Course Outline

Python and Machine Learning

Course Summary

Description

Python is a popular open source language. It has libraries for almost everything, including web programming, administrative tasks, system programming, mathematics, machine learning, and graphics. This course is intended for data scientists and software engineers. It gives them practical level of experience, achieved through a combination of about 50% lecture, 50% lab work.

Topics

- Python Introduction
- Python Language Overview and First Steps
- Python OOP
- Pandas
- NumPy
- Python – DB Programming
- Python – Web Programming
- Visualization
- NLTK
- Machine Learning (ML) Overview
- Machine Learning Environment
- Machine Learning Concepts
- Feature Engineering (FE)
- Linear regression
- Logistic Regression
- Classification : SVM (Supervised Vector Machines)
- Classification : Decision Trees & Random Forests
- Classification : Naive Bayes
- Clustering (K-Means)
- Principal Component Analysis (PCA)
- Recommendation (Collaborative filtering)
- Final workshop (time permitting)

Audience

This course is designed for Data Scientists, Developers, and Administrators.

Prerequisites

Before taking this course, students should be able to navigate Linux command line, and have familiarity with programming

Duration

Five days

Course Outline

Python and Machine Learning

Course Outline

I. Python Introduction
- A. Installing Python
- B. Python Versions
- C. IDEs
- D. Jupyter Notebook

II. Python Language Overview and First Steps
- A. Data Types
- B. NumPy
- C. Packages
- D. Pandas

III. Python OOP
- A. Classes
- B. Modules/Packages
- C. Python Packages
- D. Data Types

IV. Pandas
- A. DataFrames
- B. Schema inferences
- C. Data exploration

V. NumPy
- A. Capabilities
- B. Data types
- C. Packages

VI. Python – DB Programming
- A. Database Connectivity
- B. Pandas and DB
- C. ORM

VII. Python – Web Programming
- A. Python Web Frameworks
- B. Flask
- C. Restful API with Flask

VIII. Visualization
- A. Pandas visualization
- B. Matplotlib
- C. Seaborn
- D. Ggplot
- E. Doing Data Science with Scikit-learn
- F. Introducing Scikit-Learn
- G. Clustering Data
- H. Building a Classifier

IX. NLTK
- A. Bag-of-words (NLTK labs in python)
- B. Bag-of-n-Grams
- C. Filtering (NLTK labs, later-spacy)
- D. Stopwords
- E. Frequency-based
- F. Stemming
- G. Parsing and tokenization
- H. TF-IDF
- I. SpaCy for semantic pipeline and named entity recognition

X. Machine Learning (ML) Overview
- A. Machine Learning landscape
- B. Machine Learning applications
- C. Understanding ML algorithms & models (supervised and unsupervised)

XI. Machine Learning Environment
- A. Introduction to Jupyter notebooks / R-Studio
- B. Lab: Getting familiar with ML environment

XII. Machine Learning Concepts
- A. Statistics Primer
- B. Covariance, Correlation, Covariance Matrix
- C. Errors, Residuals
- D. Overfitting / Underfitting
- E. Cross validation, bootstrapping
- F. Confusion Matrix
- G. ROC curve, Area Under Curve (AUC)
- H. Lab: Basic stats

XIII. Feature Engineering (FE)
- A. Preparing data for ML
- B. Extracting features, enhancing data
- C. Data cleanup
- D. Visualizing Data
- E. Lab : data cleanup
- F. Lab: visualizing data

Course Outline (cont'd)

XIV. Linear regression
- A. Simple Linear Regression
- B. Multiple Linear Regression
- C. Running LR
- D. Evaluating LR model performance
- E. Lab
- F. Use case: House price estimates

XV. Logistic Regression
- A. Understanding Logistic Regression
- B. Calculating Logistic Regression
- C. Evaluating model performance
- D. Lab
- E. Use case: credit card application, college admissions

XVI. Classification : SVM (Supervised Vector Machines)
- A. SVM concepts and theory
- B. SVM with kernel
- C. Lab
- D. Use case: Customer churn data

XVII. Classification : Decision Trees & Random Forests
- A. Theory behind trees
- B. Classification and Regression Trees (CART)
- C. Random Forest concepts
- D. Labs
- E. Use case: predicting loan defaults, estimating election contributions

XVIII. Classification : Naive Bayes
- A. Theory behind Naive Bayes
- B. Running NB algorithm
- C. Evaluating NB model
- D. Lab
- E. Use case: spam filtering

XIX. Clustering (K-Means)
- A. Theory behind K-Means
- B. Running K-Means algorithm
- C. Estimating the performance
- D. Lab
- E. Use case: grouping cars data, grouping shopping data

XX. Principal Component Analysis (PCA)
- A. Understanding PCA concepts
- B. PCA applications
- C. Running a PCA algorithm
- D. Evaluating results
- E. Lab
- F. Use case: analyzing retail shopping data

XXI. Recommendation (Collaborative filtering)
- A. Recommender systems overview
- B. Collaborative Filtering concepts
- C. Lab
- D. Use case: movie recommendations, music recommendations

XXII. Final workshop (time permitting)
- A. Students will analyze a couple of datasets and run ML algorithms. This is done as a group exercise. Each group will present their findings to the class.