CSE514 Fall 2017 - Datamining
Administrative Information and Syllabus

Instructor: Professor Weixiong Zhang; email: zhang@cse.wustl.edu; homepage: `www.cse.wustl.edu/~zhang`; office: Jolley 530.

Prior knowledge required: The knowledge of the following topics is essential to do well in this course: Calculus, Probability, Statistics, Data Structure and Algorithms. Mastering one programming language (e.g., Python, MatLab, C/C++, Jave, and R) is also required.

Main objectives to achieve: You should expect to walk away from the course the knowledge of the basic concepts and methods of DM (and machine learning), as well as some working knowledge on how a data analysis problem should be approached, i.e., knowledge on what data processing and analysis methods should be adopted in order to solve the problem.

My way of teaching and how you should take the course: I will use an old school fashion of teaching, i.e., a type of chalk talk to talk through a subject and write/draw on board the key concepts and ideas, and work through examples. I will also ask questions along the way to inspire you to think and participate.

I will release my lecture notes for the topics to be discussed in class. You are not required to review the materials - in fact, I strongly *discourage* you to preview my lecture notes. Instead, you are *strongly encouraged* to bring a copy of the lecture notes to the class so that you may add your own notes to make a complete set of notes for your review. *You are required to* **review** *the materials we cover in class.* I want to emphasize that the lecture notes that I provide serve primarily as a guidance of my lectures and help you take notes. As such the original, incomplete lecture notes may not be an ideal resource for independent study. While class attendance is not mandatory,

you are responsible for all of the content we discuss in class. *We will have in-class quizzes based on the previous lectures*, which contribute to your final grade. This implies that you are better off to attend the class and participate.

We will also use an online forum on piazza for off-hour online learning. You will receive an invitation from the system to join the forum.

Textbook, reference and software tool kit: There is no required textbook for this course mainly because the topics we cover are not in a single book. Nevertheless, there are many online materials that we can use for additional reading. I will post the links to relevant topics to help leaning. Although you may choose not to read these online materials, some of the concepts and problems in homeworks/exams may appear in the online materials.

We will use a datamining tool kit called WEKA - google to find it online - and the following book has a good coverage of the main tools in the tool kit, so you may want to get a copy of it:

Ian H. Wkitten, Eibe Frank and Mark A. Hall, ***Data Mining: Practical Machine Learning Tools and Techniques***, 3rd ed, Morgan Kaufmann, 2011.

Course work and grading: In addition to class attendence, your work involves periodic in-class quizzes (20%), 4 sets of homeworks (40%), and two exams (40%). One of the lowest quiz score will be *excluded from you final grade*, so please don't ask for makeup quiz if you cannot make to a class when we schedule for a quiz. The data for each quiz will be annouced one week ahead. Some of the homework problems require programming, which combined serves as a course project.

Homework collaboration policy: Collaboration on homework is allowed and encouraged. But you need to abide by the following collaboration rules; violation of this policy may result in lossing all credit for a homework assignment. serious violation (e.g., plagiarism) will be reported to the dean's office.

- You cannot have more than 3 students collaborate on a problem.
- Your discussion MUST remain at the level of ideas and concepts, and you CANNOT work together on details where you, e.g., write equations, set parameters, etc.

- If you discuss with someone about a homework assignment, you must write that person's name on the first page of your homework submission and reveal the nature of the discussion, i.e., what your discussion is about and to what extend it covers. Following this policy is bilateral, meaning that two or more students involved in a collaboration must all reveal this in their homework submission.

Office hours:
Weixiong Zhang, office: Jolley 530, time - TBA.
TAs: TBA

Main topics (I may adjust the order and add or remove topics if needed)

1. Introduction
   - Problems that DM attempt to solve
   - The process and elements of a DM system.
   - Characteristics of data - big data means big on size, but what size?
   - Supervised vs. unsupervised learning.
   - Datamining vs. Machine Learning vs. Math vs. Computer Science
   - Similarity measure - similarity vs. distance
2. Supervised learning - Regression and classification
   - Regression
   - k-nearest neighbors
   - Decision trees
   - Random forest and bagging
   - Kernel methods and support vector machines (SVMs)
   - Discriminative method and naive Bayesian

3. Performance measure
    - Elements of performance measure and confusion matrix
    - What to compare, algorithms or problem instances?
    - Receiver Operating Characteristic (ROC) curve
4. Association rule mining
    - Frequent set and Apriori algorithm
    - Generation and selection of rules
5. Unsupervised learning - Clustering
    - Hierarchical clustering - Agglomerative vs. divisive
    - Partition-based clustering
    - k-means and the EM algorithm
    - Probabilistic model-based clustering
6. Pattern discovery through dimension reduction
    - Curse of dimensionality - The problem
    - Feature selection
    - Principle component analysis (PCA)
    - Singular value decomposition (SVD)
    - Nonnegative matrix factorization (NMF)
    - Modular or community structures of a network *
    - Spectral clustering *
    - Kernel PCA *
7. Introduction to deep learning
    - Artificial neural networks (ANNs)
    - Gradient descent for model fitting
    - Backpropagation
    - ANNs as a general model vs deep learning as special architectures

- Autoencoder
- Convolutional NNs (CNNs)
- Recurrent NNs (RNNs)

Note: The topics marked with * are optional.