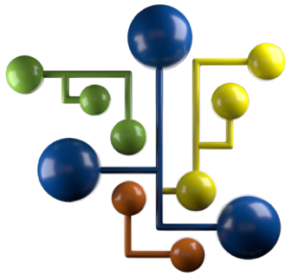


Data Science Academy

www.datascienceacademy.com.br

Big Data Real-Time Analytics com Python e Spark

Normalização x Padronização

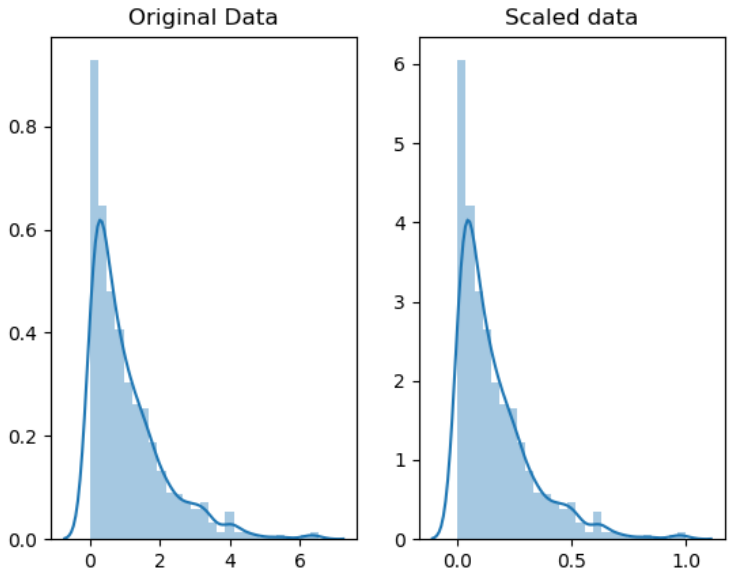
Conforme vimos nas aulas anteriores, a Normalização transforma os dados em um intervalo, digamos entre 0 e 1 ou 1 e 10, de forma que os números estejam na mesma escala. Por exemplo, podemos converter os dados de centímetros para metros para que tenhamos todos na mesma escala. A Normalização pode ser formulada como:

x <- (x - min (x)) / (max (x) - min (x))

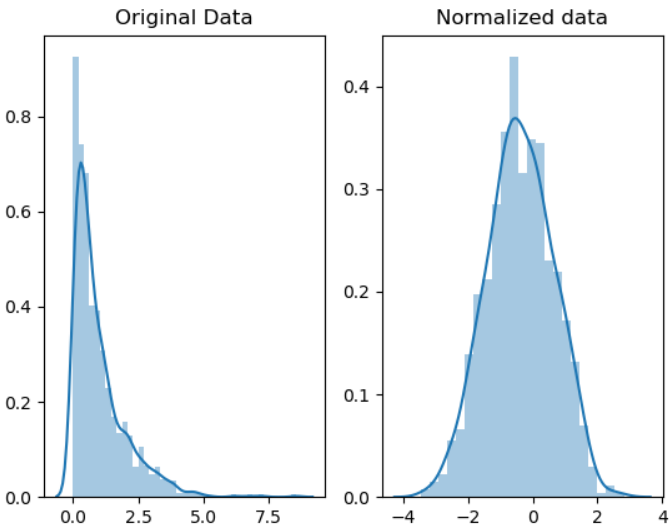
A Padronização significa transformar os dados de tal forma que eles tenham média zero e desvio padrão igual a 1. Portanto, aqui temos os dados em escala de forma padronizada, de modo que a distribuição seja aproximadamente uma distribuição normal, sendo representado da seguinte forma:

x <- (x - mean (x)) / sd (x)

Ambas as técnicas têm suas desvantagens. Se você tiver valores outliers em seu conjunto de dados, a Normalização dos dados certamente aumentará os dados "normais" para um intervalo muito pequeno. E, geralmente, a maioria dos conjuntos de dados tem outliers. Ao usar a Padronização, seus novos dados não são limitados (ao contrário da Normalização). Portanto, a Normalização é geralmente evitada quando o conjunto de dados tem outliers (desde que inclua o valor máximo). Nesses casos, preferimos a Padronização. Os gráficos abaixo resumem as diferenças quando aplicamos Normalização e Padronização:



Observe no eixo x do gráfico acima como a escala dos dados é diferentes, embora a distribuição dos dados seja a mesma. Isso é o que chamamos de Normalização.



Observe no eixo x do gráfico acima como a distribuição dos dados agora segue uma distribuição normal depois que aplicamos a Padronização. Uma distribuição normal é caracterizada por média 0 e desvio padrão 1.

Algumas considerações importantes:

1. A Normalização torna o treinamento menos sensível à escala de recursos, para que possamos resolver melhor os coeficientes.

2. O uso de um método de Normalização melhorará a análise de múltiplos modelos.

3. A Normalização assegurará que um problema de convergência não tenha uma variância massiva, tornando a otimização viável.

4. A Padronização tende a tornar o processo de treinamento bem melhor, porque a condição numérica dos problemas de otimização é melhorada.

Hora de praticar esses conceitos. Até a próxima aula!