Syllabus: Applications of Data Science

Giora Simchoni

Goals:

- Learn and practice some of the building blocks of the Data Science profession
- Move forwards with a first-degree student knowledge: realistic datasets and previously uncovered topics.
- Practice the DS tech stack: R, Python (yes both), Git, Github and Docker
- Build a deliverable Data Science project from start to finish

Topics:

Part I: Data Wrangling in the Tidyverse (R) and in pandas (Python)

- The Tidy Data philosophy
- Cleaning, manipulating, summarizing and visualizing data, in a consistent and tidy manner
- Modeling in the Tideverse
- A visit to Python's Tidyverse-equivalent: Pandasverse

Part II: Intro to Network Analysis with NetworkX (Python)

- Empirical study of networks
- Network Theory: measures and metrics of a network and its individual nodes and edges
- Network Math: linear algebra in the service of networks, implemented in NetworkX
- Basic Network Applications: Community Detection and Network models

Part III: More Predictive Modeling

- Revisit Ensemble Methods: From CART to Random Forests and XGBOOST
- Topics in Classification: class imbalance, multiple classes, active learning
- Topics in Regression: Regression's relatives (CCA, PLS)
- Feature Engineering

Part IV: Deep Learning – beyond the basics

- Demystifying Deep Neural Networks
- Learning and toying with 2-3 types of Deep Learning beyond the "basics", as time permits: Autoencoders and Representation Learning, GANs, Style Transfer, LSTM for text or sound, Reinforcement Learning -TBD and dependent on the class collective experience

Part V: Building a DS Project (if time permits)

- Upscaling visualizations of data to a more modern, interactive, browser-appropriate level
- Learn Shiny (R) or Dash (Python) to build an online "Data App"

Expected Tech Stack:

- R: tidyverse packages (e.g. dplyr, purr and ggplot2), shiny, varying modeling packages (e.g. infer, parsnip, recipes), perhaps keras
- Python: pandas, numpy, scipy, scikit-learn, keras (tensorflow-based), NetworkX
- Git and Github: all students are expected to use Git and Github for version control and for sharing code at the very basic level. For those unfamiliar with Git a self-learning introduction will be given.
- Docker: the bare necessities to make it work
- Cloud services: RStudio Cloud, Binder

Pre-requisites:

- Statistical Learning or equivalent plus all Statistical Learning pre-requisites (Calculus, Linear Algebra, Probability, Intro to Stats)
- Familiarity with R or with Python (students with neither R or Python experience might find the task of mastering both a bit daunting)

Language:

- Course will be taught in Hebrew but all written materials and assignments will be in English

Grading:

- Homework: 4-7 home assignments (10% each)
- Final Project: A deliverable, interactive “Data App”, demonstrating the student’s mastery in class topics (100% - (n_assignments - 1) * 10%)