Introduction to Machine Learning with Apache Spark

Module 2: Machine Learning with Apache Spark

Welcome! This alphabetized glossary contains many terms you'll find in this course. This comprehensive glossary also includes additional industry-recognized terms not used in course videos. These terms are essential for you to recognize when working in the industry, participating in user groups, and participating in other certificate programs.

Terms	Definition	Video
Bisecting K-means	A hierarchical clustering algorithm that recursively splits clusters into smaller subclusters until the desired number of clusters is reached.	Clustering using SparkML
Cluster monitoring	Monitoring the performance, resource usage, and overall health of the Spark cluster using built-in and third-party tools.	Spark for Data Engineers
CSV file	Comma-separated values file, a common file format for storing tabular data where a comma separates each value.	Regression using SparkML
Data ingestion	The process of importing large volumes of data from various sources into Spark for processing.	Spark for Data Engineers
DataFrame	A distributed collection of data organized into named columns, commonly used in Spark for data manipulation and analysis.	Regression using SparkML
DataFrameReader	A class in Spark that provides methods for reading data from various file formats into a DataFrame.	Classification using SparkML
Decision tree regression	A regression algorithm that uses a decision tree to model the relationship between the target variable and input features.	Regression using SparkML
Domain-Specific Language (DSL)	A programming language or syntax specifically designed to express concepts and operations within a particular domain or problem space. GraphFrames utilizes a DSL to specify search queries for motif finding.	GraphFrames on Apache Spark
Gaussian Mixture Models (GMM)	A probabilistic model for representing data distributions often used for clustering tasks. GMM assumes that the data points are generated from a mixture of Gaussian distributions.	Clustering using SparkML
Gradient-boosted tree regression	A regression algorithm that builds an ensemble of weak decision trees in a sequential manner to make accurate predictions.	Regression using SparkML

GraphFrame	A graph processing library built on top of Apache Spark that provides high-level APIs for working with graph data. It extends the capabilities of Spark's DataFrame and Dataset APIs to enable graph computation and analysis.	GraphFrames on Apache Spark
Hadoop MapReduce	A previous framework for distributed processing of large data sets that Spark overcomes.	Spark for Data Engineers
In-memory processing	Spark's capability to cache data in memory eliminating the high input/output costs associated with disk-based processing.	Spark for Data Engineers
Jupyter Notebook	An interactive web-based environment for writing and running code, visualizing data, and documenting workflows. An open-source web application that allows you to create and share documents containing live code, equations, visualizations, and narrative text.	Classification using SparkML
K-means	A popular clustering algorithm aims to partition the data into K clusters, where each data point belongs to the cluster with the nearest mean.	Clustering using SparkML
Linear regression	A regression algorithm that models the relationship between the target variable and input features as a linear equation.	Regression using SparkML
Logistic regression	A classification algorithm that models the relationship between input features and categorical outcomes using the logistic function.	Classification using SparkML
MulticlassClassificationEvaluator	The MulticlassClassificationEvaluator is a class in Apache Spark's Machine Learning (ML) library that provides evaluation metrics for multiclass classification models. It helps measure the performance of multiclass classification models by comparing their predicted class labels against the true class labels.	Classification using SparkML
Random forest regression	A regression algorithm that combines multiple decision trees to improve prediction accuracy.	Regression using SparkML
RegressionEvaluator	The RegressionEvaluator is a class in Apache Spark's Machine Learning (ML) library that provides evaluation metrics for regression models. It helps measure the performance of regression models by comparing their	Regression using SparkML

	predicted continuous values against the true labels or target values.	
Root Mean Squared Error (RMSE)	RMSE measures the average deviation between the predicted and actual values. It calculates the square root of the average of the squared differences between the predicted and true values. Lower RMSE values indicate better model performance.	Regression using SparkML
Scaling	Scaling is a common preprocessing technique in machine learning that helps to normalize or standardize the features of a dataset. It ensures that all the features have a similar scale or range, which can be beneficial for certain machine learning algorithms, such as those based on distance calculations or gradient descent optimization.	Spark for Data Engineers
Sentiment analysis	A classification task that involves determining the sentiment or emotion expressed in a piece of text, such as positive, negative, or neutral.	Classification using SparkML
Silhouette score	A metric that measures the quality of clustering by computing the average distance between data points within clusters and the average distance to the nearest neighboring cluster. Higher silhouette scores indicate well-separated and distinct clusters.	Clustering using SparkML
Spark	A distributed computing system that processes large-scale data sets and provides a unified computing engine for various data processing tasks.	Spark for Data Engineers
Spark Cluster	A Spark cluster is a group of computers or servers that work together to process data using Apache Spark, an open-source distributed computing system. Spark is designed to handle large-scale data processing tasks and provides a framework for distributing data across multiple machines and parallelizing computation.	Spark for Data Engineers
Spark ML	The machine learning library provided by Apache Spark for building and training machine learning models.	Regression using SparkML
Spark MLlib	An older name for Spark ML referring to the same machine learning library within the Spark ecosystem.	Regression using SparkML
Spark Session	A Spark Session is the entry point for programming with Apache Spark. It is a unified interface that allows you to interact with Spark and perform various data	Regression using SparkML

	processing and analysis tasks. Spark Session provides a convenient way to create and manage Spark contexts, which are required to connect to a Spark cluster and execute Spark operations.	
SparkSession	The entry point for Spark functionality serving as a unified interface to interact with various Spark features and libraries.	Regression using SparkML
Stream processing	Processing real-time data streams using Spark's capabilities.	Spark for Data Engineers
Supervised learning	A type of machine learning where the algorithm learns from labeled training data to make predictions or classify new, unseen data.	Classification using SparkML
VectorAssembler	The VectorAssembler is a feature transformation class in Apache Spark's Machine Learning Library that combines multiple input columns into a single vector column. It is often used as a preprocessing step to prepare data for machine learning algorithms that expect input features in vector format.	Regression using SparkML
Vertex	Also known as a node, it represents an object or entity in a graph. In the context of social networks, it can represent a person, while in other domains, it can represent different entities such as web pages, products, or locations.	GraphFrames on Apache Spark