Data Science with Kaggle Syllabus Fall 2016

Introduction:

Welcome to Data Science with Kaggle! Kaggle is home to an abundant source of company-volunteered data that encourage data scientists from around the world to solve proposed, and often business-related, challenges. The platform fosters a great amount of knowledge sharing, competition, and practical relevance where beginners and experts alike benefit from an exponentially expanding field.

Prerequisites:

This class is a projects-based class with a focus on machine learning. You are expected to have some programming or statistics backgrounds and so the material will be of greatest benefit to sophomores or those who have taken CS61A, DATA 8, STAT 133, or equivalent. However, the first two weeks of class will be an optional python bootcamp for those taking the course with absolutely no programming background. By the end, you can determine whether you are comfortable continuing through the course.

Note that this is not an easy class. The student facilitators intend to provide you with a comprehensive guide to data analysis with the goal of preparing you for industry and, if demonstrated superb interest, future machine learning competitions.

Learning objectives:

Students are expected to gain proficiency in modeling and coding in several topics, including but not limited to the following list:

1. Python programming
2. Data interpretation, data munging, and visual analysis
a. Numerical and Text
3. Linear and Logistic Regression
4. Clustering Techniques
5. RandomForest
6. TensorFlow
7. Assorted Topics (Microsoft Azure, AWS, SQL, etc.)

The objectives will be met by completing projects and assignments.

Project Schedule:

Each project will last for two weeks. Projects can be done in teams up to 4 or individually. Note that actual data sets may change throughout the semester.

- The first two weeks will involve an analysis of House Prices: Advanced Regression Techniques. This is a live Kaggle knowledge competition with a bunch of relevant variables that impact a home's final sale price. All basic data munging and linear modeling techniques can be used in this competition.

- MNIST Digit Recognizer – Clustering Techniques

- Auto-Librarian Classification – RandomForest, text analysis, and clustering techniques

- Company-sponsored In Class Kaggle Competition - All concepts

Assignment Schedule:
Python bootcamp assignments will involve simple practice problems to get you more familiar with Python, specifically numpy, matplotlib, and pandas.

Subsequent assignments will involve in-class kaggle competitions where students submit their model predictions on a custom arranged data set separate from lecture. This will give you a chance to apply what you have learned in class. These assignments should be done individually.

Class Logistics:
First two weeks will be an optional programming bootcamp for students to get introduced to programming and to catch up to the level of skill necessary to complete the course.

There will be 3 hours of lecture per week in addition to office hours. In-class Kaggle assignments will be given at relevant instruction periods but we will try to consistently start them at the beginning of a week and they will last for one full week.

The first hour of lecture will introduce you to the concept and the code. The remaining time will be dedicated to “sprints” where we will go on Kaggle and replicate someone’s model or build a quick model to submit on Kaggle’s leaderboard.

Grading:
Python bootcamp assignments will be graded on completeness. In-class Kaggle assignments will be graded on whether the prediction score is above a certain threshold. Final projects will be graded on completeness, quality of response, accuracy, and team mate evaluations. There will be 4 open-ended projects throughout the semester. All assignments, projects, and attendance will be assigned a point value with a general weighting scheme of 40% assignments, and 60% projects. In order to pass the class, you must pass the minimum accuracy threshold for all assignments and projects.

Required Online Reading Schedule:
These are hand-picked resources the student instructors strongly believe will help you understand lecture.

9/19 - https://www.kaggle.com/c/titanic/details/getting-started-with-python
9/26 - https://www.kaggle.com/c/titanic/details/getting-started-with-python-ii
10/3 - https://www.dataquest.io/blog/k-nearest-neighbors-in-python/
10/10 - http://opencvpython.blogspot.com/2012/12/k-means-clustering-1-basic-understanding.html
10/17 - https://www.kaggle.com/c/word2vec-nlp-tutorial/details/part-1-for-beginners-bag-of-words
10/24 - http://blog.yhat.com/posts/random-forests-in-python.html
10/31 - http://natureofcode.com/book/chapter-10-neural-networks/ (Introduction and Section 10.2)

Recommended Texts and Online Readings:

Most material will be in the form of powerpoint slides, handouts, and live demos. In addition, there are a few resources we recommend reading throughout the course to better understand concepts or a programming language. Some are really just fun reads.

The Data Science Handbook by Carl Shan, Henry Wang, William Chen, and Max Song: http://www.thedatasciencehandbook.com
Python for Data Analysis by Wes McKinney: http://shop.oreilly.com/product/0636920023784.do
The Signal and the Noise by Nate Silver: https://en.wikipedia.org/wiki/The_Signal_and_the_Noise
Book for nitty-gritty of neural networks:
http://neuralnetworksanddeeplearning.com/
Recurrent Neural Networks:
http://karpathy.github.io/2015/05/21/rnn-effectiveness/

Extra Credit:

Students will be able to receive extra credit for completing a side project with instructor approval.

Attendance:

Since this is a project-team-based class, attendance is **mandatory**. We will be keeping track at the beginning of each class. However, you may have two absences for any reason. If you are working with a team, please communicate appropriately.

Class Schedule:

9/7 - Python setup. Coding environment setup. Jupyter notebooks.

9/12 - Variables, objects, loops, numpy, matplotlib, and pandas.

9/14 – Demonstration of Kaggle Poker Induction model.

9/19 - Overview of class. Introduction of Titanic data set. Hands-on data exploration. Plots. Summary statistics. Introduce Home Prices Project.

Form Teams Here. Deadline to submit team proposals is 9/26.

9/21 - Data cleaning. Regular expressions. Home Depot data set.

9/26 - Linear Regression. Assumptions for intuition. Interpretation. Practical example on Wine dataset. Example of linear dependence (Santander data set add-on).

9/28 - Logistic Regression. Assumptions for intuition. Difference from Linear Regression. Practical example on Titanic. First Project due 10/2 midnight.

10/3 - Introduction of Digit Recognition data set. Linear and logistic regression. KNN clustering.

10/5 - K-means

10/10 - In-class implementation of KNN.

10/12 - Validation Method. Cross Validation. Review of all models. KNN assignment due 10/16

10/17 - Introduction of Auto-Librarian data set. NLP: bag-of-words model.

10/19 - RandomForest.

10/24 - Review. More practice/catchup day.

10/26 - Guest speaker. Auto-librarian project due on 10/30 midnight.

10/31, 11/2, 11/7, -11/9 - Introduction to neural networks via TensorFlow.. Final Project introduced.

11/14 - 11/30 - Miscellaneous topics: Microsoft Azure, Tableau, SQL, AWS, etc. Final project due on 12/9.

12/6, 12/8 - Kaggle Master Guest Lecturers