Where Technology Means More®

Machine Learning With Apache Spark

Course Time:
3 Days

Course Description

This course teaches doing Machine Learning at Scale with the popular Apache Spark framework. We assume no previous knowledge of Machine Learning—we teach popular Machine Learning algorithms from scratch. For each concept, we discuss foundations, applicability, and limitations. Then we explain the implementation and use as well as specific use cases. This is achieved through a combination of about 50% lecture, 50% lab work. This course is taught using Spark & Python.

Learning Objectives

- Learn popular Machine Learning algorithms, their applicability, and limitations
- Practice the application of these methods in the Spark machine learning environment
- Learn practical use cases and limitations of algorithms

You will learn:

- ML Concepts
- Regressions
- Classifications
- Clustering
- Principal Component Analysis (PCA)
- Recommendations

Who Should Attend

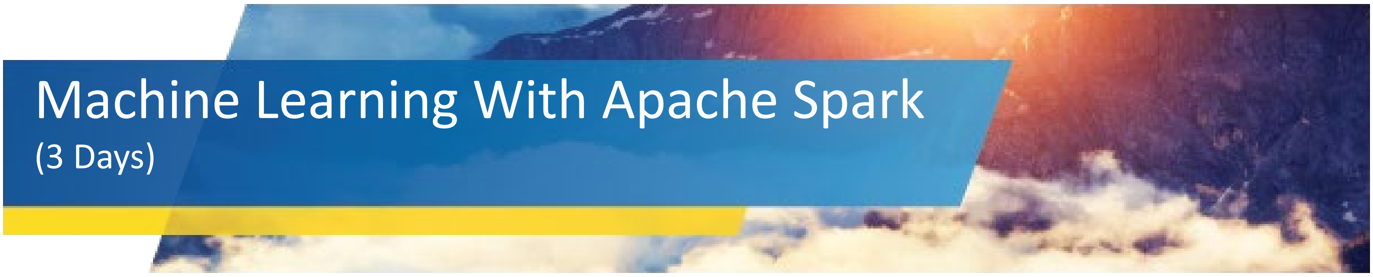
- Data Scientists
- Software Engineers

Prerequisites

- Programming background
- Familiarity with Python would be a plus, but not required
- No machine learning knowledge is assumed

Lab Environment

- Working Spark environment will be provided for students. Students only need an SSH client and a browser.
- Zero Install: There is no need to install software on students' machines.



Machine Learning With Apache Spark

(3 Days)

Course Content

Section 1: Machine Learning (ML) Overview

- Machine Learning landscape
- Machine Learning applications
- Understanding ML algorithms and models

Section 2: ML in Python and Spark

- Spark ML Overview
- Introduction to Jupyter notebooks
- Lab: Working with Jupyter + Python + Spark
- Lab: Spark ML utilities

Section 3: Machine Learning Concepts

- Statistics Primer
- Covariance, Correlation, Covariance Matrix
- Errors, Residuals
- Overfitting / Underfitting
- Cross-validation, bootstrapping
- Confusion Matrix
- ROC curve, Area Under Curve (AUC)
- Lab: Basic stats

Section 4: Feature Engineering (FE)

- Preparing data for ML
- Extracting features, enhancing data
- Data cleanup
- Visualizing Data
- Lab: data cleanup
- Lab: visualizing data

Section 5: Linear regression

- Simple Linear Regression
- Multiple Linear Regression
- Running LR
- Evaluating LR model performance
- Lab
- Use case: House price estimates

Section 6: Logistic Regression

- Understanding Logistic Regression
- Calculating Logistic Regression
- Evaluating model performance
- Lab
- Use case: credit card application, college admissions

Section 7: Classification: Supervised Vector Machines

- SVM concepts and theory
- SVM with kernel
- Lab
- Use case: Customer churn data

Section 8: Classification: Decision Trees & Random Forests

- Theory behind trees
- Classification and Regression Trees (CART)
- Random Forest concepts
- Labs
- Use case: predicting loan defaults, estimating election contributions

Section 9: Classification: Naive Bayes

- Theory
- Lab
- Use case: spam filtering

Section 10: Clustering (K-Means)

- Theory behind K-Means
- Running K-Means algorithm
- Estimating the performance
- Lab
- Use case: grouping cars data, grouping shopping data

Section 11: Principal Component Analysis (PCA)

- Understanding PCA concepts
- PCA applications
- Running a PCA algorithm
- Evaluating results
- Lab
- Use case: analyzing retail shopping data

Section 12: Recommendations (Collaborative filtering)

- Recommender systems overview
- Collaborative Filtering concepts
- Lab
- Use case: movie recommendations, music recommendations

Section 13: Performance

- Best practices for scaling and optimizing Apache Spark
- Memory caching
- Testing and validation

Section 14: Final workshop (time permitting)

- Students will analyze datasets and run ML algorithms as a group exercise with presentations of findings