



Big Data Real-Time Analytics Com Python e Spark 3.0

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

O Que é Feature Scaling?

Muitos algoritmos de aprendizado de máquina funcionam melhor quando as variáveis de entrada são dimensionadas para um intervalo padrão. Esse processo de dimensionamento ou mudança de escala é chamado de Feature Scaling.

Isso inclui algoritmos que usam uma soma ponderada da entrada, como regressão linear, e algoritmos que usam medidas de distância, como k-vizinhos mais próximos (KNN).

As duas técnicas mais populares para dimensionar dados numéricos antes da modelagem são a normalização e a padronização.

A normalização dimensiona cada variável de entrada separadamente para o intervalo 0-1, que é o intervalo para valores de ponto flutuante em que temos mais precisão.

A padronização dimensiona cada variável de entrada separadamente subtraindo a média (chamada de centralização) e dividindo pelo desvio padrão para deslocar a distribuição para ter uma média de zero e um desvio padrão de um.

O dimensionamento de recursos (Feature Scaling) consiste em transformar os valores de diferentes recursos numéricos para que caiam em um intervalo semelhante entre si. O dimensionamento de recursos é usado para evitar que os modelos de aprendizado supervisionados sejam tendenciosos em relação a um intervalo específico de valores.

Por exemplo, se seu modelo é baseado em regressão linear e você não dimensiona recursos, alguns recursos podem ter um impacto maior do que outros, o que afetará o desempenho das previsões, dando vantagem indevida a algumas variáveis sobre outras. Isso coloca certas classes em desvantagem durante o treinamento do modelo. É por isso que se torna importante usar algoritmos de dimensionamento para que você possa padronizar seus valores de recursos.

Esse processo de dimensionamento de recursos é feito para que todos os recursos possam compartilhar a mesma escala e, portanto, evitar problemas como: perda de precisão e aumento no custo computacional à medida que os valores dos dados variam amplamente em diferentes ordens de magnitude.

A ideia é transformar o valor dos recursos em um intervalo semelhante como outros para que os algoritmos de aprendizado de máquina se comportem melhor, resultando em modelos ideais.