



Big Data Real-Time Analytics Com Python e Spark 3.0

Big Data Real-Time Analytics Com Python e Spark Versão 3.0

Normalização x Padronização

As duas abordagens comuns para trazer diferentes recursos para a mesma escala são a normalização e a padronização.

O que é Normalização?

A normalização refere-se ao reescalonamento dos recursos para um intervalo de [0, 1], que é um caso especial de escalonamento mínimo-máximo. Para normalizar os dados, o dimensionamento mínimo-máximo pode ser aplicado a uma ou mais colunas de recursos. Abaixo está a fórmula para normalizar os dados com base no dimensionamento mínimo-máximo. A normalização é útil quando os dados são necessários nos intervalos limitados.

$$x_{norm}^{(i)} = \frac{x^{(i)} - x_{min}}{x_{max} - x_{min}}$$

O método `MinMaxScaler()` do pacote `sklearn.preprocessing` oferece uma forma simples de realizar o procedimento.

O que é Padronização?

A técnica de padronização é usada para centralizar as colunas de recursos na média 0 com um desvio padrão de 1 para que as colunas de recursos tenham os mesmos parâmetros de uma distribuição normal padrão. Ao contrário da Normalização, a padronização mantém informações úteis sobre valores discrepantes e torna o algoritmo menos sensível a eles em contraste com o dimensionamento mínimo-máximo, que dimensiona os dados para um intervalo limitado de valores. Aqui está a fórmula para a padronização.

$$x_{std}^{(i)} = \frac{x^{(i)} - \mu_x}{\sigma_x}$$

O método `StandardScaler()` do pacote `sklearn.preprocessing` oferece uma forma simples de realizar o procedimento.