CBE 143 – Data Analytics for Chemical Engineers

Spring 2020

Instructors

Profs. Ali Mesbah, Karthik Shekhar, and David Graves

GSI

Alberto Nava (alberto_nava@berkeley.edu)

Description

Businesses, institutions, and individuals routinely create massive collections of data as a by-product of their activity. Data science is the study of generalizable approaches to translate such large-scale, multi-dimensional datasets into useful knowledge that can build insight and guide decisions. It involves the synthesis of mathematics, statistics, and computer science into a set of practical tools that can be applied to large datasets involving tens to tens of thousands of measured variables (features) arising in a variety of contexts. The aims of such analyses involve one or more of the following tasks: fitting parameters of a specific model using the data (inference), choosing the best model (model selection), learning patterns in the data (unsupervised learning), identifying groups (clustering) and/or identifying low-dimensional representations of the data (manifold learning). This course will introduce chemical engineering students to this rapidly growing field, and equip them with the basic foundations as well as practical tools to process, explore, integrate, model and visualize data. While students will be introduced to work with real world datasets arising in a variety of disciplines, particular emphasis will be laid on contextualizing students to applications in systems diagnosis and monitoring, as well as biological discovery, areas close to chemical engineering. In addition, we will work hands-on with the Python programming language and associated data analysis libraries.

Prerequisite knowledge, skills, and resources

- Familiarity with linear algebra, elementary probability and statistics, and multivariate calculus will be assumed. We will not teach these subjects in class, except reviewing concepts that are necessary.
- Some proficiency with computer programming (data structures; for and while loops; if/else statements, reading/writing file), preferably in Python. Please consult the instructors if you have doubts. Without basic fluency in Python programming, you are likely to struggle with the homeworks and the class project. An online introduction to the basics of Python is available at https://learnpython.org.
- **Important:** You must have access to a computer that can install software. Please install Miniconda (https://docs.conda.io/en/latest/miniconda.html), and Jupyter notebook (https://cs205uiuc.github.io/guidebook/resources/python-miniconda.html)
- A great online resource is Google Colaboratory (http://colab.research.google.com). It is a free Jupyter notebook environment that requires no setup, runs entirely on the cloud and is linked to your google drive. While the resources may not suffice for intensive computation, it can be a great starting tool to familiarize yourself with Jupyter notebooks and Python scripting.

Course Objectives and Outcomes– Students learn:

- Data analysis approaches in open source platforms such as Python and Jupyter notebooks
- How to read in, preprocess and explore large-scale datasets from different sources
- How to generate compelling visualizations of the data
- A variety of probability distributions used in data science, and using statistical packages to fit models to data, and assess the quality of fits

Outcomes – Students must be able to:

- Explain in basic terms what statistical inference and machine learning mean, and distinguish between unsupervised and supervised learning
- Access and read in a variety of publicly available datasets from a variety of sources (the web, MS excel files, csv files, text files, hdf5 files) into Python, and preprocess them into standard formats (e.g. lists, matrices, data tables) for further analysis
- Use Jupyter notebooks and python packages (numpy, scipy, scikit-learn, pandas, matplotlib)
- Perform exploratory data analysis. Apply basic tools (graphs and summary statistics) to data
- Identify probability distributions commonly used for statistical modeling. Fit a model to data and estimate model quality, and perform model comparisons and assess overfitting and generalizability
- Apply basic machine learning algorithms such as linear regression, logistic regression, k-nn. Explain the importance of covariates and regularization in inference
- Be able to compare the performance of alternative algorithms/approaches – e.g. linear regression with and without regularization, knn classifier vs. random -forests
- Explain the concept of Bayesian inference, and the importance of probabilistic models as a formal way to quantify uncertainty of model parameters. Difference between Maximum Likelihood (MLE) and Maximum A-Posteriori (MAP) estimates
- Develop empirical (e.g., data-driven) models from time series data
- Analyze correlated data using multivariate statistical methods

Tentative Course Outline

Background (5 lectures)

- Intro to course, Python and Jupyter notebooks
- Basics of probability and statistics (t-test, ANOVA)
- Exploratory data analysis: Handling tabular data using the Pandas package
- Exploratory data analysis: Summary Statistics, Covariance, Entropy and Mutual Information
- Exploratory data analysis: Visualization techniques
- Foundations of Statistical Inference

Regression (4 lectures)

- Linear Regression, Kernel trick, Regularization (Conceptual)
- Cross-validation, Model selection
- Model selection, A brief overview of nonlinear regression

Parameter estimation (2 lectures)

- Frequentist versus Bayesian approach to parameter estimation
- Maximum likelihood and maximum a-posteriori estimation

Classification (4 lectures)
- Introduction, Parametric classification using Logistic Regression
- Non-parametric, knn-classifier, decision-trees, random forests
- Training and cross-validation, evaluation, AUC
- Boosting and Bagging. Robustness (Advanced)

Dimensionality reduction (3 lectures)
- Concepts, Linear algebra, Matrix factorization
- PCA/SVD, Non-negative matrix factorization
- Non-linear dimensionality reduction: tSNE/UMAP/Diffusion Maps

Ethics in data science (1 lecture)

Advanced applications (6 lectures throughout the course)
- Statistical process monitoring, fault diagnosis, and predictive maintenance (PCA, PLS, and Fisher discriminant analysis)
- Parameter estimation (MLE and MAP)
- Empirical modeling (Subspace identification and Recurrent Neural Networks)
- Single-cell biology
- Learning Regulatory interactions
- Computational chemistry

Guest lecturers (2 lectures)

Class schedule

Lectures: Tuesday and Thursday 12:30 – 2:00 pm, 425 Latimer Hall
Lab Section: Friday 2:00 – 4:00 pm, Hildebrand B56

Office hours

Ali Mesbah: Tuesdays 4:00 – 5:00 pm, 316 Gilman Hall
Karthik Shekhar: Thursdays 4:00 – 5:00 pm, 201F Gilman Hall

Grading

Five homework assignments (40%)
Project + report (40%)
Peer review an article (20%)

All coursework should be submitted on Gradescope
piazza will be used as the online Q&A platform

Useful sources and references (not required to purchase textbooks)

G. James, D. Witten, T. Hastie, and R. Tibshirani. An introduction to statistical learning: with Applications to R. Springer, 2017.

C. Bishop. Pattern recognition and machine learning. Springer, 2006.