Syllabus for Data Mining CS363D

Adam Klivans

Spring 2016

1 Course Overview

Using programs to automatically find structure in complex data sets has become fundamental in science and industry. This course will give introductory techniques for building programs that can model data. Topics include classification, regression, clustering, ensemble, and Bayesian methods. Prerequisites: knowledge of Linear Algebra and Probability.

2 Instructors

Instructor: Adam Klivans (office hours: Monday 1–2:00pm in GDC 4.826). TA: Xueyu Mao (office hours: Tuesday 3-4:30pm GDC 1.302 Desk 1 of TA stations).

3 Classroom

PAR 203 (MW 9:30–11)

4 Syllabus

TSK refers to the text "Introduction to Data Mining," by P. Tan, M. Steinbach, and V. Kumar. JWHT refers to "An Introduction to Statistical Learning in R."

- Overview and Preliminaries on Working with Data (1 week) [TSK Chap. 2, JWHT Chap. 1]
- Probability and Linear Algebra Review (1-2 weeks interspersed) [TSK Appendices A and C]
- Classification: Decision Trees (1 week) [TSK Chapter 4]
- Classification Issues: Overfit, Cross-Validation (1 week) [TSK Chapter 4]
- Classification: Nearest Neighbor (1 lecture) [TSK Chapter 5.2]
- Classification: Naive Bayes (1 lecture) [TSK Chapter 5.3]
- Regression (1-2 weeks): [TSK Appendix D, JWHT Chapter 3]
- Clustering (1-2 weeks): [TSK Chapter 8, JWHT Chapter 10]

- Ensemble Methods (1-2 weeks): [TSK 5.6, JWHT Chapter 8]
- Advanced Methods (Kernel Methods, Online Learning, Neural Networks) (time permitting)

The syllabus is subject to change depending on students' background and interests.

5 Assignment, Assessment, Evaluation

- There will be six homeworks (50%), one midterm (20%), and one final (30%).
- Homework is due by the beginning of class. Students are encouraged to use Latex to typeset their solutions and email them to the instructor. You may have two free "late days" that you can use all on one homework or on two homeworks. Homework that is one day late is worth 50%. Homework is worth 0% if turned in more than one day late.
- Some homeworks will have programming assignments. We will be programming in Python and use the scikit-learn package. You may work in groups of two for any programming assignment, clearly indicating who you collaborated with. Each non-programming assignment must be written up individually. You may discuss a non-programming assignment with at most two other students in the class (or use Piazza). YOU MUST WRITE UP YOUR OWN SOLUTIONS BY YOURSELF.
- You may not under any circumstances search for a solution to homework on the internet, in a book or from any other resource. Each student in this course is expected to abide by the University of Texas Honor Code.

6 Other University Notices and Policies

6.1 Documented Disability Statement

Any student with a documented disability who requires academic accommodations should contact Services for Students with Disabilities (SSD) at (512) 471-6259 (voice) or 1-866-329-3986 (video phone). Faculty are not required to provide accommodations without an official accommodation letter from SSD. Please notify me as quickly as possible if the material being presented in class is not accessible (e.g., instructional videos need captioning, course packets are not readable for proper alternative text conversion, etc.). Contact Services for Students with Disabilities at 471-6259 (voice) or 1-866-329-3986 (video phone) or reference SSDs website for more disability-related information: `http://www.utexas.edu/diversity/ddce/ssd/for_cstudents.php`

6.2 Behavior Concerns Advice Line

If you are worried about someone who is acting differently, you may use the Behavior Concerns Advice Line to discuss by phone your concerns about another individuals behavior. This service is provided through a partnership among the Office of the Dean of Students, the Counseling and Mental Health Center (CMHC), the Employee Assistance Program (EAP), and The University of Texas Police Department (UTPD). Call 512-232-5050 or visit `http://www.utexas.edu/safety/bcal`.