Introduction to Machine Learning

Prepared for Intuit

Last Updated: Aug 3, 2023

Course Overview:

Machine Learning (ML) is the killer app for Big Data. The increased availability of compute resources through myriad cloud providers such as AWS and the proliferation of tools — from software libraries to ML "as a service" providers — have brought the power of ML to all kinds of programmers. However, to use ML effectively, one needs to understand the models, how to use them, how to evaluate their performance, and how to position ML models into a larger software ecosystem.

This course is intended for data scientists and software engineers. It maintains an optimal balance of theory and practice. For each machine learning concept, we first discuss the foundations, its applicability and limitations. Then we explain the implementation and use, and specific use cases. This is achieved through a combination of about 50% lecture and 50% lab work.

Duration: This course will be delivered in 3 days.

Audience: Data Scientists and Software Engineers

Prerequisites:

- Background in high school level statistics and math
- B.S. Computer Science degree or equivalent professional experience
- Basic Python experience (functions, classes, packages)
- Basic understanding of cloud computing and core cloud services
- Familiarity with programming in at least one language
- Be able to navigate Linux command line

Lab environment: Students may install crucial Python libraries on their system or work on Google Colaboratory to complete this training without installing anything on their own machines.

Objectives :

- Attain thorough understanding of popular machine learning algorithms, their applicability and limitations.
- Practice the application of these methods. Achieve clarity in the real-world use of machine learning by illustrating each method with practical use cases.

This Class DOES Include:

- Introduction to machine learning principle
- Supervised learning algorithms focus, with minor coverage of unsupervised learning.
- Linear regression, Logistic regression, Random Forests, Gradient Boosted Decision Trees, Neural Net basics
- Introduction to data processing using NumPy and Pandas
- Introduction to data visualization using Matplotlib
- Intermediate hands-on experience with Scikit Learn
- Beginner-Intermediate hands-on experience with Tensorflow 2.x
- Hands-On Labs and Exercise

This Class Does Not Include:

- Amazon SageMaker
- Advanced AWS
- Deep learning
- Proofs or theoretic motivation

Course Outline:

Introductions and overviews

- Data Collection and Preparation
    - o Extract/Transform/Load (ETL)
    - o Exploratory Data Analysis (EDA)
    - o “Data cleaning” vs “data preprocessing”
    - o Data as a source of bias.

- Machine learning
    - o Supervised vs unsupervised.
    - o Training, Test, and Validation datasets.
    - o Evaluation criteria.

- The ML product lifecycle.

Supervised Learning

- Linear regression
- Logistic regression and multinomial logistic regression
- decision trees, random forests, gradient boosting
- Overview of neural networks
- Labs for every section above, in Jupyter.

Unsupervised learning

- K-Means clustering

Data visualization

- Visualization examples for the models above
- Links to other visualizations for self-study