Math 166 / CSCI 145 - Data Mining

Claremont McKenna College, Department of Mathematical Sciences

Blake Hunter - Spring 2015

Course Information

Instructor : Blake Hunter	**Office :** Adams 212
Email : bhunter@cmc.edu	**Office Hours :** TBA
Lectures : TBA from TBA to TBA in TBA	**Course Website :** Sakai

Text Book

The Elements of Statistical Learning: Data Mining, Inference, and Prediction 2nd Ed., Hastie, Tibshirani, and Friedman.
Matlab a practical introduction to programming and problem solving 3rd Ed., Stormy Attaway.

Course Description

Data mining is the process for discovering patterns in big data using techniques from mathematics, computer science and statistics with applications ranging from biology and neuroscience to history and economics. The goal of the course is to teach students fundamental data mining techniques that are commonly used in practice. Students will learn advanced data mining techniques (including linear classifiers, support vector machines, clustering, dimension reduction, transductive learning and topic modeling). The course is designed to be applicable and fulfilling for both strongly motivated students who have taken Linear Algebra and advanced mathematics or computer science majors. **Prerequisite:** MATH 60 CM (Linear Algebra); and a proof based course above MATH 100 or CSCI 62 CM (Data Structures and Advanced Programming); or consent of the instructor.

Note : Familiarity with a computer programming language (C++, Java, Matlab, Python or R) or CSCI 51 CM (Intro to Computer Science) is suggested but not required.

Grading

Homework - 20%
Midterms (2) - 40%
Final Project - 10%
Final Exam - 30%

Grading Scale

100-90% **A±,** 89-80% **B±,** 79-70% **C±**, 69-60% **D±**, else **F**
Grades will only be curved up at the end of the course, if need be, to insure everyone above the median student's grade will receive C or better.
You can check your point totals online at sakai.

Tentative Weekly Schedule

1. Overview, programming in Matlab (variables, logic, loops, functions, vectors, matrices, class/objects...),
    a. (optional) parallel computing in Matlab, C++ and CUDA
2. Linear algebra review (vectors, matrices, eigenvalues/vectors, norms, inner products, Matlab matrix operations,...), mathematical logic/proofs review
3. Supervised Learning, regression, linear classifiers
4. Support vector machines (SVM)
5. kernel SVM, soft SVM, **Midterm 1**
6. Clustering and k nearest neighbor
7. *k*-means and spectral clustering
8. Dimension reduction - singular value decomposition (SVD)
9. Image processing, image features. object recognition
10. Unsupervised/supervised image segmentation, **Midterm 2**
11. Image classification, Computer vision
12. Principal component analysis (PCA)
13. Topic modeling and non-negative matrix factorization
14. More on topic modeling / natural language processing
15. Makeup, (if time allows) Transductive learning, Diffusion, motion by mean curvature
16. Student presentations
17. **Final Exam**

Due Dates	Exam Dates
week 5 - select team	week 5 - midterm 1
week 6 - select a data set	
week 7 - problem statement	
week 10 - progress report	week 10 - midterm 2
week 13 - update report	
week 16 - presentation slides/poster	
	week 17 - final

Exams

There will be **two midterm** in class exams on **Friday** of **week 5** and **Friday** of **week 10** each accounting for 20% of your course grade. The **final exam** is scheduled for **TBA** in class and will cover all the material we covered in the course accounting for 30% of your course grade. The exams will be given during our regular lecture time. If you are unable to take the exam at the scheduled time, you must contact the professor **before** the exam date and will be asked to provide documentation for your absence.

Homework

There will be 8 bi-weekly homework assignments. The homework will be made of three parts including

1. a math component
2. a computational component
3. an advanced theory component. You can choose between
    - 3.1. a theoretical math option (requiring MATH 100+)
    - 3.2. **or** advanced computer science option (requiring CSCI 62).

Homework assignments are due in class two weeks from the date assigned. The computational assignments can be done in any programming language but Matlab, C++, Java or Python are the only support languages for the provided supporting code, examples, example code and pre-formatted data for the course.

Final Projects

Each student will work in a group of 3-5 students, on a final project related to the material covered in the course. The goal of the final project is to have each student demonstrate their knowledge of the material by applying and extending the algorithms presented in class to an application of their choice (in consultation with the course instructors). Each group will submit a written proposal, progress report, final report and deliver a 20 minute oral presentation of their results with poster/sides. Students are expected to meet regularly with the course instructor to discuss project development.

Final Project Topics

Enron email - email text mining and time series analysis
Yahoo finance - stock and market analysis, clustering, visualization, auto-summary
Hyperspectral Imaging - pixel classification and object detection
Audio - voice command classification (OK Google / Siri)
Enron Social Network - Email network, social network analysis
MRI brain images - image noise statistical analysis and reduction
Video - Scene detection and object recognition
LAPD field interview cards - community detection
Robot survey data - object classification (xBox kinect data)
Google election polling data - trend detection and topic model fitting
data acquired from outside the course - from other classes, labs, internet or industry

Academic Honesty

You are allowed to discuss homework assignments with your fellow classmates in general terms, but do not share your homework solutions with other students. All homework assignments should be written up independent of math programs, the internet and other students. Sharing your homework or copying someone else's homework may result in a zero on the assignment and further discipline. If you have specific questions ask the professor or tutor. **Collaboration of any kind on the exams is strictly forbidden** and will result in a failing grade for the exam and academic discipline. Please consult the documents related to Academic Integrity,Statement of Academic Integrity for details.

Special Needs

It is recommended that Disability Support Services students contact the professor as soon as possible to discuss and make any special arrangements.

The syllabus is subject to change. See the sakai website for any updates, edits or changes.