Machine Learning Using R, Python and Apache Spark

Training Course Outline

What is Machine Learning?

Arthur Samuel, a pioneer in Artificial Intelligence (AI) defined machine learning as "field of study that gives computers the ability to learn without being explicitly programmed". The field of machine learning is concerned with the question of how to construct computer programs that automatically improve with experience. In the recent years, many machine learning applications have been developed ranging from data mining that learn to detect fraudulent credit card transactions to information filtering systems that learn reading preferences, to autonomous vehicles that learn to drive on public roads, optical character recognition systems that identify and read the text and objects from images, medical diagnostics and many more. Many researchers also think it is the best way to make progress towards human-level AI.

Overview of the Training

This course provides a broad introduction to machine learning, data mining, and statistical pattern recognition. Topics include: (i) Supervised learning (parametric/non-parametric algorithms, support vector machines, kernels, neural networks). (ii) Unsupervised learning (clustering, dimensionality reduction, recommender systems, deep learning). (iii) Best practices in machine learning (bias/variance theory; innovation process in machine learning and AI). You'll learn about not only the theoretical underpinnings of learning, but also gain the practical know-how needed to quickly and powerfully apply these techniques to new problems. You will spend 40% time on theory lectures and case-studies and 60% on hands on lab exercises.

Programming Language:

This course will be taught using **R and Python** programming languages for exploring machine learning models. A few demonstrations will be shown in Scala as well.

Duration of the Training Course: 5 Days

Mode: Instructor led classroom based training, hands on oriented.

Prerequisite

- Basic knowledge of statistics, mathematics and computer programming are required to take full advantage of this course.
- Basic programming knowledge in Python, R or Scala will be helpful.

Day 1 – Introduction to Spark, Python and R for machine learning

Topic	Description
Introduction to Machine Learning	• Machine Learning vs Statistics vs Data Science vs Data Engineering vs Data Analysis • Real world scenarios to understand the perspective • History, Pioneers and Modern Trends • Types of Machine Learning ◦ Supervised Learning ◦ Unsupervised Learning ◦ Reinforced learning
Introduction to R and Python for descriptive statistics	• Inspecting Data using R and Python ◦ Examine data types – quantitative, qualitative, continuous, discrete, ordinal, nominal etc. ◦ Probability distribution • Common Data Pre-processing Tasks ◦ Feature Transformation ◦ Encode Categorical features ◦ Handle Missing Data ◦ Handle Outliers ◦ Remove Problematic Features
Lab Exercises	• Getting started with R, R-studio and R packages • Getting started with Python Anaconda distribution for data analysis – Pandas, Scikit, Numpy, Scipy, Matplotlib etc. • Explore built in datasets in R and Python
Introduction to Spark, SparkR and Spark MLlib	• Apache Spark – a Swiss army knife for data processing at scale • Apache Spark architectural overview • Data Abstraction for massively parallel processing – RDD, DF, DS • Building and submitting Spark application • Highly available, high throughput data sources • Efficient data formats for large scale data processing
Lab Exercise	• Working with stocks market data for S&P 500 companies

Learning Goals

- Spot the use cases of machine learning
- Explore and transform data using R, Python
- Understands Spark framework for large scale, distributed data processing for machine learning

Day 2 – Supervised Algorithms (Regression)

Topic	Description
Linear Regression	• Assumptions of Linear Regression • Simple Linear Regression • Multiple Linear Regression • Cross validation • Tuning and hyper parameter selection • Regularization – Ridge, Lasso, Elastic Net for robustness • Running gradient boosting techniques to solve regression problems
Lab Exercises	• Build a regression model using on startup dataset. Measure performance, tune model. • Apply Linear Regression to predict the demand of Electricity Power Demand • [Advanced] Build regression model on housing price prediction problem from kaggle.com

Learning Goals

- Learn to predict numeric data (for example, revenue, price etc.)
- Learn to evaluate a regression model
- Learn to validate and tune a regression model for best performance

Day 3 – Supervised Algorithms (Classifications)

Topic	Description
Classification	• What are classification problems and types of classification algorithms • Decision Tree Classification ◦ Intuition of tree models ◦ CART regression trees ◦ Tree pruning ◦ Missing Data • Classification using Random Forrest • Logistic Regression ◦ Generalized linear models ◦ Interpreting coefficients in logistic regression ◦ Performance metrics – confusion matrix ◦ AUC value, ROC plot • Support Vector Classifiers • Compare algorithms
Lab Exercise	• Evaluate of credit status using customer profile • Segmentation of insurance customers • Wine class prediction based on the dataset from UCI ML library

Learning Goals

- Learn to predict categorical data types, for example – customer segmentation, spam detection etc.
- Evaluate and tune classification models
- Hyper parameter tuning

Day 4 – Unsupervised algorithms – clustering, PCA and Text Analytics

Topic	Description
Clustering	• What is clustering and use cases • Proximity Matrices – find dissimilarity between two observations o Choice of attributes o Units of measure of attributes o Importance of scaling • Determining number of clusters • K-means clustering • Practical Issues with K-means clustering • Measure performance of clustering • Hierarchical clustering • Practical Issues in clustering
Lab Exercises	• Clustering movies into genres based on movie attributes • Apply clustering on images for clinical diagnostics
Feature Engineering	• Rescaling • Feature selection techniques • Principle Component Analysis
Lab Exercise	• Running PCA on wine classification dataset from UCI ML dataset
Text Analytics	• Text Analysis Steps • Term Frequency – Inverse Term Frequency • POS tagging, lemmatization, stemming • Collecting raw text • Representing text • Categorize Documents by type • Determining sentiments • Gaining insights
Lab Exercises	• Classification on IMDB comments dataset

Learning Goals

- Apply clustering algorithms to gain insights about latent groups inside a dataset
- Identify properties of the features that impact model performance
- Transform data structure to facilitate model performance and computation speed
- Learn technique to analyze textual data and prepare it for further analysis such as classification, regression, clustering etc.
- Extract information from text in the form of word cloud, sentiment

Day 5 – Scalable Machine Learning and Machine Learning Deployments

Topic	Description
Scalable Machine Learning	• Challenges with large scale machine learning • Deep dive into Spark MLlib • Spark R integrate between R with Spark • Spark ML packages for regression, classification, clustering, recommender system • Saving the trained model on persistent storage • Serving model for online prediction • [Optional] Using Prediction IO for deployment machine learning pipelines • [Optional] Machine learning for streaming application
Lab Exercise	• Process data from stack-overflow dataset • Find community from stack-overflow dataset using tag analysis

Learning Goals

- Learn how to build machine learning pipeline on massive data using Spark framework
- Deployment options for machine learning applications for online and offline applications.