SML 312 Tentative Syllabus for Research Projects in Data Science Fall 2022

Instructor: Jonathan Hanke
Instructor E-mail: jonhanke@princeton.edu
Meeting Times: Mon/Tues 7:30-8:50pm
Office Hours: Mon/Tues 9-10pm and by appointment

Preceptor: Daniel Melese
Preceptor E-mail: dm3862@princeton.edu
Precept Meeting Times: Friday 1:30-2:20pm (possibly subject to change)
Office Hours: TBD

Course Description: This course is designed from the ground up to give students a broad immersive experience with the core algorithms, main tools, and many important choices involved in doing real-world Data Science -- the art of using data to gain insights and answer compelling questions.

The course pursues two tracks simultaneously:

1. Give a broad perspective of how to approach the main types of data science problems (Classification and Regression), as well as some cutting-edge speciality areas (Text / Natural Language Processing and Image Classification / Neural Networks).

2. Develop the technical competencies and proficiencies with core tools to carry out (often messy) real-world analyses -- both in the context of class "Mini-Project" assignments, and also for a personal data science "Final Project" designed and executed by each student to answer a compelling self-initiated question with data over the course of the semester.

At the end of the course, students will create a 12-15 page writeup of their original data science work, present their work to their peers with a 15-20-minute "Data Science Seminar" talk, and post their work as code in a publicly accessible industry-standard format (Github) that can be used later to showcase their data science skills to other researchers and prospective employers.

Overview of Course Topics:

- The Language of Probability and Statistics for Data Science
- Exploratory Data Analysis (EDA) & Data Visualization
- Regression Models -- Linear, multivariate, polynomial
- Exact Regression solutions and Implicit Variable Bias
- Regularization and Robust models to handle outliers
- Loss Functions and the Mean Squared Error
- (Stochastic) Gradient Descent

- Finding "good" models and measures of "goodness"
- Feature Selection / Multicollinearity
- Cross Validation via Test / Train splits
- Explaining variable significance in models
- Overfitting / Underfitting and the role of noise in modelling
- One-Hot Encoding and Label Encoding
- Classification Models
- Decision Trees / Random Forests / Splitting Criteria
- K-Nearest Neighbors
- Naive Bayes models
- Logistic Regression
- Neural Networks / Network Architectures / Activation functions
- Training Neural Networks
- Hyperparameters and Learning Rates
- Text Processing -- stop words / stemming / lemmatization
- Sentiment Analysis and Lexicons
- Vector embeddings to encode semantic meaning / Word2Vec
- Cosine Similarity
- Topic Modelling / Latent Dirichlet Allocation / TF-IDF scores
- Confusion matrices -- Accuracy / Precision / Recall
- Prediction Cutoffs and the Receiver Operating Characteristic Curve
- Choosing appropriate measures of goodness for specific applications
- Image encoding and Convolutional Neural Networks (CNNs).

Course Grading Policy:

Your final course grade will be determined from a combination of your final project (with parts for proposals, code, final writeup, project content, and a final presentation to your peers), 3 course mini-projects, homework assignments, and participation (in Slack and in Zoom) according to the following weights:

Final Project Code: (5%)
Final Project Proposals (10%)
Final Project/Writeup: (35%)
Final Project Presentations: (10%)
Class Mini-Projects: (30%)
Attendance/Participation: (10%)

All assignments will be given a due date, and are required to be submitted by e-mail to jonhanke@princeton.edu and dm3862@princeton.edu by 7pm (i.e. 30 minutes before class starts) on the due date unless otherwise specified. Late assignments may be accepted for up to one week after the due date under exceptional circumstances.

Course Technology:

Slack: One of our main tools for communicating as a group is the "Princeton SML312 -- Fall 2022" Workspace at https://princeton-jt88710.slack.com. The instructor and preceptor will be regularly looking at the Slack channels to answer any questions you have about the course, assignments, or data science in general! You're also highly encouraged to use these as a way of communicating with your peers about project ideas, related notes articles online, or the latest meme of the week.

Github: As part of the class you will be asked/guided to create an account at http://github.com to share the code for some of your class projects (e.g your final project). Having well-documented code available online is a great way to demonstrate experience and projects to other researchers and potential employers interested in your work.

Python/Jupyter/Anaconda/Colab: While there are many languages available for doing data science, one of the most popular is the Python programming language (https://www.python.org/), often run in your browser in a Jupyter notebook (https://jupyter.org/). One of the easiest ways to install this on your computer is with the Anaconda distribution (https://www.anaconda.com/products/individual) which includes Python, Jupyter, and many other useful Python libraries (e.g. sklearn). If you'd prefer to run Python in the cloud instead, there's a free Google Colab service for running Jupyter notebooks on their servers (https://colab.research.google.com/).

Course Etiquette: In these unusual times it's worth giving some thought to best practices for how to create a welcoming respectful environment where everyone can learn effectively and share ideas freely. Here are some thoughts about how best to do that:

- Please silence your phones and other devices and/or put them away in another room to not distract you or your peers.
- Please give your colleagues your full attention during our class meetings.
- Please feel free to share your ideas and questions with the class. This can be a spoken comment (please feel free to do this if you have a question!), or a posting to our Slack channel. If you have a question, it's likely that at least a few other people do as well, so please ask!

Course COVID Policies: Due to the continuing risks from COVID-19 and related variants, this is a mask-mandatory class (see e.g. https://ehs.princeton.edu/covid19-faqs-faculty). Please be sure to wear your masks while in the classroom for your own safety and the safety of others.