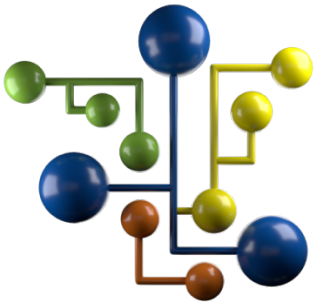

Data Science Academy

www.datascienceacademy.com.br

Machine Learning

Padronização de Dados

Muitos algoritmos de aprendizado de máquina fazem suposições sobre os dados e geralmente é uma boa ideia preparar os dados para melhor expor a estrutura do problema aos algoritmos de aprendizado de máquina que você pretende usar, sendo esta uma parte importante da etapa de pré-processamento de dados, quando trabalhamos com Data Science.

Você quase sempre precisa pré-processar seus dados. É um passo praticamente obrigatório.

Uma dificuldade é que algoritmos diferentes fazem suposições diferentes sobre os dados e podem exigir transformações diferentes. Além disso, quando você segue todas as regras e prepara seus dados, às vezes os algoritmos podem oferecer melhores resultados sem o pré-processamento. Não há fórmulas mágicas, pois tudo depende dos dados e os dados sempre mudam.

Geralmente, recomendamos a criação de muitos modos de exibição e transformações diferentes dos dados e, em seguida, exercitar um punhado de algoritmos em cada bloco do seu conjunto de dados. Isso ajudará você a identificar quais transformações de dados podem ser melhores para expor a estrutura do problema em geral. Vamos listar agora as 4 diferentes técnicas de pré-processamento de dados para aprendizado de máquina, pois esta é uma das maiores dúvidas de quem começa em Data Science.

Vamos considerar um problema de classificação binária em que todos os atributos são numéricos (representando dados qualitativos e quantitativos) e possuem escalas diferentes. É um ótimo exemplo de dados que pode se beneficiar do pré-processamento.

1. Aplicar Escala aos dados

Quando seus dados são compostos de atributos com escalas variáveis, muitos algoritmos de aprendizado de máquina podem se beneficiar do reescalonamento dos atributos para que todos tenham a mesma escala.

Muitas vezes isso é conhecido como normalização (embora não seja o termo ideal) e os atributos são frequentemente redimensionados no intervalo entre 0 e 1. Isso é útil para algoritmos de otimização usados no núcleo de algoritmos de aprendizado de máquina como o gradiente descendente. Também é útil para algoritmos que pesam entradas como regressão e redes neurais e algoritmos que usam medidas de distância como K-Nearest Neighbors (KNN).

2. Padronização dos Dados

A padronização é uma técnica útil para transformar atributos com uma distribuição gaussiana e diferentes médias e desvios padrão para uma distribuição Gaussiana padrão com uma média de 0 e um desvio padrão de 1.

É mais adequado para técnicas que pressupõem uma distribuição gaussiana nas variáveis de entrada e funcionam melhor com dados reescalonados, como regressão linear, regressão logística e análise discriminante linear.

3. Normalização dos Dados

A normalização refere-se ao reescalonamento de cada observação (linha) para ter um comprimento de 1 (chamado de norma unitária em álgebra linear).

Esse pré-processamento pode ser útil para conjuntos de dados esparsos (muitos zeros) com atributos de escalas variadas ao usar algoritmos que ponderam valores de entrada, como redes neurais e algoritmos que usam medidas de distância, como K-Nearest Neighbors (KNN).

4. Binarização dos Dados

Você pode transformar seus dados usando um limite binário. Todos os valores acima do limite são marcados como 1 e todos iguais ou inferiores são marcados como 0.

Isso é chamado de binarizar seus dados ou limitar seus dados. Pode ser útil quando você tem probabilidades que você deseja tornar valores nítidos para leitura e interpretação. Também é útil quando na engenharia de recursos adicionamos novos recursos que indicam algo significativo.