Introduction to Data Science: CptS 483-06 – Syllabus
First Offering: Fall 2015

Course Information

Credit Hours: 3
Semester: Fall 2015
Meeting times and location: MWF, 12:10–13:00, Sloan 163
Course website: `http://www.eecs.wsu.edu/~assefaw/CptS483-06/`
Relevant course material, including this syllabus, and course related resources will be made available at the course website. Additionally, the online portal OSBLE (`https://osble.org`) will be used for posting lecture material, assignments, announcements, etc and for handling submissions.

Instructor Information

Assefaw Gebremedhin
Office: EME 59
Email: assefaw AT eecs DOT wsu DOT edu
Homepage: `www.eecs.wsu.edu/~assefaw`
Office Hours: Tuesdays 2:00–3:00pm, or by appointment.

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 mathematics, statistics, machine learning, databases and other branches of computer science along with a good understanding of the craft of problem formulation to engineer effective solutions. This course will introduce students to this rapidly growing field and equip them with some of its basic principles and tools as well as its general mindset. Students will learn concepts, techniques and tools they need to deal with various facets of data science practice, including data collection and integration, exploratory data analysis, predictive modeling, descriptive modeling, data product creation, evaluation, and effective communication. The focus in the treatment of these topics will be on breadth, rather than depth, and emphasis will be placed on integration and synthesis of concepts and their application to solving problems. To make the learning contextual, real datasets from a variety of disciplines will be used.

Learning Outcomes

At the conclusion of the course, students should be able to:

- Describe what Data Science is and the skill sets needed to be a data scientist.
- Explain in basic terms what Statistical Inference means. Identify probability distributions commonly used as foundations for statistical modeling. Fit a model to data.
- Use R to carry out basic statistical modeling and analysis.
- Explain the significance of exploratory data analysis (EDA) in data science. Apply basic tools (plots, graphs, summary statistics) to carry out EDA.
- Describe the Data Science Process and how its components interact.
- Use APIs and other tools to scrap the Web and collect data.
- Apply EDA and the Data Science process in a case study.

- Apply basic machine learning algorithms (Linear Regression, k-Nearest Neighbors (k-NN), k-means, Naive Bayes) for predictive modeling. Explain why Linear Regression and k-NN are poor choices for Filtering Spam. Explain why Naive Bayes is a better alternative.
- Identify common approaches used for Feature Generation. Identify basic Feature Selection algorithms (Filters, Wrappers, Decision Trees, Random Forests) and use in applications.
- Identify and explain fundamental mathematical and algorithmic ingredients that constitute a Recommendation Engine (dimensionality reduction, singular value decomposition, principal component analysis). Build their own recommendation system using existing components.
- Create effective visualization of given data (to communicate or persuade).
- Work effectively (and synergically) in teams on data science projects.
- Reason around ethical and privacy issues in data science conduct and apply ethical practices.

Audience

The course is suitable for upper-level undergraduate (or graduate) students in computer science, computer engineering, electrical engineering, applied mathematics, business, computational sciences, and related analytic fields.

Prerequisites

Students are expected to have basic knowledge of algorithms and reasonable programming experience (equivalent to completing a data structures course such as CptS 223), and some familiarity with basic linear algebra (e.g. solution of linear systems and eigenvalue/vector computation) and basic probability and statistics. *If you are interested in taking the course, but are not sure if you have the right background, talk to the instructor. You may still be allowed to take the course if you are willing to put in the extra effort to fill in any gaps.*

Course work

The course consists of lectures (three times a week, 50 min each), and involves a set of assignments (about 3 or 4) and a project. A project could take one of several forms: analyzing an interesting dataset using existing methods and software tools; building your own data product; or creating a visualization of a complex dataset. Students are encouraged to work in teams of two or three for a project. Assignments, on the other hand, are to be completed and submitted individually. Besides the assignments and a project, there will be frequent opportunities for in-class exercises and “thought experiments”.

Grading

Your final grade will be determined based on your performance on each of the following items; the percentages in parenthesis show the weight each item carries to the final grade.

- Class participation (10%)
- Assignments (30%)
- Project (30%)
- Final exam (30%)

Letter grades: A (93–100%), A- (90–92.99%), B+ (87–89.99%), B (83–86.99%), B- (80–82.99%), C+ (77–79.99%), C (70–76.99%), C- (67–69.99%), D (60–66.99%), F (less than 60%). Grading scale may be adjusted depending on class average.

Topics and course outline:

1. Introduction: What is Data Science?
    - Big Data and Data Science hype – and getting past the hype
    - Why now? – Datafication
    - Current landscape of perspectives
    - Skill sets needed
2. Statistical Inference
    - Populations and samples
    - Statistical modeling, probability distributions, fitting a model
    - Intro to 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
    - Case Study: RealDirect (online real estate firm)
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
    - Data Wrangling: APIs and other tools for scrapping the Web
6. Feature Generation and Feature Selection (Extracting Meaning From Data)
    - Motivating application: user (customer) retention
    - Feature Generation (brainstorming, role of domain expertise, and place for imagination)
    - Feature Selection algorithms
        - Filters; Wrappers; Decision Trees; Random Forests
7. Recommendation Systems: Building a User-Facing Data Product
    - Algorithmic ingredients of a Recommendation Engine
    - Dimensionality Reduction
    - Singular Value Decomposition
    - Principal Component Analysis
    - Exercise: build your own recommendation system
8. Mining Social-Network Graphs
    - Social networks as graphs
    - Clustering of graphs
    - Direct discovery of communities in graphs
    - Partitioning of graphs
    - Neighborhood properties in graphs
9. Data Visualization
    - Basic principles, ideas and tools for data visualization

- Examples of inspiring (industry) projects
- Exercise: create your own visualization of a complex dataset

10. Data Science and Ethical Issues
    - Discussions on privacy, security, ethics
    - A look back at Data Science
    - Next-generation data scientists

Books

The following book will be used as a textbook and primary resource to guide the discussions, but will be heavily supplemented with lecture notes and reading assignments from other sources. The lecture notes and reading material will be posted on the course's website or the associated OSBLE page as the course proceeds.

- Cathy O'Neil and Rachel Schutt. Doing Data Science, Straight Talk From The Frontline. O'Reilly. 2014.

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)
- Kevin P. Murphy. Machine Learning: A Probabilistic Perspective. ISBN 0262018020. 2013.
- Foster Provost and Tom Fawcett. Data Science for Business: What You Need to Know about Data Mining and Data-analytic Thinking. ISBN 1449361323. 2013.
- Trevor Hastie, Robert Tibshirani and Jerome Friedman. Elements of Statistical Learning, Second Edition. ISBN 0387952845. 2009. (free online)
- Avrim Blum, John Hopcroft and Ravindran Kannan. Foundations of Data Science.
  (Note: this is a book currently being written by the three authors. The authors have made the first draft of their notes for the book available online. The material is intended for a modern theoretical course in computer science.)
- Mohammed J. Zaki and Wagner Miera Jr. Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press. 2014.
- Jiawei Han, Micheline Kamber and Jian Pei. Data Mining: Concepts and Techniques, Third Edition. ISBN 0123814790. 2011.

Policies

Missing or late work

Submissions will be handled via the OSBLE page of the course. Students are expected to submit assignments by the specified due date and time. Assignments turned in up to 48 hours late will be accepted with a 10% grade penalty per 24 hours late. Except by prior arrangement, missing or work late by more than 48 hours will be counted as a zero.

Academic Integrity

Academic integrity will be strongly enforced in this course. Any student who violates the University's standard of conduct relating to academic integrity will receive an F as a final grade in this course, will not have the option to withdraw from the course and will be reported to the Office of Student Standards and Accountability. Cheating is defined in the Standards for Student Conduct WAC 504-26-010 (3). You can learn more about Academic Integrity on the WSU campus at `http://conduct.wsu.edu`. Please also read this link carefully: EECS Academic Integrity Policy (`http://www.eecs.wsu.edu/~schneidj/Misc/academic-integrity.html`). Use these resources to ensure that you do not inadvertently violate WSU's standard of conduct.

Safety on Campus

Washington State University is committed to enhancing the safety of the students, faculty, staff, and visitors. It is highly recommended that you review the Campus Safety Plan (`http://safetyplan.wsu.edu/`) and visit the Office of Emergency Management web site (`http://oem.wsu.edu/`) for a comprehensive listing of university policies, procedures, statistics, and information related to campus safety, emergency management, and the health and welfare of the campus community.

Students with Disabilities

Reasonable accommodations are available for students with a documented disability. If you have a disability and need accommodations to fully participate in this class, please either visit or call the Access Center (Washington Building 217; 509-335-3417) to schedule an appointment with an Access Advisor. All accommodations MUST be approved through the Access Center. For more information, consult the webpage `http://accesscenter.wsu.edu` or email at `Access.Center@wsu.edu`.

Important Dates and Deadlines

Students are encouraged to refer to the academic calendar often to be aware of critical deadlines throughout the semester. The academic calendar can be found at `www.registrar.wsu.edu/Registrar/Apps/AcadCal.ASPX`.

Weather Policy

For emergency weather closure policy, consult: `http://alert.wsu.edu`.

Changes

This syllabus is subject to change. Updates will be posted on the course website.