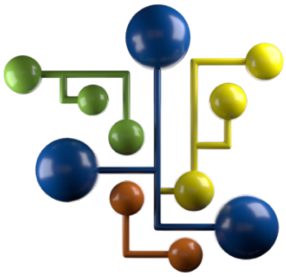

Data Science Academy

www.datascienceacademy.com.br

Big Data Real-Time Analytics com Python e Spark

O Que é Normalização e Quando Aplicar?

A normalização é uma técnica frequentemente aplicada à preparação dos dados em aprendizado de máquina. O objetivo da normalização é alterar os valores das colunas numéricas no conjunto de dados para uma escala comum, sem distorcer as diferenças nos intervalos de valores. Não precisamos aplicar normalização a todo conjunto de dados. É necessário apenas quando os recursos (variáveis) tiverem intervalos diferentes.

Por exemplo, considere o conjunto de dados contendo dois recursos, idade (x1) e receita (x2). Onde a faixa etária varia de 0 a 100 anos, enquanto a renda varia de 0 a 20.000 ou mais. A renda é cerca de 1.000 vezes maior do que a idade e com uma variação de valores muito maior. Então, esses dois recursos estão em intervalos muito diferentes. Quando fazemos análises adicionais, como regressão linear multivariada, por exemplo, a renda atribuída influenciará muito mais o resultado devido ao seu valor maior. E isso causa problemas durante o treinamento do algoritmo.

A normalização também é chamada simplesmente de ***Scaler Min-Max*** e basicamente reduz o intervalo dos dados de forma que o intervalo seja fixo entre 0 e 1 (ou -1 a 1, se houver valores negativos). Funciona melhor para casos em que a padronização (que veremos no próximo item de aprendizagem) pode não funcionar tão bem. Se a distribuição não for gaussiana ou o desvio padrão for muito pequeno, o Scaler Min-Max funciona melhor. Aqui a fórmula que define a normalização:

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Quando a Normalização é Importante?

A normalização é principalmente necessária no caso de algoritmos que usam medidas de distância como clustering, sistemas de recomendação que usam semelhança de cosseno, etc. Isto é feito de forma que uma variável que está em uma escala maior não afeta o resultado apenas porque está em uma escala maior.

Abaixo listamos alguns algoritmos de Machine Learning que requerem a normalização dos dados:

1. KNN com medida de distância euclidiana se quiser que todos os recursos contribuam igualmente no modelo.

2. Regressão Logística, SVM, Perceptrons, Redes Neurais.

3. K-Means

4. Análise discriminante linear, análise de componentes principais, análise de componentes principais do kernel.

Classificadores baseados em modelo gráfico, como Fisher LDA ou Naive Bayes, bem como Árvores de Decisão e métodos baseados em árvore, como Random Forest, são invariantes ao dimensionamento de recursos, mas ainda assim pode ser uma boa ideia redimensionar os dados.

A normalização eliminará a capacidade de interpretação do modelo e, portanto, dependerá, em última instância, da necessidade do negócio.