Spark MLlib

Apache Spark offers a Machine Learning API called MLlib. PySpark has this machine learning API in Python.

Spark MLlib is used to perform machine learning in Apache Spark. MLlib consists of popular algorithms and utilities. MLlib in Spark is a scalable Machine learning library that discusses both high-quality algorithm and high speed. The machine learning algorithms like regression, classification, clustering, pattern mining, and collaborative filtering. Lower level machine learning primitives like generic gradient descent optimization algorithm are also present in MLlib. It supports different kind of algorithms

- **mllib.classification** − The **spark.mllib** package supports various methods for binary classification, multiclass classification and regression analysis. Some of the most popular algorithms in classification are **Random Forest, Naive Bayes, Decision Tree**, etc.
- **mllib.clustering** − Clustering is an unsupervised learning problem, whereby you aim to group subsets of entities with one another based on some notion of similarity.
- **mllib.fpm** − Frequent pattern matching is mining frequent items, itemsets, subsequences or other substructures that are usually among the first steps to analyze a large-scale dataset. This has been an active research topic in data mining for years.
- **mllib.linalg** − MLlib utilities for linear algebra.
- **mllib.recommendation** − Collaborative filtering is commonly used for recommender systems. These techniques aim to fill in the missing entries of a user item association matrix.
- **spark.mllib** − It ¬currently supports model-based collaborative filtering, in which users and products are described by a small set of latent factors that can be used to predict missing entries. spark.mllib uses the Alternating Least Squares (ALS) algorithm to learn these latent factors.
- **mllib.regression** − Linear regression belongs to the family of regression algorithms. The goal of regression is to find relationships and dependencies between variables. The interface for working with linear regression models and model summaries is similar to the logistic regression case.

  Machine learning is closely related to computational statistics, which also focuses on prediction-making through the use of computers.

There are three categories of Machine learning tasks:

1. **Supervised Learning**: Supervised learning is where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output.
2. **Unsupervised Learning**: Unsupervised learning is a type of machine learning algorithm used to draw inferences from datasets consisting of input data without labeled responses.
3. **Reinforcement Learning**: A computer program interacts with a dynamic environment in which it must perform a certain goal (such as driving a vehicle or playing a game against an opponent). The program is provided feedback in terms of rewards and punishments as it navigates its problem space. This concept is called reinforcement learning.

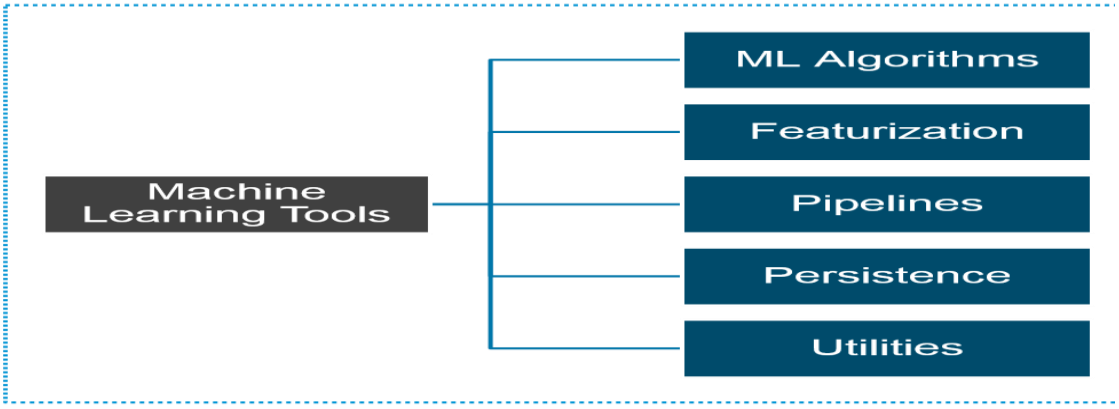
Spark MLlib Overview

Spark MLlib is used to perform machine learning in Apache Spark. MLlib consists popular algorithms and utilities.

MLlib Overview:

- *spark.mllib* contains the original API built on top of RDDs. It is currently in maintenance mode.
- *spark.ml* provides higher level API built on top of DataFrames for constructing ML pipelines. *spark.ml* is the primary Machine Learning API for Spark at the moment.

Spark MLlib Tools



Spark MLlib provides the following tools:

- **ML Algorithms:** ML Algorithms form the core of MLlib. These include common learning algorithms such as classification, regression, clustering and collaborative filtering.

- **Featurization:** Featurization includes feature extraction, transformation, dimensionality reduction and selection.
- **Pipelines:** Pipelines provide tools for constructing, evaluating and tuning ML Pipelines.
- **Persistence:** Persistence helps in saving and loading algorithms, models and Pipelines.
- **Utilities:** Utilities for linear algebra, statistics and data handling.

MLlib Algorithms

The popular algorithms and utilities in Spark MLlib are:

1. Basic Statistics
2. Regression
3. Classification
4. Recommendation System
5. Clustering
6. Dimensionality Reduction
7. Feature Extraction
8. Optimization

Regression

Regression analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables when the focus is on the relationship between a dependent variable and one or more independent variables. More specifically, regression analysis helps one understand how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed.

Classification

Classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known. It is an example of pattern recognition.

Clustering

Clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters).

Predictive Analysis

Predictive analytics is a branch of advanced analytics that makes predictions about future outcomes using historical data combined with statistical modeling, data mining techniques and machine learning.

IBM SPSS Statistics is a powerful solution that can meet any statistical needs of any business industry.

Predictive analytics process

Predictive analysis is used:

- Online Retail
- Healthcare
- Education
- Reduces Risks
- Fraud Detection
- Improvised market campaigning
- weather forecasting
- Social Media Analysis
- cyber security
- Recommendation and search engines
- Government Sector etc.

Steps To Perform Predictive Analysis:

1. **Define Problem Statement:**
   Define the project outcomes, the scope of the effort, objectives, identify the data sets that are going to be used.
2. **Data Collection:**
   Data collection involves gathering the necessary details required for the analysis. It involves the historical or past data from an authorized source over which predictive analysis is to be performed.
3. **Data Cleaning:**
   Data Cleaning is the process in which we refine our data sets. In the process of data cleaning, we remove un-necessary and erroneous data. It involves removing the redundant data and duplicate data from our data sets.
4. **Data Analysis:**
   It involves the exploration of data. We explore the data and analyze it thoroughly in order to identify some patterns or new outcomes from the data set. In this stage, we discover useful information and conclude by identifying some patterns or trends.
5. **Build Predictive Model:**
   In this stage of predictive analysis, we use various algorithms to build predictive

models based on the patterns observed. It requires knowledge of python, R, Statistics and MATLAB and so on. We also test our hypothesis using standard statistic models.

6. **Validation:**

   It is a very important step in predictive analysis. In this step, we check the efficiency of our model by performing various tests. Here we provide sample input sets to check the validity of our model. The model needs to be evaluated for its accuracy in this stage.

7. **Deployment:**

   In deployment we make our model work in a real environment and it helps in everyday discussion making and make it available to use.

8. **Model Monitoring:**

   Regularly monitor your models to check performance and ensure that we have proper results. It is seeing how model predictions are performing against actual data sets.