CSC 566: Advanced Data Mining
Spring 2017
Course Syllabus

April 3, 2017

Instructor: Alexander Dekhtyar
email: dekhtyar@calpoly.edu
office: 14-215

What	When	Where
Lecture	MW 4:10 – 6:00pm	14-257

Note: the class will not have a written final exam, but the exam time will be used for student presentations.

Office Hours

	When	Where
Wednesday	8:10am - 10:00am	14-210
Friday	8:10am - 10:00am	14-210

Additional appointments can be scheduled by emailing the instructor at *dekhtyar@calpoly.edu*.

Description

The course is designed to cover in depth a variety of mathematically complex methods and techniques for data mining and machine learning.

Texbook

There isn't a single required textbook for the class. The material will come from a number of existing textbooks, academic papers, and a few other sources.

Some of the books used in the class are

- Charu C. Aggrawal, *Data Mining: The Handbook*, Springer, 2015, ISBN 978-3-319-14141-1.
- Mohammed J. Zaki, Wagner Meira Jr., *Data Mining And Analysis: Fundamental Concepts and Algorithms*, Cambridge University Press, 2014, ISBN: 978-0-521-76633-3.
- Charu C. Aggrawal, *Recommender Systems: The Textbook*, Springer, 2016, USBN: 978-3-319-29657-9.
- Ian Goodfellow, Yoshua Bengio, Aaron Courville, *Deep Learning*, MIT Press, 2016, ISBN: 9780262035613.
- Jure Leskovec, Anand Rjarman, Jeffrey Ullman, *Mining Massive Datasets*, 2014.
- Richard Duda, Peter Hart, David Stork, *Pattern Classification*, Wiley, 2001, ISBN: 0-471-05669-3.
- Amy Langville, Carl Meyer, *Google's PageRank and Beyond: The Science of Search Engine Rankings*, Princton University Press, 2006, ISBN: 978-0-691-12202-1.

In addition, we will be reading a number of papers.

Topics

We will cover the following topics in the course:

1. Numerical Analysis:
    - Function approximation
    - Gradient descent
2. Regression:
    - Linear Regression
    - Logistic Regression
3. Classification
    - Perceptron
    - Support Vector Machines
    - Neural Networks
4. Dimensionality Reduction Techniques
    - Singular-Value Decomposition of Matrices
    - Principal Components Analysis
    - Latent Semantic Indexing
5. Mathematical Foundations of PageRank

Time permitting, we will talk about a few more topics, including recommender systems, clustering, generative methods for text mining, Gibbs sampling, deep learning and feature embeddings.

The key component of this class is that we choose to study methods and techniques that are based on mathematics that is more complex that what we could afford to cover in CSC 466. Where CSC 466 was broad and shallow, CSC 566 is expected to be deep and narrow. We may not be able to finish discussing all the topics on the list above, but those we do study, will be covered in detail.

Grading

Assignments	50-60%
Project	20-30%
Reports	10-20%

Course Policies

Prerequisites

This class naturally extends CSC 466 and therefore is most appropriate for those who already took CSC 466. Students who took CSC 582 have also seen some of the background and therefore, it is a good prerequisite course. DATA 401 is another course that is considered a valid prerequisite.

All cases will be dealt with on individual basis.

Exams

The course has **no exams.** Instead, there is a quarter-long team project with a number of deliverables (proposal, presentation, report), which accounts for 20-30% of the course grade. The details of the course project will be released some time in the first two-three weeks of the class.

The final exam time (Friday, June 16, 4:10-7:00pm) is reserved for project presentations. Given how late it is in the finals exam week, we may use the last day of classes for the presentations instead.

Reports

It is customary for students taking graduate classes to read papers and present them in class. It is possible that some students will get presentation assignments as we go through the quarter, however, I do not expect having every person present in the class.

However, I want every student to have read a number of seminal and cutting edge papers on the topic of data mining. Some time during the first three weeks of classes, I will put together the reading list for everyone. Every student in the class will bid on papers, and write reports discussing the papers and reviewing them. There will be multiple individual writing assignments in the class.

Programming Assignments

The bulk of the coursework starting week 2 of the class will be programming assignments. You will learn how to implement specific methods and techniques from scratch, and you will run your methods on a battery of datasets and will evaluate the accuracy and the performance of your implementations.

The goal, by the end of the quarter is to have a comparative study of a variety of machine learning techniques that looks at the accuracy vs. performance trade-offs and identifies techniques that work better on different types of problems.

Communication

The class will have an official mailing list. The email address for the mailing list is *csc-566-01-2174@calpoly.edu*. All students enrolled in the class are automatically subscribed to the mailing list (using the email addresss that the CS department has on file).

I encourage questions during classtime and questions via email. My answers to email questions may be broadcast to the entire class via the mailing list, if the answer may be relevant to everyone (e.g. a correction in a text of a handout, or a clarification of a homework problem), and may also appear on the web page. The questions can also be posted to the mailing list directly. The mailing list will also be used for all annoucements related to the course. It is your responsibility to read your class-related email. Failure to read email posted to the mailing list cannot be used as an excuse in the class.

Web Page

Class web page can be found at

http://www.csc.calpoly.edu/∼dekhtyar/566-Spring2017

Through this page you will be able to access all class handouts including homeworks, lab assignments, project information, lab/project data and lecture notes.

Links to additional information, and notes and announcements will also be posted.

Academic Integrity

University Policies

Cal Poly's Academic Integrity policies are found at

http://www.academicprograms.calpoly.edu/academicpolicies/Cheating.htm

In particular, these policies define *cheating* as (684.1)

> *"...obtaining or attempting to obtain, or aiding another to obtain credit for work, or any improvement in evaluation of performance, by any dishonest or deceptive means. Cheating includes, but is not limited to: lying; copying from another's test or examination; discussion*

> *of answers or questions on an examination or test, unless such discussion is specifically authorized by the instructor; taking or receiving copies of an exam without the permission of the instructor; using or displaying notes, "cheat sheets," or other information devices inappropriate to the prescribed test conditions; allowing someone other than the officially enrolled student to represent same.*"

Plagiarism, per University policies is defined as (684.3)

> "... the act of using the ideas or work of another person or persons as if they were one's own without giving proper credit to the source. Such an act is not plagiarism if it is ascertained that the ideas were arrived through independent reasoning or logic or where the thought or idea is common knowledge. Acknowledgement of an original author or source must be made through appropriate references; i.e., quotation marks, footnotes, or commentary."

University policies state (684.2): "Cheating requires an "F" course grade and further attendance in the course is prohibited." (appeal process is also outlined, see the web site above for details.). Plagiarism, per university policies (684.4) can be treated as a form of cheating, although a level of discretion is given to the instructor, allowing the instuctor to determine the causes of plagiarism and effect other means of remedy. It is the obligation of the instructor to inform the student that a penalty is being assessed in such cases.

Course Policies

All homeworks are to be completed by each student **individually**. Lab assignments are to be completed by the appropriate units (individual, pair, group), and no code/solution-sharing between units is permitted. Students are encouraged to discuss class content among themselves but NOT in a manner that constitutes plagiarism and cheating as defined above (e.g., you can solve together a problem from the textbook that had not been assigned in the homework, but you should solve assigned problems individually).