Course Outline

Introduction to Machine Learning with Apache Spark

Course Summary

Description

This course teaches Machine Learning from a practical perspective. In-depth coverage of Math / Stats is beyond the scope of this course.

Machine Learning (ML) is changing the world. To use ML effectively, one needs to understand the algorithms and how to utilize them. This course provides an introduction into the most popular machine learning algorithms.

We will also use Apache Spark as our ML platform. Apache Spark provides scalable ML platform, that makes it possible to analyze large amount of data.

Objectives

By the end of this course, students will learn:

- Spark ecosystem
- Spark ML Library
- ML Concepts
- Regressions
  - Linear Regression
  - Logistic Regressions
- Classifications
  - Naive Bayes
  - SVM
  - Decision Trees
  - Random Forest
- Clustering algorithms (K-Means)
- Principal Component Analysis (PCA)
- Recommendations

Topics

- Spark
- Machine Learning (ML) Overview
- ML in Python and Spark
- Feature Engineering and Exploratory Data Analysis (EDA)
- Machine Learning Concepts
- Linear regression
- Logistic Regression
- Classification: SVM (Supervised Vector Machines)
- Classification: Decision Trees & Random Forests
- Classification: Naive Bayes
- Unsupervised Algorithms
- Unsupervised: Clustering: K-Means
- Unsupervised: Principal Component Analysis (PCA)
- Recommendations
- Final workshop (time permitting)

Audience

This course is designed for Data analysts, Software Engineers, and Data scientists.

Course Outline

Introduction to Machine Learning with Apache Spark

Course Summary (cont.)

Prerequisite

- Good programming background
- familiarity with Python would be a plus, but not required
- No machine learning knowledge is assumed
- No Spark knowledge is assumed

Duration

Four Days

Course Outline

Introduction to Machine Learning With Apache Spark

Course Outline

I. Spark
A. Spark ecosystem
B. Spark data models
C. Spark ML

II. Machine Learning (ML) Overview
A. Machine Learning landscape
B. Understanding Deep Learning use cases
C. Understanding AI / Machine Learning / Deep Learning
D. Data and AI
E. AI vocabulary
F. Hardware and software ecosystem
G. Understanding types of Machine Learning (Supervised / Unsupervised / Reinforcement)

III. ML in Python and Spark
A. Spark ML Overview
B. Introduction to Jupyter notebooks
- Lab: Working with Jupyter + Python + Spark
- Lab: Spark ML utilities

IV. Feature Engineering and Exploratory Data Analysis (EDA)
A. Preparing data for ML
B. Statistics Primer
C. Data cleanup
D. Extracting features, enhancing data
E. Visualizing Data
- Labs:
  - Data cleanup
  - Exploring data
  - Visualizing data

V. Machine Learning Concepts
A. Training and Testing
B. Gradient Descent
C. Overfitting / Under-fitting
D. Cross validation, bootstrapping
E. Confusion Matrix
F. ROC curve, Area Under Curve (AUC)

VI. Linear regression
A. Linear Regression
B. Errors, Residuals
C. Multiple Linear Regression
D. Evaluating model performance
- Labs:
  - Use case: House price estimates

VII. Logistic Regression
A. Understanding Logistic Regression
B. Calculating Logistic Regression
C. Evaluating model performance
- Labs:
  - Credit card application
  - college admissions

VIII. Classification: SVM (Supervised Vector Machines)
A. SVM concepts and theory
B. SVM with kernel
- Labs: -Customer churn data

IX. Classification: Decision Trees & Random Forests
A. Classification and Regression Trees (CART) introduction
B. Decision Tree concepts
C. Pruning trees
D. Gini index
E. Bias Variance Tradeoff
F. Random Forest concepts
G. Random Forests features and examples
- Labs:
  - Predicting loan defaults
  - Estimating election contributions

X. Classification: Naive Bayes
A. Naive Bayes theory
B. Running Naive Bayes algorithm
C. Evaluating model performance
- Lab
  - Spam filtering

XI. Unsupervised Algorithms
A. Overview of unsupervised algorithms
B. Supervised vs. unsupervised
C. Understanding unsupervised algorithms

Introduction to Machine Learning With Apache Spark

Course Outline (cont.)

XII. Unsupervised: Clustering: K-Means

A. Theory behind K-Means
B. Running K-Means algorithm
C. Estimating the performance
   - Labs:
     - Predicting Uber demand
     - Clustering shopping trips

XIII. Unsupervised: Principal Component Analysis (PCA)

A. Understanding dimensions
B. 'Curse of dimensionality'
C. Reducing dimensions
D. Overview of Principal Component Analysis (PCA)
E. Eigen vectors and values
F. Implementing PCA algorithm
   - Labs:
     - Predicting wine quality
     - Predicting income from census data

XIV. Recommendations

A. Recommendation use cases
B. Recommender systems
C. Collaborative Filtering (CF)
D. Implementing CF algorithm
   - Lab:
     - Movie ratings recommendation
     - Songs rating recommendation

XV. Final workshop (time permitting)

A. This is a group workshop
B. Each group will analyze a couple of real-world datasets and run ML algorithms
C. Each group will present their findings to the class