Cálculo numérico

Trabalho Prático 1

Heitor Lourenço Werneck
heitorwerneck@hotmail.com

5 de Dezembro de 2019

Conteúdo

1 Introdução **1**

2 Base de dados **1**
2.1 Análise 1

3 Modelagem e solução **2**

4 Análise de resultados **4**

5 Conclusão **6**

1 Introdução

Regressão linear múltipla é um método estatístico que permite a criação de um modelo linear com diversas variáveis que descreve ou se a aproxima do comportamento dos dados de entrada.

A aplicação de regressão linear múltipla é das mais diversas como por exemplo: previsão de preço de casas, previsão de temperatura máxima em um dia e previsão de valor de itens com base em algumas de suas características.

Esse trabalho tem como objetivo aplicar o método de regressão linear com múltiplas variáveis em uma base, que no caso é uma base de 50 *startups* e o objetivo será predizer o ganho das *startups*.

2 Base de dados

A base de dados contem as informações mostradas na tabela 1, um problema para a regressão seria o tipo texto porém essa variável pode ser transformada para variáveis *dummies* tal que se o valor na tabela original for "New York" somente uma variável "State_New York" terá valor 1 e as outras 0, essa lógica é mostrada na tabela 2.

Coluna	Tipo
R&D Spend	Ponto flutuante
Administration	Ponto flutuante
Marketing Spend	Ponto flutuante
State	Texto
Profit	Ponto flutuante

Tabela 1: Dados na base.

State	State_New York	State_California	State_Florida
New York	1	0	0
California	0	1	0
Florida	0	0	1

Tabela 2: Variáveis *dummies*, mapeamento de valores.

2.1 Análise

O número de amostras dessa base é 50 como dito na introdução, para o contexto de aprendizado de maquina não é muito porém será o suficiente para esse caso.

Uma visão geral da base de dados é dada na figura 1, como pode ser visto pela média a cidade com menos *startups* é a florida.

A magnitude dos valores são grandes como pode ser visto nas outras colunas.

O desvio padrão do gasto em "Administration" é o menor comparado aos outros valores.

Em "R&D Spend" e "Marketing Spend" existe *startup* com 0, que pode ser considerado uma empresa que não tem motivo de gastar nesses campos e talvez seja um *outlier* da base de dados.

Já em "Profit" e "Administration" todos tiveram valores maiores que 0.

A maior média de gasto de todas empresas é em "Marketing" mesmo que algumas empresas não gastem com essa parte o valor de todas outras empresas que gastam compensa.

Após ser feito uma regressão será possível inferir, não com completa exatidão, qual o gasto que mais gera lucro.

	mean	std	min	max
R&D Spend	73721.62	45902.26	0.0	165349.2
Administration	121344.64	28017.8	51283.14	182645.56
Marketing Spend	211025.1	122290.31	0.0	471784.1
Profit	112012.64	40306.18	14681.4	192261.83
State_California	0.34	0.48	0.0	1.0
State_Florida	0.32	0.47	0.0	1.0
State_New York	0.34	0.48	0.0	1.0

Figura 1: Medidas da base de dados.

3 Modelagem e solução

Para ser feito a regressão linear com múltiplas variáveis será utilizado o modelo $G^T \cdot G \cdot \alpha = G^T \cdot y$ que dará o modelo com menor distância dos pontos dados (método dos mínimos quadrados) tal que y é uma matriz coluna da variável a ser predita, G é a matriz da base de dados, porém sem a coluna da característica a ser predita, e α é os parâmetros a serem descobertos para criação da função preditora.

No caso para ser realizado essas operações é necessário definir o transposto de uma matriz e a multiplicação de matrizes.

Para ser feito a transposta da matriz é utilizado o algoritmo 1.

O algoritmo 2 descreve o funcionamento da multiplicação de matrizes. O ponto fundamental do algoritmo ocorre na linha 6 que é o laço que soma a multiplicação dos elementos da linha i da matriz 1 com os elementos da coluna j da matriz 2 e atribui essa soma na linha 8 a posição i e j da matriz resultante.

Após ser feito a transposta e multiplicação da matriz só falta um método para solucionar um sistema de equação linear.

O método escolhido foi o Gauss-Seidel pois o mesmo é eficiente e converge rápido. O algoritmo 3 mostra o funcionamento.

Algorithm 1 Matriz transposta.

```
1: procedure TRANSPOSTA(Matriz)
2:     Matriz^T ← NovaMatriz(Matriz.colunas, Matriz.linhas)  ▷ Cria matriz com "Matriz.colunas" colunas e "Matriz.linhas" linhas
3:     for j = 0 to Matriz.colunas − 1 do
4:         for i = 0 to Matriz.linhas − 1 do
5:             Matriz^T[j, i] = Matriz[i, j]
6:     return Matriz^T
```

Algorithm 2 Multiplicação de matrizes.

```
1: procedure MULTIPLICACAO(M1, M2)
2:     MResultante ← NovaMatriz(M1.linhas, M2.colunas)        ▷ Cria matriz com zeros
3:     for i = 0 to MResultante.linhas − 1 do
4:         for j = 0 to MResultante.colunas − 1 do
5:             soma = 0
6:             for k = 0 to M1.colunas − 1 do
7:                 soma ← soma + M1[i, k] · M2[k, j]
8:             MResultante[i, j] ← soma
9:     return MResultante
```

Algorithm 3 Método de resolução de sistema de equação linear.

```
1: procedure GAUSSSEIDEL(A, b, Erro = 0.0001)
2:     X_Velho ← [0..b.linhas − 1]        ▷ Começa com todos elementos iguais a zero
3:     DistanciaMaxima ← ∞
4:     while DistanciaMaxima > Erro do
5:         X_Novo ← [0..b.linhas − 1]
6:         for i = 0 to A.linhas − 1 do
7:             X_Novo[i] ← b[i][0]
8:             for j = 0 to A.colunas − 1 do
9:                 if i ≠ j then
10:                    if j < i then
11:                        X_Novo[i] ← X_Novo[i] − A[i][j] · X_Novo[j]
12:                    else
13:                        X_Novo[i] ← X_Novo[i] − A[i][j] · X_Velho[j]
14:            X_Novo[i] ← X_Novo[i]/A[i][i]
15:        DistanciaMaxima ← 0
16:        for i = 0 to b.linhas − 1 do        ▷ Calcula o erro pela distância
17:            DistanciaMaxima ← max(DistanciaMaxima, |X_Novo[i] − X_Velho[i]|)
18:        X_Velho ← X_Novo
19:    return X_Novo
```

Com todos esses componentes e métodos basta utilizá-los para solucionar o problema. O algoritmo 4 é a solução final para o problema.

Algorithm 4 Regressão linear com múltiplas variáveis.

1: **procedure** RegressaoLinear(G, y)
2: $G^T \leftarrow Transposta(G)$
3: $G^T G \leftarrow Multiplicacao(G^T, G)$
4: $G^T y \leftarrow Multiplicacao(G^T, y)$
5: $Parametros \leftarrow GaussSeidel(G^T G, G^T y)$

4 Análise de resultados

Primeiro foi separado os dados em 70% para treino e 30% para teste.

O treino foi utilizado para treinamento do modelo com o algoritmo de regressão linear com múltiplas variáveis apresentado.

Depois do treinamento a função preditora da equação 4.1 foi obtida.

$$\begin{aligned} Profit(R\&DSpend, Administration, ...) = (0.823) * R\&DSpend + (-0.0483) * Administration \\ +(0.0301) * MarketingSpend + (5.17e+04) * dummy \\ +(-9.32e+02) * State\_California + (-5.92e+02) * State\_Florida \\ +(1.98e+03) * State\_New\ York \end{aligned} \tag{4.1}$$

Os parâmetros da função são mostrados na figura 2 de maneira mais intuitiva, primeiro é possível observar que a cidade que está mais relacionada com um lucro alto é New York, as outras cidades se relacionam muito menos com um lucro alto que New York.

A variável dummy é proximo de 50000 o que mostra que a linha de base de lucro das empresas dessa base de dados é esse valor.

Já as outras variáveis que tem a ver com o gasto em certas áreas mostrou que o gasto em "R&D" é o que da maior lucro para as empresas nesse contexto. Isso mostra que as empresas gastam muito em Marketing(Discutido na secção da base de dados) e pouco em pesquisa e desenvolvimento(R&D) e um investimento em R&D poderia melhorar seus lucros em um longo prazo.

Essa análise dos parâmetros não é muito precisa porém são algumas ideias interessantes que na prática podem ajudar empresas a onde investir, se forem comparadas a empresas que tem segmentos semelhantes, pois pode existir empresa que em certo campo realmente não faz sentido investir. Nesse caso as empresas podem não ser semelhantes e as análises feitas podem não ser verdadeira para todas.

A figura 3 mostra o erro relativo e absoluto para os dados do teste. Como pode-se ver o erro relativo foi bem baixo mostrando que os valores foram bem proximos dos valores reais.

Já o erro absoluto é grande pois ele não é normalizado e isso torna difícil a análise porém pela coluna de valores reais e preditos é possivel ver que o erro foi baixo.

A figura 4 mostra os erros obtidos com medidas estatísticas, a média do erro relativo é 0.04 o que é bem baixo juntamente com o desvio padrão que também é 0.04, ou seja, o modelo criado se mostrou bem preciso.

Na figura 5 os erros são mostrados, tanto com a predição no teste e no treino. Por ela pode-se ver que o erro realmente foi baixo e a função de predição está descrevendo bem os dados. Essa comprovação também foi feita pelo R^2, isso pode ser visto na tabela 3 que mostra um R^2 proximo de 1, que significa que o modelo descreve bem os dados tanto no teste como no treino.

	R^2	Adjusted R^2
Test	0.93	0.85
Train	0.94	0.92

Tabela 3: R^2 obtidos aplicando no treino e teste.

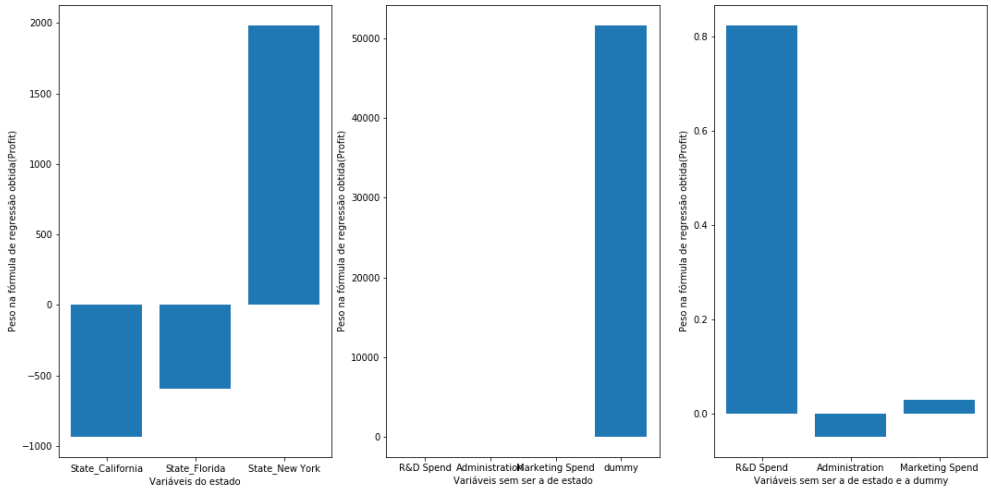

Figura 2: Parâmetros.

	real	predicted	absolute_error	relative_error
0	156122.51	158242.74	2120.23	0.01
1	124266.9	129926.73	5659.83	0.05
2	156991.12	168262.16	11271.04	0.07
3	192261.83	197294.33	5032.5	0.03
4	191792.06	190578.09	1213.97	0.01
5	97427.84	98053.78	625.94	0.01
6	122776.86	117346.54	5430.32	0.04
7	134307.35	127486.86	6820.49	0.05
8	108733.99	110699.66	1965.67	0.02
9	125370.37	133030.66	7660.29	0.06
10	129917.04	149839.73	19922.69	0.15
11	152211.77	155031.48	2819.71	0.02
12	108552.04	116476.5	7924.46	0.07
13	103282.38	100155.78	3126.6	0.03

Figura 3: Erros na predição dos dados no teste.

	real	predicted	absolute_error	relative_error
count	14.0	14.0	14.0	14.0
mean	136001.0	139458.93	5828.12	0.04
std	30264.18	31463.91	5047.52	0.04
min	97427.84	98053.78	625.94	0.01
25%	112244.71	116694.01	2295.1	0.02
50%	127643.7	131478.7	5231.41	0.04
75%	155144.83	157439.92	7450.34	0.06
max	192261.83	197294.33	19922.69	0.15

Figura 4: Erros na predição dos dados no teste com medidas estatísticas.

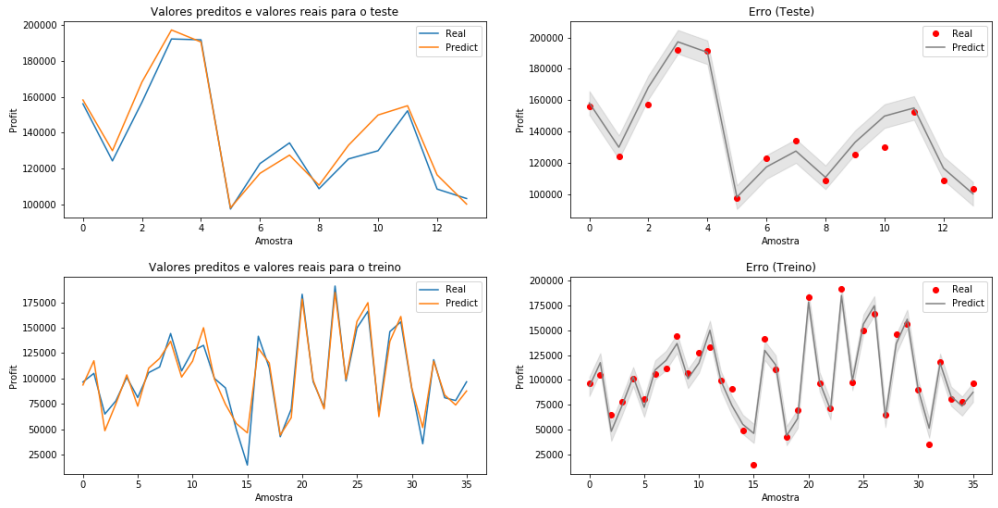

Figura 5: Erros.

Outros métodos de se calcular o erro foram abordados também e todos deram resultados positivos de que os erros foram baixos. (tabela 4)

MSE	57624666.284417234
RMSE	7591.091244637838
RRMSE	0.06956706362289085
MAE	5828.124952610224
MAPE	0.04422778830100514
RMSLE	0.05537110299114828

Tabela 4: Diversas métricas de erro.

5 Conclusão

Foi possível observar que a regressão linear múltipla pode dizer muito sobre variáveis que influenciam em outras e para o problema a regressão se mostrou capaz de descrever os dados bem e com baixo erro, ou seja, o problema pode ser resolvido com regressão.

Para trabalhos futuros seria interessante comparar com outros métodos de aprendizado de maquina como por exemplo gradiente descendente e redes neurais para ver se os erros podem ser diminuídos com modelos mais complexos.