CSE514 Fall 2018 - 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++, Java, 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 analytic 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 use an old school fashion of teaching, i.e., a type of chalk talk to talk through a subject and write and 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 get involved.

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 and understanding on a topic to make a complete set of notes for yourself. 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 to my lectures and help you take notes. As such, the original, incomplete lecture notes may not be an ideal resource for independent study.
The combination of my notes and your notes taken during the lectures should be your textbook for the course. 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 learning. 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:


**Course work and grading:** In addition to class attendance, 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 your final grade, so please don’t ask for makeup quiz if you cannot make to a class when we schedule for a quiz. The topics to be covered in each quiz and the date for the quiz will be announced one week ahead to help you prepare. Some of the homework problems require programming, which as a whole serves as a course project. Also note that there is no final exam for the course, instead we have two exams - the first will be scheduled around the Fall Break and the second in the last class of the course.

In case you have a problem with the grading of your work (including any homework, quiz or exam), you have until the next due date for the same type of work to ask for a regrading. For example, if you have a problem with homework 1 grading, you may ask for a regrading of that problem before homework 2 is due. The same rule applies to quiz. We will never regrade
any previous work if it has passed the regarding deadline. Also note that the result of a regarding can go either way - you may get more points or you may lose more points depending on the problem and your answer(s).

**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 losing 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 as 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. Supervised learning - Neural networks
• Artificial neural networks (ANNs)
• Gradient descent for model fitting
• Backpropagation

4. Performance measure
• Elements of performance measure and confusion matrix
• What to compare, algorithms or problem instances?
• Receiver Operating Characteristic (ROC) curve

5. Association rule mining
• Frequent set and Apriori algorithm
• Generation and selection of rules

6. Unsupervised learning - Clustering
• Hierarchical clustering - Agglomerative vs. divisive
• Partition-based clustering
• $k$-means and the EM algorithm
• Probabilistic model-based clustering
7. Pattern discovery through dimension reduction
   • Curse of dimensionality - The problem
   • Feature selection
   • Principle component analysis (PCA)
   • Singular value decomposition (SVD)
   • Nonnegative matrix factorization (NMF) *
   • Identification of modular structures of network *

8. Deep learning
   • ANNs as a general model vs deep learning as special architectures
   • Autoencoder
   • Convolutional NNs (CNNs)
   • Recurrent NNs (RNNs)
   • Insights into deep neural networks

Note: The topics marked with * are optional.