Machine Learning using Python Deep Dive (5 Days)

Overview

In this Python for ML training course, attendees take a deep dive into machine learning, including supervised and unsupervised learning, regression, classification, and clustering. Students also learn how to implement ML algorithms in Python, a popular programming language for machine learning.

Skills Gained

- 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

Prerequisites

All attendees should have completed the Comprehensive Data Science with Python class or have equivalent experience.

Course Outline

Introduction

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

An Overview of Machine Learning

Machine Learning Theory
Data pre-processing
Supervised Versus Unsupervised Learning

Modeling for explanation (descriptive models)

Understanding the linear model
Describing model fit
Adding complexity to the model
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
K-Means Clustering
Visualizing Clusters
Comparison of Clustering Methods
Agglomerative Clustering and DBSCAN
Evaluating Cluster Performance with Silhouette Scores
Scaling
Mean Shift, Affinity Propagation and Birch
Scaling Clustering with mini-batch approaches

Clustering for Treatment Effect Heterogeneity

Understand average versus conditional treatment effects
Estimating conditional average treatment effects for a sample
Summarizing and Interpreting

Data Munging and Machine Learning Via H20

Intro to H20
Launching the cluster, checking status
Data Import, manipulation in H20
Fitting models in H20
Generalized Linear Models
naïve bayes
Random forest
Gradient boosting machine (GBM)
Ensemble model building
automl
data preparation
leaderboards
Methods for explaining modeling output

Introduction to Natural Language Processing (NLP)

Transforming Raw Text Data into a Corpus of Documents
Identifying Methods for Representing Text Data
Transformations of Text Data
Summarizing a Corpus into a TF—IDF Matrix
Visualizing Word Frequencies

NLP Normalization, Parts-of-speech and Topic Modeling

Installing And Accessing Sample Text Corpora
Tokenizing Text
Cleaning/Processing Tokens
Segmentation
Tagging And Categorizing Tokens
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