MACHINE LEARNING LIBRARY

CHEAT SHEET

MLlib Basics

MLlib

It is an Apache Spark machine learning library which is scalable; it consists of popular algorithms and utilities

MLlib contains two packages

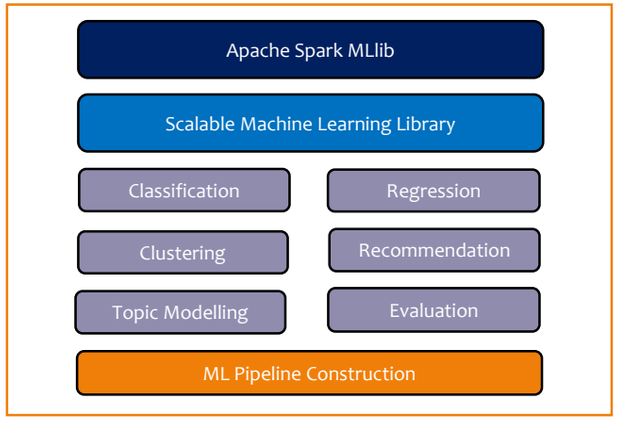
- Spark.mllib
- Spark.ml

To add the **MLlib** the following library is imported:

- In Scala: import org.apache.spark.mllib.linalg.{Vector, Vectors}
- In Java: import org.apache.spark.mllib.linalg.Vector;
  import org.apache.spark.mllib.linalg.Vectors;
- In python: from pyspark.mllib.linalg import SparseVector
  from pyspark.mllib.regression import LabeledPoint

Data Source

Access to HDFS and HBase can be done using **MLlib**, which enables **MLlib** to be plugged in Hadoop Work process



Main Concepts In Pipeline

MLlib is used to standardize the APIs for easy use of multiple algorithms being used as a single pipeline or a workflow

- Data frame: The **ML API** uses Dataframe from **Spark SQL** as a dataset, which can be used to hold a variety of datatypes
- Transformer: This is used to transform one Dataframe to another Dataframe. Examples are
  - Hashing Term Frequency: This calculates how word occurs
  - Logistic Regression Model: The model which results from trying logistic regressions on a dataset
  - Binarizer: This changes a given threshold value to 1 or 0
- Estimator: It is an algorithm which can be used on a Dataframe to produce Transformer. Examples are:
  - Logistic Regression: It is used to determine the weights for the resulting Logistic Regression Model by processing the dataframe
  - StandardScaler: It is used to calculate the Standard deviation
  - Pipeline: Calling fit on a pipeline produces pipeline model, and the pipeline contains only transformers and not the estimators
- Pipeline: A pipeline chains multiple Transformers and Estimators together to specify the ML workflow
- Parameters: To specify the parameters a common API is used by the Transformers and Estimators

Observations

The items or data points used for learning and evaluating

Features

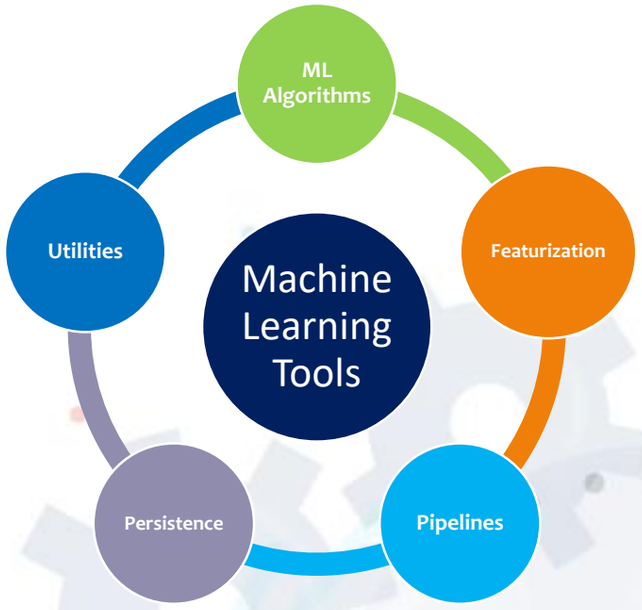
The characteristic or attribute of an observation

Labels

The values assigned to an observation is called a **Label**

Training or test data

A learning algorithm is an observation used for training and testing of the data



Spark MLlib Tools

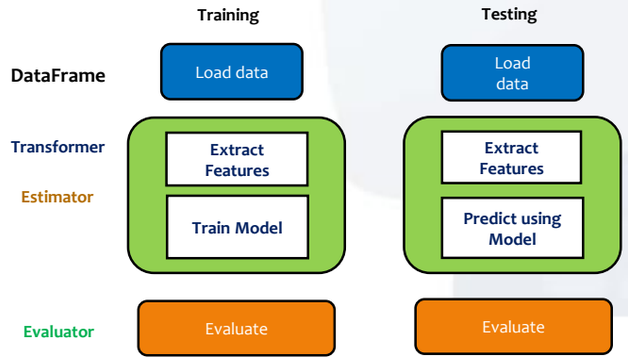
ML Algorithm: These include common learning algorithms such as classification, clustering, regression and collaborative filtering. These algorithms form the core of MLlib

Featurization: It includes feature extraction, transformation, dimensionality reduction and selection

Pipelines: Pipelines provide tools for constructing, evaluating and tuning ML pipelines

Persistence: It helps in saving and loading algorithms, models and pipelines

Utilities: It provides utilities for linear algebra, statistics and data handling



MLlib Algorithms

These include the popular algorithms and utilities

- Basic statistics: It includes the most basic of the machine learning techniques such as:
  - Summary statistics
  - Correlation
  - Stratified sampling
  - Hypothesis testing

Regression: It is a statistical approach to estimate the relationship among variables. It is widely used for prediction and forecasting

Classification: It is used to identify to which set of categories a new observation belongs to.

- K-means classification: It is used for classification using MLlib in Java. It is used to classify every observation, experiment or a vector into one of the cluster

Recommendation system: it is a sub class of information filtering system that seeks to predict the preference or rating a person can give to an item. This can be done in two ways

- Collaborative filtering: It approaches in building a model from a user's past behavior as well as similar decisions made by the user. The model is then used to predict the items in which the user might have interest
- Content-based filtering: It approaches to utilize a series of discrete characteristics of an item in order to recommend more items with similar properties

Clustering: It is a task to group set of objects in a way that the objects in the same group is more similar to each other when compared to the objects in the other group.

Dimensionality Reduction: It is a process of reducing a set of random variables under consideration by obtaining a set of principal variables. It can be divided into two types

- Feature selection: It finds a subset of original variables called attributes
- Feature Extraction: This will transform the data from in a high dimensional space to a space of fewer dimensions.

Feature extraction: It starts from initial set of derived data and builds derived values.

Optimization: It is a selection of best element from the set of available alternatives