/CursoAmadeus20...

Machine Learning

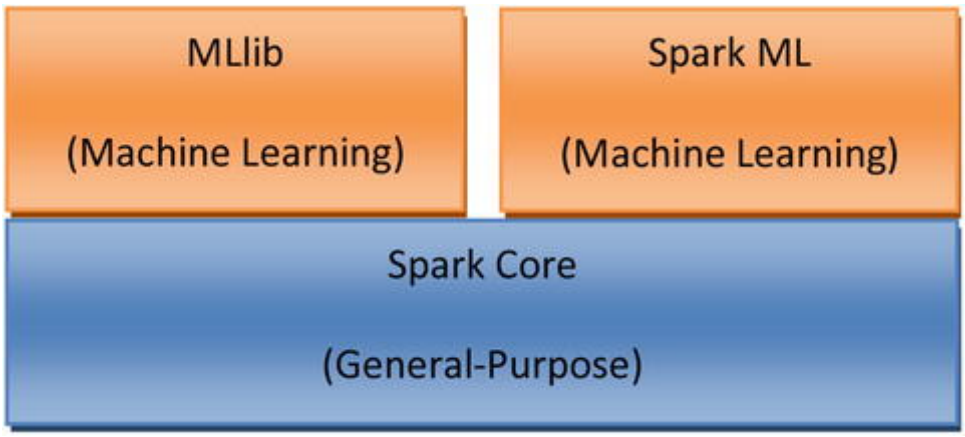
FINISHED

Took 1 sec. Last updated by anonymous at March 02 2017, 6:12:39 AM.

Introduction

FINISHED

General Concepts



MLlib Overview

- MLlib extends Spark for machine learning and statistical analysis.
- It provides a higher-level API than the Spark core API for machine learning and statistical analysis.
- It comes prepackaged with commonly used machine learning algorithms used for a variety of machine learning tasks.
- It also includes statistical utilities for different statistical analysis.
- MLlib integrates with other Spark libraries such as Spark Streaming and Spark SQL

What includes Spark MLib

Regression and Classification

```
Linear regression
Logistic regression
Support Vector Machine
Naïve Bayes
Decision tree
Random forest
Gradient-boosted trees
Isotonic regression
```

Clustering

```
K-means
Streaming k-means
Gaussian mixture
Power iteration clustering (PIC)
Latent Dirichlet allocation (LDA)
```

Dimensionality Reduction

```
Principal component analysis (PCA)
Singular value decomposition (SVD)
```

Feature Extraction and Transformation

```
TF-IDF
Word2Vec
Standard Scaler
Normalizer
Chi-Squared feature selection
Elementwise product
```

Frequent pattern mining

```
FP-growth
Association rules
PrefixSpan
```

Recommendation

```
Collaborative filtering with Alternating Least Squares (ALS)
```

Took 0 sec. Last updated by anonymous at March 02 2017, 6:24:01 AM. (outdated)

https://goo.gl/dH6Fde (https://goo.gl/dH6Fde)

FINISHED

Took 1 sec. Last updated by anonymous at March 02 2017, 7:10:40 AM.

Basic Types

FINISHED

Vector

The Vector type represents an indexed collection of Double-type values with zero-based index of type Int. It is generally used for representing the features of an observation in a dataset. Conceptually, a Vector of length n represents an observation with n features. In other words, it represents an element in an n-dimensional space.

In order to use the Vector you need to import:

```
import org.apache.spark.mllib.linalg.Vector
```

DenseVector

An instance of the DenseVector class stores a double-type value at each index position. It is backed by an array. A dense vector is generally used if a dataset does not have too many zero values. It can be created, as shown here.

```
import org.apache.spark.mllib.linalg._
val denseVector = Vectors.dense(1.0, 0.0, 3.0)
```

SparseVector

The SparseVector class represents a sparse vector, which stores only non-zero values. It is an efficient data type for storing a large dataset with many zero values. An instance of the SparseVector class is backed by two arrays; one stores the indices for non-zero values and the other stores the non-zero values.

A sparse vector can be created, as shown here.

```
import org.apache.spark.mllib.linalg._
val sparseVector = Vectors.sparse(10, Array(3, 6), Array(100.0, 200.0))
```

LabeledPoint

The LabeledPoint type represents an observation in a labeled dataset. It contains both the label (dependent variable) and features (independent variables) of an observation. The label is stored as a Double-type value and the features are stored as a Vector type.

```
import org.apache.spark.mllib.linalg.Vectors
import org.apache.spark.mllib.regression.LabeledPoint

val positive = LabeledPoint(1.0, Vectors.dense(10.0, 30.0, 20.0))
val negative = LabeledPoint(0.0, Vectors.sparse(3, Array(0, 2), Array(200.0, 300.0)))
```

Rating

The Rating type is used with recommendation algorithms. It represents a user's rating for a product or item. A training dataset must be transformed to an RDD of Ratings before it can be used to train a recommendation model.

```
import org.apache.spark.mllib.recommendation._

val rating = Rating(100, 10, 3.0)
```

Took 1 sec. Last updated by anonymous at March 02 2017, 7:27:23 AM.

Example of predicting loan credit rick using random forest & Spark MLib

FINISHED

https://goo.gl/6ddNim (https://goo.gl/6ddNim)

Took 0 sec. Last updated by anonymous at March 02 2017, 6:31:25 AM.

Spark GraphX & Graphframes

FINISHED

https://goo.gl/NV7utl (https://goo.gl/NV7utl)

Took 0 sec. Last updated by anonymous at March 02 2017, 9:31:16 AM.

Examples of Graphframes

FINISHED

scala

https://goo.gl/qDp2xj (https://goo.gl/qDp2xj)

Python

https://goo.gl/OTUN24 (https://goo.gl/OTUN24)

https://goo.gl/ZcYb8O (https://goo.gl/ZcYb8O)

Took 0 sec. Last updated by anonymous at March 02 2017, 7:27:10 AM. (outdated)

```
%md
```

READY