Comprehensive Machine Learning for CyberSecurity and GEN AI (5 Days)

By Dr. Vishwanath Rao

Overview

Comprehensive Machine Learning (ML) with Python training course builds on our
Comprehensive Data Science with Python class and teaches attendees how to write machine
learning applications in Python.

Prerequisites

Basic knowledge on python

Objectives

- Understand machine learning as a useful tool for predictive models
- Know when to reach for machine learning as a tool
- Implement data preprocessing for an ML workflow
- Understand the difference between supervised and unsupervised tasks
- Implement several classification algorithms
- Evaluate model performance using a variety of metrics
- Compare models across a workflow
- Implement regression algorithm variations
- Understand clustering approaches to data
- Interpret labels generated from clustering
- Transform unstructured text data into structured data
- Understand text-specific data preparation
- Visualize frequency data from text sources
- Perform topic modeling on a collection of documents
- Use labeled text to perform document classification

Outline

- Introduction
- Review of Core Python Concepts
  - o Anaconda Computing Environment
  - o Importing and manipulating Data with Pandas
  - o Exploratory Data Analysis with Pandas and Seaborn
  - o NumPy ndarrays versus Pandas Dataframes

- An Overview of Machine Learning
  - Machine Learning Theory
  - Data pre-processing
    - Missing Data
    - Dummy Coding
    - Standardization
    - Data Validation Strategies
  - Supervised Versus Unsupervised Learning

- Modeling for explanation (descriptive models)
  - Understanding the linear model
  - Describing model fit
  - Adding complexity to the modelo Explaining the relationship between model inputs and the outcome
  - Making predictions from the model

- Supervised Learning: Regression
  - Linear Regression
  - Penalized Linear Regression
  - Stochastic Gradient Descent
  - Decision Tree Regressor
  - Random Forest Regression
  - Gradient Boosting Regressor
  - Scoring New Data Sets
  - Cross Validation
  - Variance-Bias Tradeoff
  - Feature Importance

- Supervised Learning: Classification
  - Logistic Regression
  - LASSO
  - Support Vector Machine
  - Random Forest
  - Ensemble Methods
  - Feature Importance
  - Scoring New Data Sets
  - Cross Validation

- Unsupervised Learning: Clustering
  - Preparing Data for Ingestion

o K-Means Clustering
o Visualizing Clusters
o Comparison of Clustering Methods
o Agglomerative Clustering and DBSCAN
o Evaluating Cluster Performance with Silhouette Scores
o Scaling
o Mean Shift, Affinity Propagation and Birch
o Scaling Clustering with mini-batch approaches

• Clustering for Treatment Effect Heterogeneity
o Understand average versus conditional treatment effects
o Estimating conditional average treatment effects for a sample
o Summarizing and Interpreting

• Data Munging and Machine Learning Via H20
o Intro to H20
o Launching the cluster, checking status
o Data Import, manipulation in H20
o Fitting models in H20
o Generalized Linear Models
o naïve bayes
o Random forest
o Gradient boosting machine (GBM)
o Ensemble model building
o automl
o data preparation
o leaderboards
o Methods for explaining modeling output• Introduction to Natural Language Processing (NLP)
o Transforming Raw Text Data into a Corpus of Documents
o Identifying Methods for Representing Text Data
o Transformations of Text Data
o Summarizing a Corpus into a TF—IDF Matrix
o Visualizing Word Frequencies

• NLP Normalization, Parts-of-speech and Topic Modeling
o Installing And Accessing Sample Text Corpora
o Tokenizing Text
o Cleaning/Processing Tokens
o Segmentation
o Tagging And Categorizing Tokens
o Stopwords

- Vectorization Schemes for Representing Text
- Parts-of-speech (POS) Tagging
- Sentiment Analysis
- Topic Modeling with Latent Semantic Analysis

- NLP and Machine Learning
  - Unsupervised Machine Learning and Text Data
  - Topic Modeling via Clustering
  - Supervised Machine Learning Applications in NLP
- Conclusion