

**Machine Learning With Spark**

**Course Number:** SPRK-106  
**Duration:** 3 days

**Overview**

This Machine Learning With Spark training course teaches attendees how to leverage machine learning at scale with the popular Apache Spark framework. This class dives into foundations, applicability, and limitations, as well as implementation, use, and specific use cases. Students don't just learn the APIs, they learn the theory behind it and work with real-world sample datasets from leading companies.

**Prerequisites**

This course is intended for data scientists and software engineers, however, we assume no previous knowledge of Machine Learning. Students should have a programming background, and familiarity with Python would be a plus but is not required. If students are new to Apache Spark, we can offer a 1-day Introduction to Spark training primer.

**Materials**

All Spark training students receive comprehensive courseware.

**Software Needed on Each Student PC**

* Windows, Mac, or Linux PCs with the current Chrome or Firefox browser.
  + Most class activities will create Spark code and visualizations in a browser-based notebook environment. The class also details how to export these notebooks and how to run code outside of this environment.
* Internet access

**Objectives**

* Learn popular machine learning algorithms, their applicability, and their limitations
* Practice the application of these methods in the Spark machine learning environment
* Learn practical use cases and limitations of algorithms
* Apply ML Concepts
* Use Regressions, Classifications, and Clustering
* Perform Principal Component Analysis (PCA)

**Outline**

* Introduction
* Machine Learning (ML) Overview
  + Machine Learning landscape
  + Machine Learning applications
  + Understanding ML algorithms & models
* ML in Python and Spark
  + Spark ML Overview
  + Introduction to Jupyter notebooks
* Machine Learning Concepts
  + Statistics Primer
  + Covariance, Correlation, Covariance Matrix
  + Errors, Residuals
  + Overfitting / Underfitting
  + Cross-validation, bootstrapping
  + Confusion Matrix
  + ROC curve, Area Under Curve (AUC)
* Feature Engineering (FE)
  + Preparing data for ML
  + Extracting features, enhancing data
  + Data cleanup
  + Visualizing Data
* Linear regression
  + Simple Linear Regression
  + Multiple Linear Regression
  + Running LR
  + Evaluating LR model performance
  + Use case: House price estimates
* Logistic Regression
  + Understanding Logistic Regression
  + Calculating Logistic Regression
  + Evaluating model performance
  + Use case: credit card application, college admissions
* Classification: SVM (Supervised Vector Machines)
  + SVM concepts and theory
  + SVM with kernel
  + Use case: Customer churn data
* Classification: Decision Trees & Random Forests
  + Theory behind trees
  + Classification and Regression Trees (CART)
  + Random Forest concepts
  + Use case: predicting loan defaults, estimating election contributions
* Classification: Naive Bayes
  + Theory
  + Use case: spam filtering
* Clustering (K-Means)
  + Theory behind K-Means
  + Running K-Means algorithm
  + Estimating the performance
  + Use case: grouping cars data, grouping shopping data
* Principal Component Analysis (PCA)
  + Understanding PCA concepts
  + PCA applications
  + Running a PCA algorithm
  + Evaluating results
  + Use case: analyzing retail shopping data
* Recommendations (Collaborative filtering)
  + Recommender systems overview
  + Collaborative Filtering concepts
  + Use case: movie recommendations, music recommendations
* Performance
  + Best practices for scaling and optimizing Apache Spark
  + Memory caching
  + Testing and validation
* Conclusion