Universidade Federal do Rio de Janeiro
Bacharelado em Ciência da Computação
Inteligência Artificial

Relatório:
Aprendizagem de Máquina com “Discrimination in salaries” dataset

Tainá da Silva Lima - DRE: 116165607
Rio de Janeiro
2019

SUMÁRIO

I. INTRODUÇÃO

Este relatório é referente ao trabalho da disciplina de Inteligência Artificial, ministrada pelo Prof. Mário Benevides. O trabalho foi passado com o intuito de colocar-se em prática os ensinamentos dados em aula sobre aprendizagem de máquina. Sendo assim, este consiste em utilizar o dataset "Discrimination in salaries" fornecido para tal objetivo. Para resolução deste trabalho foi aplicado o conceito de regressão linear e seu código fonte foi feito usando a linguagem Python.

O propósito deste documento é relatar o objetivo do trabalho, a metodologia utilizada, bem como quais foram os resultados encontrados e as conclusões geradas por estes.

II. OBJETIVO

O objetivo deste trabalho é, utilizando algum algoritmo de aprendizagem de máquina apresentado em aula e baseado num conjunto de dados de entrada, prever quais seriam os salários dos professores de uma escola de pequeno porte.

Tal banco de dados consiste de 41 entradas, onde cada linha é contém informações sobre sexo, cargo ou rank (professor assistente, associado ou integral), tempo em que esteve no cargo (em anos), maior título de ensino superior (mestre ou doutor), tempo desde que este este título foi adquirido (em anos) e seu salário anual (em dólares) de um professor da escola. Cada entrada é formatada da seguinte maneira:

```
  sx        rk yr        dg yd    sl
male      full 25 doctorate 35 36350
```

Figura 1: Exemplo de entrada

- A primeira coluna representa o sexo, feminino ou masculino, em que foi associado 0 e 1 aos dois respectivamente,
- A segunda coluna representa o cargo, onde 0,1 e 2 estão relacionados respectivamente com professor assistente, associado e integral.
- A terceira coluna representa o título de ensino superior, onde 0 e 1 estão relacionados respectivamente com mestre e doutor.
- A última e quarta coluna representa o salário dos professores, visto como um número real (float)

III. METODOLOGIA

Para solucionar este problema, foi utilizado o método de regressão linear, especificamente mínimo quadrados. Além disso, o algoritmo foi criado usando a linguagem de programação Python (Versão 3.x) e algumas de suas bibliotecas tais como Numpy, OS e matplotlib.

A. Regressão e mínimos quadrados

Tal método na verdade é um conceito estatístico, vindo da análise de regressão. Esta é uma área da estatística que estuda a relação entre um conjunto de variáveis independentes e uma variável dependente destas. Em específico nesse problema, temos que queríamos construir um modelo de regressão linear múltiplo, ou seja, queremos construir um modelo que nos diga como se dá tal relação, onde há mais de uma variável independente mas que esta seja linear.

De maneira menos conceitual, dado um conjunto de dados onde para cada um deles temos valores para as variáveis dependentes e os valores associados para as variáveis independentes, queremos encontrar uma "reta" (hiperplano) que aproxima tais dados.

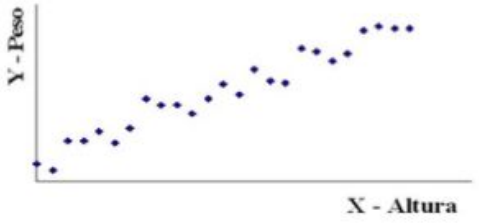

Figura 2: Exemplo de dados que tendencionam a aproximação por uma reta

O método de mínimos quadrados foi utilizado pois não só queremos uma reta mas sim a reta que melhor aproxima esses dados. Este método utiliza a ideia de que a "reta" que tem uma boa aproximação é aquele que possui o menor erro, onde este é calculado como a distância dos pontos até a "reta" e essa distância elevado ao quadrado.

$$SQ(\alpha, \beta) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

Figura 3: Fórmula para o cálculo do erro

B. O algoritmo

Algoritmicamente falando, o problema foi resolvido da maneira apresentada em sala de aula e escrita nos slides sobre aprendizado de máquina disponibilizado aos alunos, sendo criado uma função “linearRegression(w,xTrData, yTrData)” que retorna o w (os coeficientes) do hiperplano que aproxima os dados.

Além disso, foi criado funções para o tratamento e leitura dos dados para o formato de matriz e outra função “test” onde é verificado, para as entradas de teste, qual seria o valor do salário dos professores previsto pela regressão linear feita. Nessa mesma função também é calculado o erro para cada instância, bem como o erro total (acumulado).

IV. RESULTADOS

Analisando os resultados gerados pelo algoritmo de regressão linear feito, é plausível dizer que obtive um resultado mediano. Os salários previstos utilizando esta metodologia de aproximação foram relativamente próximos aos esperados, mas creio que não o suficiente.

Existem alguns casos em que o erro ultrapassa a casa dos 10.000 de diferença, e, ao mesmo tempo, há outras instâncias em que tal erro não chega a ser nem superior a 100.

```
-------------------- Results -------------------------
Salary expected: 24900.0 | Salary provided: 23306.5200145 | Error: 1593.47998546
Salary expected: 31909.0 | Salary provided: 32017.7419789 | Error: 108.741978931
Salary expected: 32850.0 | Salary provided: 27343.8597202 | Error: 5506.14027978
Salary expected: 24750.0 | Salary provided: 30651.7836994 | Error: 5901.78369942
Salary expected: 23712.0 | Salary provided: 21652.2369034 | Error: 2059.76309657
Salary expected: 31114.0 | Salary provided: 31868.7737374 | Error: 754.77373744
Salary expected: 25400.0 | Salary provided: 27095.9384979 | Error: 1695.93849788
Salary expected: 26182.0 | Salary provided: 7034.19883489 | Error: 19147.8011651
Salary expected: 21600.0 | Salary provided: 14607.0965281 | Error: 6992.9034719
Salary expected: 20850.0 | Salary provided: 18420.9611841 | Error: 2429.0388159
Salary expected: 18075.0 | Salary provided: 14817.2875485 | Error: 3257.71245151
------------------------------------------------------
Total error with w = [  1182.66623617   1515.19249881    447.4366801   11088.2939
106
    596.6708994 ] calculated is: 81359661.6245
```

Figura 4: Teste do algoritmo

Observando o resultado mostrado na figura acima, é possível ver que o salário previsto para a segunda instância gerou um erro de 108, enquanto para o salário da oitava pessoa, o erro foi quase 20.000.

Executando o programa várias vezes, o que faz a regressão linear ser executada com dados de entrada diferentes, não gera resultados muito melhores que este. Até então, o erro de 108 foi o menor conseguido.

Deve-se levar em consideração que a escolha das instâncias utilizadas como dados de alimentação da regressão linear, bem como aquelas utilizadas para teste foram escolhidas de maneira aleatória. O primeiro grupo corresponde a 80% do dataset original e o que resta é o segundo.

V. CONCLUSÕES

Este trabalho teve como objetivo colocarmos em prática os algoritmos de machine learning aprendidos em sala, bem como testar sua eficiência e eficácia. Ele foi útil para que pudéssemos ver como funciona o código em si da regressão linear. Este último já havia sido aprendido por nós em outra disciplina obrigatória do curso, porém foi visto de maneira relativamente teórica, logo foi bom poder vê-lo funcionando de fato.

Esperava que obtivesse melhores resultados, mas creio que talvez a disposição dos dados fez com que o resultado da regressão não alcançasse minhas expectativas, portanto considerei-o mediano/bom.

VI. REFERÊNCIAS BIBLIOGRÁFICAS

- https://data.princeton.edu/wws509/datasets/#salary
- https://www.tutorialspoint.com/python/index.htm
- https://docs.python.org/3/tutorial/