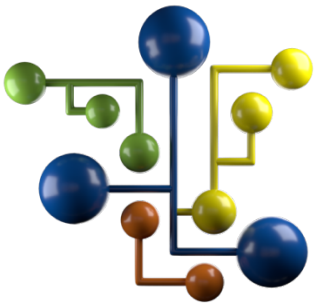

Data Science Academy

www.datascienceacademy.com.br

Engenharia de Dados com Hadoop e Spark

Definindo Machine Learning

A palavra classificação faz parte das nossas vidas e do nosso dia a dia. Desde a época das aulas de português ou biologia ouvimos a palavra classificação o tempo todo. O conceito é simples: baseado em algumas características classificamos determinado objeto em uma categoria ou outra. Vejamos o conceito de Classificação nos algoritmos de Machine Learning.

Primeiro vamos definir Machine Learning. Machine Learning (ML) é uma área da Inteligência Artificial onde criamos algoritmos para ensinar a máquina a desempenhar determinadas tarefas. Um algoritmo de ML basicamente recebe um conjunto de dados de entrada e baseado nos padrões encontrados gera as saídas. Cada entrada desse conjunto de dados possui suas features (ou atributos) e ter um conjunto delas é o ponto inicial para qualquer algoritmo de ML.

Feature é uma característica que descreve um objeto. Qualquer atributo de um objeto pode ser tratado como feature, seja um número, um texto, uma data, um booleano etc. Como no objeto pessoa, temos vários atributos que o descreve, esses atributos são suas features. As features são as entradas dos algoritmos de ML, quanto mais detalhes o algoritmo tiver sobre uma entrada, mais facilmente achará padrões nos dados. Features ruins podem prejudicar o desempenho do algoritmo. Features boas são a chave para o sucesso de um algoritmo. Boa parte do trabalho em ML é conseguir trabalhar os dados e gerar boas features em cima deles, o que é conhecido como **engenharia de features** ou **feature engineering**. Existem diversas técnicas para gerar features, seja através do conhecimento da natureza dos dados ou da aplicação de matemática e estatística para criá-las sobre os dados. Tendo nossas features em mãos podemos aplicar diversos algoritmos de aprendizado nelas. Existem dois grandes grupos de algoritmos em ML, os de aprendizagem supervisionada e os de aprendizagem não supervisionada.

Aprendizagem Supervisionada

Quando você tem um conjunto de entradas que possuem as saídas que deseja prever em outros dados. Com conhecimento das entradas e saídas de um número suficiente de dados, os algoritmos desse grupo podem achar os padrões que relacionam as entradas com as saídas. Dessa forma, se tivermos novos dados apenas com as entradas, podemos prever as saídas com base nesses padrões previamente encontrados. São divididos em dois grupos: classificação e regressão.

Aprendizagem Não-Supervisionada

Quando você tem um conjunto de entradas sem as saídas que você deseja. Com base nas características desses dados podemos gerar um agrupamento ou processá-los a fim de gerar novas formas de expressar essas características. Dois grupos comuns de aprendizagem não supervisionada são: redução de dimensionalidade e clusterização.