Data Science Boot Camp

Presented by: Paul F. Roysdon, Ph.D.

Summary

This 2-week boot camp is for aspiring data scientists. We will take each student from an introduction to Jupyter Notebooks and Python, to the state-of-the-art in Deep Learning. Each section is organized with 30% lecture + 70% hands-on training. Each student will develop a uniquely tailored Capstone Project and present it at the end of the course.

Pre-boot camp Preparation

- Download and install Anaconda v1.9.7.
  - https://www.anaconda.com/distribution/
- Run Jupyter Notebook "Welcome to DSBC" and report the Student ID on the first day of class.
  - https://github.com/dsbc2020/bootcamp/tree/master/notebooks/_welcome_

Time & Location

- **Dates**: September 23rd – October 4th, 2019.
- **Times**: 8:30am – 5:00pm, 30 min lunch @ 12:00pm, 15 min breaks @ 10:30am & 2:00pm.*
- **Location**: National Security Collaboration Center, North Paseo Building, Suite 4.212 (4th floor), University of Texas San Antonio.

Format

- 80 hours: 30% lecture + 70% Labs, Exercises and Demos.
- 20 hours: Capstone Project (assigned as homework).
- 38 modules, with a quiz at the end of each module.
- Cumulative final exam based on quizzes. Passing score is 70% for JQS (beginner-level).

Prerequisites

- Interest in data science, data engineering, and programming.
- No experience necessary in data science. I will teach you what you need to know.
- Familiarity with Python and Microsoft Excel.

Requirements

- 100% attendance required to pass the course. No exceptions.
- Be 30 min early each day for security processing and check-in.
- Personal laptop to run Jupyter Notebooks on Google Collab.

* If we finish early, there will be cubicles available to work on the Capstone Projects.

Schedule (draft)

Day 1: Introduction to the Data Science

- AM: Introduction to Data Science, and Predictive Analytics, Anaconda and Python Programming.
- PM: EDA (Exploratory Data Analysis), Visualization, and Feature Engineering.

Day 2: Classification and Regression Algorithms

- AM: Linear algebra review. Evaluating Performance, Visualizing Features and Parameters.
- PM: Introduction to Predictive Modeling. Building a Classifier.

Day 3: Intro Machine Learning & Communicating Results

- AM: Statistics review. Intro to time series modeling, intro to Bayes Theorem.
- PM: Hypothesis testing, Communicating results with a customer.

Day 4: Unsupervised Learning

- AM: Calculus review, gradient descent. Dimension reduction (PCA, SVD, LSI).
- PM: Clustering (k-means, k-modes, hidden Markov models, expectation-maximization).

Day 5: Supervised Learning – Classification

- AM: k-neighbors, decision tree, naïve Bayes.
- PM: gradient boosting, bootstrapping, bagging, AdaBoost.
- **Start Capstone Project.**

Day 6: Supervised Learning – Regression & Prediction

- AM: random forest, least squares (LS), weighted least squares (WLS).
- PM: linear and integer linear programming (LP and ILP), extended Kalman filtering (EKF).

Day 7: Outlier Detection and Neural Networks

- AM: L1/L2 regression.
- PM: artificial neural networks (ANN).

Day 8: Deep Learning

- AM: Support vector machines (SVM), convolutional neural networks (CNN).
- PM: recurrent neural networks (RNN), long short-term memory (LSTM).

Day 9: Deep Learning

- AM: gated recurrent units (GRU), generative adversarial networks (GAN).
- PM: transformers.

Day 10: ML Ethics & Capstone Presentations

- AM: ML Ethics. Presentations.
- PM: Exam

Detailed Schedule

Day 1

- Introduction to Data Science
- IPython
  - Beyond Normal Python
  - Help and Documentation
  - Shell Keyboard Shortcuts
  - Magic Commands
  - Input Output History
  - IPython and Shell Commands
  - Errors and Debugging
  - Timing and Profiling
- Python Summary
- NumPy
  - Introduction to NumPy
  - Understanding Data Types
  - The basics of NumPy Arrays
  - Computation on Arrays UFuncs
  - Computation on Arrays Aggregates
  - Computation on Arrays Broadcasting
  - Boolean Arrays and Masks
  - Fancy Indexing
  - Sorting
  - Structured Data
- Pandas
  - Introduction to Pandas
  - Introduction to Pandas Objects
  - Data Indexing and Selection
  - Operations in Pandas
  - Missing Values
  - Hierarchical Indexing
  - Concat and Append
  - Merge and Join
  - Aggregation and Grouping
  - Pivot Tables
  - Working with Strings
  - Working with Time Series
  - Performance Eval and Query
- MatPlotLib and Seaborn
  - Introduction to Matplotlib
  - Simple Line Plots
  - Simple Scatter Plots
  - Error Bars
  - Density and Contour Plots
  - Histograms and Binnings
  - Customizing Legends
  - Customizing Colorbars
  - Multiple Subplots
  - Test and Annotation
  - Customizing Ticks
  - Settings and Stylessheets
  - Three Dimensional Plotting
  - Geographic Data with Basemap
  - Visualization with Seaborn

Day 2

- Linear Algebra
  - Why Linear Algebra?
  - Background Removal with PCA
  - Linear Regression
- Exploratory Data Analysis
  - Introduction to EDA
  - Data Cleaning
  - Data Visualization
  - Data and Models
  - Example: Vaccine Analysis
- Model Validation
- Introduction to SciPy
- Visualization Examples
  - NetworkX
  - Airports
  - GIS
  - KDE
  - GPS

Day 3

- Introduction to Machine Learning
  - Machine Learning
  - What is Machine Learning?
  - Introduction to SciKit-Learn
  - Hyper-parameters & Model Validation
  - Feature Engineering
- Statistics
  - Statistics for Hackers
  - Introduction to NumPy Stats
  - Introduction to Statsmodels
- Clustering

    - K-Means and EM
    - Gaussian Mixtures
- Decision Tree & Random Forest
- Naïve Bayes

Day 4

- Calculus & Gradient Descent
- Linear Regression
- Principle Component Analysis
- Manifold Learning

Day 5

- K-Neighbors
- Kernel Density Estimation
- Logistic Regression
- Neural Networks

Day 6

- Linear Programming
- Filtering
- Multi-Layer Perceptron
- Natural Language Translation
    - NLTK
    - Embedding's
- Support Vector Machines

Day 7

- Introduction to Deep Learning
- Convolutional Neural Networks
- Recurrent Neural Networks

Day 8

- LSTM
- Introduction to Adversarial Machine Learning

Day 9

- Generative Adversarial Network
- Transformers

Day 10

- Ethics in Machine Learning
- Exam
- Capstone Presentations

Capstone Project

The Capstone Project is designed for the student to apply what they learned to a specific problem using a supplied data-set. From the list below, each student will

- Select **one** dataset to use for their project (each student will use different datasets).
- The students are **encouraged** to **enrich their data** with other sources of information, e.g. combine LA traffic collision data with NOAA weather service data to correlate collisions with inclement weather conditions.
- The student shall create a Jupyter Notebook, and use any and all methods discussed in the lectures and tutorials to perform an analysis as a Data Scientist.
- The Notebook shall include **all** of the following sections:
    - Exploratory Data Analysis
    - Feature Engineering
    - Model Selection
    - At least 3 analysis techniques, e.g. PCA, linear regression, manifold learning.
    - Visualizations
    - Summary/ Conclusions
- The notebook should tell a story to their "customer" about the data, and derive conclusions based on the algorithms applied to the data. Any assumptions, caveats, or important features should be noted, so that the customer is informed and not misled.

The available data-sets are:

- 188 million US wildfires
- AMEX, NYSE, and NASDAQ stock histories
- Cars: manufacturer, models, prices
- Craigslist Car & Trucks ads
- Credit Card Fraud in the US
- Boston Crime data
- Denver Crime data
- Earthquakes
- Hourly energy consumption in major US cities
- Intel Corp. image classification
- Los Angeles traffic collision data
- Mt. Renier weather and climbing data
- NYC AirBnB data
- PGA Tour data
- CDC data 1918-2008
- TMDB 5000 movie data
- US minimum wage by state 1968-2017
- YouTube trends

The datasets are available at:
https://drive.google.com/open?id=1dSEP0UdYXGDWeCqIIca67KTalP93axt4