DATA 1030 COURSE PROPOSAL SYLLABUS
FALL 2019
DATA SCIENCE INITIATIVE
INSTRUCTOR: ANDRAS ZSOM
BROWN UNIVERSITY

INFORMATION Course title: *Hands-on Data Science*. Meets TTh 13:00 to 14:20 in CIT 227.

COURSE DESCRIPTION Develops all aspects of the data science pipeline: data acquisition and cleaning, handling missing data, exploratory data analysis, visualization, feature engineering, modeling, interpretation, presentation in the context of real-world datasets. Fundamental considerations for data analysis are emphasized (the bias-variance tradeoff, training, validation, testing). Classical models and techniques for classification and regression are included (linear regression and logistic regression, support vector machines, decision trees, ensemble methods). Uses the Python data science ecosystem (numpy, pandas, matplotlib, plotly, scikit-learn).

RATIONALE This course equips students with the wide variety of general skills they will need to solve data science problems as a researcher or practitioner.

LEARNING GOALS Students will be able to complete data science projects from the initial question to final presentation. Students will be able to acquire and clean data, explore the data visually, apply models, discuss the advantages and disadvantages of particular techniques, and interpret and present their findings.

ASSESSMENT AND EVALUATION CRITERIA Assessment in DATA 1030 is based on weekly homework assignments (including written and computational exercises, 20%), class participation (10%), two exams (25%), and one project (45%). The project will entail building machine learning pipeline which applies the ideas developed in the course to solve a real-world problem. Due dates for the midterm are as indicated in the schedule below and will be released with at least three weeks of lead time. Students will be evaluated on the basis of how effectively they implement the relevant models and discuss issues surrounding the application of machine learning to solve real-world problems.

SOURCES This course will be based on notes produced for the course by the instructor. Recommended secondary sources include *Python Data Science Handbook* by Jake VanderPlas and *Hands-on Machine Learning with Scikit-Learn and TensorFlow* by Aurélien Géron.

COURSE-RELATED WORK EXPECTATIONS Students will meet three hours per week in class (42 total hours), and homework and other assignments will take about seven hours per week (98 hours). The project will take about 28 hours, and final exam review will take 12 hours, for a total of 180 hours.

Calendar

2019-09-05	Course overview and admin
2019-09-10	Overview of ML
2019-09-12	Brief intro of software packages
2019-09-17	Data preprocessing, part 1, categorical and continuous features
2019-09-19	Data preprocessing, part 2, missing data
2019-09-24	Exploratory data analysis, part 1, EDA in regression and clustering
2019-09-26	Exploratory data analysis, part 2, EDA in classification
2019-10-01	Dimensionality reduction
2019-10-03	Feature engineering
2019-10-08	Evaluation metrics in supervised ML, part 1, hard predictions
2019-10-10	Evaluation metrics in supervised ML, part 2, soft predictions and regression metrics
2019-10-15	Supervised ML algorithms, part 1, Linear and Logistic regression
2019-10-17	Supervised ML algorithms, part 2, other ML algorithms
2019-10-22	Midterm presentations
2019-10-24	Midterm presentations
2019-10-29	ML pipelines, part 1, Cross-validation to evaluate model performance
2019-10-31	ML pipelines, part 2, hyper-parameter tuning
2019-11-05	Missing data revisited
2019-11-07	Interpretable ML
2019-11-12	Ethical issues in supervised ML
2019-11-14	Unsupervised ML
2019-11-19	Deployment and continuous monitoring
2019-11-21	Review
2019-11-26	Review
2019-12-03	Final presentations
2019-12-05	Final presentations