CS481 - Data Science

Course Outline (Spring 2018)

Instructor Name: Irfan Younas
Email address: irfan.younas@nu.edu.pk
Office Location/Number: Civil building, C-146
Office Hours: Mon 10-12 AM WED, FRI 9 -11 AM

Course Information

Program: BS/MS **Credit Hours:** 3 **Type:** Core/Elective
Course Website: piazza.com/fast_lahore/spring2018/cs481
Class Meeting Time: Section A: Wed, Fri 2 – 3:30 PM Section B: Wed, Fri 11 AM – 12:30 PM
Class Venue:

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)
    - Populations and samples
    - Statistical modeling, probability distributions, fitting a model
    - Intro to Python
3. Data Wrangling
    - Data cleaning, data reshaping, data integration
4. Exploratory Data Analysis and the Data Science Process
    - Basic tools (plots, graphs and summary statistics) of EDA
    - Philosophy of EDA
    - The Data Science Process
    - Relationship between variables, varibles data distribution
    - Residual plots and Model improvement
    - Data transformation
5. Machine Learning Algorithms
    - Linear Regression
    - Logistic Regression
    - Regularization
    - Advice for apply ML and ML System Design
    - Support Vector Machines
    - Dimensionality Reduction (Principal Component Analysis)
    - Clustering Algorithms

6. Data Visualization
    - Basic principles, ideas and tools for data visualization
    - Examples of inspiring (industry) projects
    - Exercise: create your own visualization of a complex dataset

Grading

Quiz(s)	10 %	
Assignments, Homeworks and Project		*15* %
Midterms	30 %	
Final Exam	45 %	
Total:	**100 %**	