Week2_SparkML_Fundamentals

SparkML Fundamentals

What is Machine Learning

- Applies algorithms that automtically learn features from data
- Are not explicitly programmed to do a task
- Learn from data/experience and improve with more data
- Apply statistical tools that enables AI
- Uses models trained with data and performs predictions on new data

MLLib

- Applies practical machine learning algorithms and is scalable
- Performs machine learning operations using DataFrame-based APIs
- Enables multiple capabilities for implementing a machine learning systems:
  - ML algorithms
  - Featurization
  - Functions
    - Pipelines
    - Persistence
    - Utilities
- Supports satandard Spark supported data sources including:
  - Parquet
  - CSW
  - JSON
  - JDBC
- Has special libraries to support images and LIBSVM data types
- Supports both feature vector and label column data
- Images are a common data source and create a DataFrame and an image schema
- LIBSVM loads the `libsvm` data files and creates a DataFrame with two columns including the feature vector and label

Spark MLLib Inbuild Utilities

- Linear Algebra (`spark.ml.linalg`): used for basic linear algebra data holders such as matrices and vectors
- Statistics (`spark.ml.stat`): used for statistics operations such as correlation, hypothesis, testing, summarizing, etc.
- Feature (`spark.ml.feature`): powerful toolbox to convert raw data into useful features for ML model fitting

ML Pipelines

- Exposes a single Spark ML pipeline API
- Combines multiple algorithms into a single workflow or pipeline using Transformers and Estimators as building blocks
- Transformers can transform one DataFrame into another DataFrame using the `transform()` function
  - For example: A model that converts feature data into predictions
- Estimators encapsulate the algorithm and model learning on the features using `fit()` function

Classification and Regression Using Apache Spark

Supervised Learning

- Subset of machine learning algorithms
- Learn from data and labels (supervision)
- Usually require a lot of effort in data labelling and improve in performance with more labelled data

Classification

- Produces a prediction from a discrete set of possible outbomes the task is called classification
- The model predicts each object's target category or class
- Examples: Predicting a sports tournament winner, heads or tails on a coin toss, classifying images with a pre-set number of distinct categories

Regression

- Regression is a form of an implicit function approximation where the model predicts real valued outputs for a given input
- The predicted value is usually a continuous real number, such as a float or integer
- Examples: weather predictions, stock market price predictions, house value estimation, and others

Supervised ML on Spark

Classification : `spark.ml.classification`

Provides convenient functions to perform classification tasks and algorithms such as logistic regression classifier, decision trees, random forests, multilayer perceptron, support vector machines (svm), naive bayes, one-versus-all classifier and others

Regression: `spark.ml.regression`

Provides convenient functions to perform many regression task or algorithms such as linear regression, generalized linear regression, decision tree and random forest regression, survival regression and isotonic regression algorithms and others

SparkML Clustering

Unsupervised Learning

- Does not require explicit labels mapped to features
- Automatically learns patterns and latent spaces in the data
  - Ex: clustering, recommender system

Clustering

- Is a subset of UL
- Groups data into clusters
- All elements within a cluster share similar characteristics

Clustering on Spark
`spark.ml.clustering`

SparkML provides the following functions that perform common clustering algorithms: k-means, Latent Dirichlet Allocation, Gaussian Mixture Models, and others

Summary

Machine learning is a branch of computer science and statistics in which algorithms are not explicitly programmed to do a task but learn from data and experience. Spark MLlib is prepackaged with standard machine learning algorithms for clustering, classification, regression, and other tasks.

Spark MLlib offers featurization with functions that easily extract features of interest from raw data, including functions for feature extraction, dimensionality reduction, and string tokenization, and others.

Spark MLlib offers functions for pipelines and persistence and utilities for day-to-day machine learning tasks with handy functions for statistics, linear algebra, image processing, and others.

Supervised learning, a subset of machine learning, includes classification and regression. Classification produces a prediction from a discrete set of possible outcomes. Regression is an implicit function approximation where the model predicts real-valued outputs for a given input. Spark MLlib supports both classification and regression with an abundance of parallel, scalable algorithms

A popular subset of Machine Learning is unsupervised learning. The models automatically learn patterns and groups during unsupervised learning, also known as latent spaces, within the data. Unsupervised learning is generally more advanced and more complex to perform than supervised learning. Examples of unsupervised learning include clustering algorithms and recommender systems.

In clustering, a machine learning model tries to divide the dataset into clusters or groups. All elements within a cluster share similar characteristics. The k-means algorithm helps you train the data model by splitting the dataset into a specified number of clusters. You can then use the fit function—with k-means as an estimator—to train the model.