Data Science

Course Outline (Spring 2017)

Course Description

Data Science is the study of the generalizable extraction of knowledge from data. Being a data scientist requires an integrated skill set spanning computer science, mathematics, statistics, and domain expertise along with a good understanding of the art of problem formulation to engineer effective solutions. The goal of this course is to teach students to answer questions with data. To do this, we will learn the necessary skills to manage and analyze data with case studies. In this course student learn concepts such as data collection and integration, exploratory data analysis, statistical inference and modeling, machine learning, and high-dimensional data analysis.

Prerequisite

Programming competence, Discrete Maths, Linear Algebra, Probabilty& Statistics

Books

There is no standard one "textbook" for this course. The following book will be used as a primary text to guide much of the discussions, but it will be heavily supplemented with lecture notes and reading assignments from other sources.

Cathy O'Neil and Rachel Schutt. Doing Data Science, Straight Talk From The Frontline. O'Reilly. 2014. ISBN 978-1-449-35865-5.

Additional references and books related to the course:

Jure Leskovek, Anand Rajaraman and Jeffrey Ullman. Mining of Massive Datasets. v2.1, Cambridge University Press. 2014. (Free online.)

Jiawei Han, Micheline Kamber and Jian Pei. Data Mining: Concepts and Techniques, Third Edition. Morgan Kaufmann Publishers. 2012. ISBN 978-0-12-381479-1.

Kevin P. Murphy. Machine Learning: A Probabilistic Perspective. MIT Press. 2013. ISBN 0262018020. (Online info available here.)

Foster Provost and Tom Fawcett. Data Science for Business: What You Need to Know about Data Mining and Data-analytic Thinking. O'Reilly 2013. ISBN 978-1-449-36132-7.

Outline

1. Introduction: What is Data Science?
   - Big Data and Data Science hype -- and getting past the hype
   - Current landscape of perspectives
   - Skill sets needed
2. Statistical Inference and (Python or R)
   - Populations and samples
   - Statistical modeling, probability distributions, fitting a model
   - Intro to Python/R
3. Exploratory Data Analysis and the Data Science Process
   - Basic tools (plots, graphs and summary statistics) of EDA
   - Philosophy of EDA
   - The Data Science Process
4. Three Basic Machine Learning Algorithms
   - Linear Regression
   - k-Nearest Neighbors (k-NN)
   - k-means
5. One More Machine Learning Algorithm and Usage in Applications
   - Motivating application: Filtering Spam
   - Why Linear Regression and k-NN are poor choices for Filtering Spam
   - Naive Bayes and why it works for Filtering Spam
6. Data Wrangling
   - Data cleaning, data resahping, data integration
7. Feature Generation
   - Motivating application: user (customer) retention
   - Feature Generation (brainstorming, role of domain expertise, and place for imagination)
8. Feature Selection
   - Filters; Wrappers
   - Decision Trees; Random Forests
9. Recommendation Systems
   - Algorithmic ingredients of a Recommendation Engine
   - Dimensionality Reduction
   - Singular Value Decomposition
10. Data Visualization
    - Basic principles, ideas and tools for data visualization
    - Examples of inspiring (industry) projects
    - Exercise: create your own visualization of a complex dataset
11. Design of Experiments
    - Factorial Designs, Latin Hypercube Designs etc.

Grading

Assignments/Quiz(s)/Homeworks	20 - 25 %
Project (Implementation & Presentation) / Research Paper	*10* %
Midterms	25 - 30 %
Final Exam	40 - 45%
Total:	**100 %**