CSE 5334: 002 Data Mining Spring 2014

Course Description and Goals:
This course is designed to teach you, at the graduate level, data mining algorithms for analyzing very large amounts of data. The topics include data extraction, exploratory data analysis, visualization, classification, clustering, frequent itemsets, search engine basics, recommender systems, dimensionality reduction etc. At the end of the semester you should:

- have a solid understanding of the basic concepts, principles, and techniques in data mining
- be familiar with most of the classical data mining algorithms
- be able to perform systematic analysis of real world data mining problems end to end
- be able to model data mining problems and evaluate, visualize and communicate statistical models

Books:
We cover a wide variety of topics that are not covered in a single book. Almost all the topics covered in the class will be found in one of the text books given below. If they are not, the instructor will point to the appropriate resources and also provide the slides. The mnemonics MMDS, ISLR and IIR will be used refer to these books in the readings.

- **[MMDS]** *Mining of Massive Datasets* by Jure Leskovec, Anand Rajaraman, Jeff Ullman. Free eBook can be found here.
- **[DMA]** *Data Mining and Analysis: Fundamental Concepts and Algorithms* by Mohammed Zaki and Wagner Meira. Free eBook can be found here.
- **[ISLR]** *An Introduction to Statistical Learning with Applications in R* by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. Free eBook can be found here.
- **[IIR]** *Introduction to Information Retrieval* by Christopher D. Manning, Prabhakar Raghavan and Hinrich Schutze. Free eBook can be found here.

In addition, there are some "classic" data mining books that are great references. These books are optional but great to have anyway!

- *Data Mining: Concepts and Techniques*, 3rd ed. by Jiawei Han, Micheline Kamber and Jian Pei.
- *Introduction to Data Mining* by Pang-Ning Tan, Michael Steinbach, and Vipin Kumar.
- *Pattern Recognition and Machine Learning* by Christopher M. Bishop.

Topics Covered:

The tentative list of topics to be covered in the class include:

- Data Science process: Extraction, Exploration, Visualization and Communication
- Classification models such as Decision trees, kNN, Naive Bayes, and others
- Ensemble learning such as Random Forests and AdaBoost
- Clustering algorithms such as k-Means, k-medoids, k-center, hierarchical clustering

•Frequent itemsets and Apriori algorithm
•Search Engine basics
•Recommender systems
•Dimensionality reduction techniques such as PCA, LSH and MDS
•Practical techniques such as sampling, feature selection, model testing, hypothesis testing etc

Grading:
Your grade will be based on the following weights:

• **Programming Projects: 30%**
- There will be 5 programming projects where you will implement some of the key ideas covered in the class
- The coding will be in Python and will done on IPython notebooks
- You can form team of 1-3 members. It is recommended to form teams as some of the projects could be challenging

•**Capstone Project: 10%**
- There will be a single team based capstone project at the end of the course
- This project could be used to demonstrate mastery of the topics covered by the course
- The instructor will provide a sample project. You can also create a equivalent project after consulting with the instructor.
- The hope is that this project (along with some of programming projects) will be part of the portfolio that you could share with your employers during interviews.

•**Midterm: 30%**
- There will one midterm exam during the semester covering first half of the course.
-There will be no make up exams except under extenuating circumstances!

•**Final : 30%**
- There will one non-comprehensive final exam during the semester covering second half of the course.
- There will be no make up exams except under extenuating circumstances!

Make-ups: Make-ups for graded activities may be arranged if your absence is caused by illness or work/personal emergency. A written explanation (including supporting documentation) must be submitted to the Instructor. If the explanation is acceptable, an alternative to the graded activity will be arranged. Make-up arrangements must be arranged *prior* to the scheduled due date.

Late days: Each student has **5** late days to use at his or her discretion for the problem sets and programming assignments. One cannot use more than **2** days for one assignment or problem set without prior approval from the instructors. Please reserve your late days for legitimate emergencies. Each late day constitutes a 24-hour extension; you cannot split late days into smaller increments. If two partners on a programming exercise want to take a late day, they must contribute two days from their allotment; either one from each of them, or, with permission, two days from the person with extra

late days.

Late penalties: Once a student runs out of late days, any late submissions are penalized at a rate of **50%** per day. No assignment may be handed in more than **2** days late.

Prerequisites:

- Algorithms & Data Structures (CSE 2320) or equivalent
- Basic knowledge of databases, linear algebra, probability and statistics
- Prior knowledge of Python (and packages such as Pandas, Scikit-learn, matplotlib) is optional but would be helpful in the course. If not, we will spend first few days brushing up these topics.