0.36429822 -0.00543758  0.27284419  0.21301947
   0.06704366 -0.43882246 -0.06192075 -0.25538832 -0.22920632 -0.26738177
  -0.09867891 -0.37656515 -0.10484804 -0.12430615  0.31199081 -0.00933673
  -0.09947466 -0.09877469  0.07911366  0.25471385  0.03612966  0.07736594
  -0.05387175  0.10222934 -0.09912758  0.01686792 -0.16245258 -0.36097325
   0.20454085  0.20824875  0.17441627 -0.24374583 -0.4034675   0.17258552
  -0.14845282  0.1304157   0.3034679   0.05498149 -0.42352829  0.1995814
   0.16239531 -0.29657653  0.64147597  0.21316951]]

Valores encontrados com o gradiente:

[-0.18491378 -0.14832175 -1.29020232  1.02528439  0.80067857 -0.19895164
 -0.30741847 -0.29525139 -0.26670773  0.74461977 -0.16891062 -0.18482113
  2.18868968 -0.25042381  0.94105721 -1.22665759 -0.30657526  0.9879418
  1.08059357 -0.25697481 -1.26867759 -0.26633339  0.76324427  0.04343546
 -0.25978238 -0.05894735 -0.05473599 -0.04912085 -0.21298951 -1.17986471
 -2.36198715  0.81471644  0.04905061  0.78757657 -0.25978238 -0.27849954
  0.84288575  1.03380069  0.78860601 -1.16245776 -0.18491378  0.81387417
 -0.21383178 -1.33896049  0.80067857 -0.2695153 ]
```

A imagem acima mostra os valores encontrados utilizando as duas abordagens, tanto o gradiente quanto a equação normal.

```
Erro relativo da função usando o gradiente: 3.1762416614724747
Erro relativo da função usando a equacao normal: 3.2669913428338906

Diferença entre os erros encontrados: 0.09074968136141592
```

Vale destacar que a diferença entre os erros é bem próxima usando qualquer um dos valores reais, ou seja, se usarmos tanto o tamanho do imóvel ou a quantidade de quartos ou preço do imóvel, a diferença dos erros será mínima. Portanto, o uso da equação normal é um recurso mais eficiente para traçar o ajuste linear pois, não é necessária a normalização dos valores reais que, dispostos em ordem diferentes, fazem com que o uso do recurso do gradiente para traçar o ajuste linear seja inconsistente.

Convém lembrar, que o fato de ter colocado "erro relativo" é por causa do uso de algum dos valores reais na função de erro quadrático, se os valores forem preço ou tamanho do imóvel, os erros relativos ficarão maiores, entretanto, como mencionado anteriormente, a diferença entre os erros continuará pequenas.