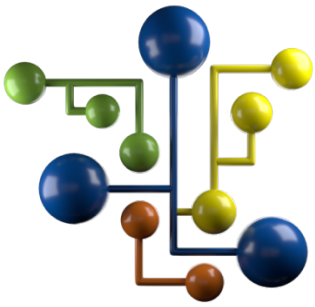

Data Science Academy

www.datascienceacademy.com.br

Engenharia de Dados com Hadoop e Spark

Apache Spark MLLib

O objetivo dos algoritmos de Machine Learning é tentar fazer previsões sobre conjuntos de dados, frequentemente otimizando uma função matemática que melhor descreve o relacionamento entre os dados. Existem diversos tipos de problemas que podem ser resolvidos com Machine Learning, tais como classificação, regressão e clusterização, cada qual com diferentes técnicas. O Spark MLLib é a biblioteca do Spark responsável pelo aprendizado de máquina. Vejamos o que é o MLLib.

O módulo MLLib do Spark contém as funções que implementam Machine Learning e foi criado para ser executado em paralelo, através de um cluster, assim como os demais módulos do Spark. Os algoritmos de Machine Learning do MLLib podem ser executados por todas linguagens de programação suportadas pelo Spark, bem como este módulo é intercambiável com os outros módulos do Framework Spark. O conceito por trás do MLLib é simples. Podemos invocar os algoritmos de Machine Learning e aplicar os modelos nos RDD's. O MLLib introduz mais alguns tipos de dados como vetores e label points, que são funções que aplicamos ao conjunto de dados.

Como já vimos, os dados no Spark são representados por RDD's, que são objetos que armazenam o conjunto de dados e podem ser particionados e distribuídos em paralelo através de um cluster. Aplicamos as funções do MLLib aos RDD's, como por exemplo, funções para feature extraction, que nos permitem converter valores de strings por representações numéricas (mais apropriadas para os algoritmos de Machine Learning). Ao aplicar uma função ao RDD, temos como retorno um vetor de RDD's. Nós então aplicamos os algoritmos de Machine Learning a esses vetores e criamos nosso modelo preditivo. O MLLib possui ainda funções de avaliação do modelo preditivo.