CSCI-599 DATA MINING AND STATISTICAL INFERENCE

Course Information

- *Course ID and title*: CSCI-599 Data Mining and Statistical Inference
- *Semester and day/time/location*: Spring 2013/ Mon/Wed 3:30 -4:50pm
- *Instructor*: Yan Liu
- *Office and office hours*: PHE 336/ Wed 5:30pm – 6:30pm
- *Phone and email*: (213)740-4371; yanliu.cs@usc.edu
- *Blackboard address, homepage (if relevant)*: http://www-bcf.usc.edu/~liu32/spring2012.html
- *Recommended preparation*: Probabilistic Methods (EE464, EE465, MATH 407, MATH 408) and Applied Linear Algebra (EE441, MATH 370 or MATH 471) or any undergraduate level classes that provide the foundation of statistics and linear algebra.

Introduction and Purposes

Data mining is the process of uncovering meaningful correlations, patterns, and trends from large amounts of data. Statistical inference is to build predictive models from the data, make inference and generate predictions for the value of unseen data. Data mining and statistical inference easily find applications in social media analysis, mobile data analysis, biology, climate modeling, health care analytics, astronomy, business analytics and so on.

This class aims to provide an introduction of the fundamental techniques in data mining and statistical inference as well as their applications in social media analysis, recommendation system, and massive data analytics. Topics include density estimation (parametric and nonparametric approach), linear and nonlinear regression, classification algorithms, clustering algorithms, association rules, dimension reduction, anomaly detection, graph mining, and time-series analysis. The class consists of lectures by the instructor, homework, course projects, final exam and one invited talk.

Required text/readings materials

There are two suggested textbook:

1. The elements of statistical learning: data mining, inference and prediction. Trevor Hastie, Robert Tibshirani, Jerome. H. Friedman. Springer, 2009 (Short name as ESL). Freely available at: http://www-stat-class.stanford.edu/~tibs/ElemStatLearn/

2. Data Mining: Concepts and Techniques, 3rd ed. Jiawei Han, Micheline Kamber and Jian Pei. The Morgan Kaufmann Series in Data Management Systems (Short name DM).

and students may find the following textbooks useful:

All of statistics: a concise course in statistical inference. Larry Wasserman. Springer, 2004.

Grading

The grade will be derived from four parts:

(1) Homework (25%)
(2) Course project (30%)
(3) Exams (40%): screening exam (5%), mid-term exam (15%) and final exam (20%)
(4) Class participation (5%)

- *Course project*: the purpose of the class project is for the students to learn hands-on experience of solving data mining problems. Students are encouraged to identify new applications, but sample topics will be provided to the students with less experience in data mining and machine learning. Working as a group is permitted, and a team can consist of 1-2 persons.

  Timeline:
  Jan 14 – Feb 10: Identifying team members and project topics
  Feb 11: Proposal due (team member, topics and milestone)
  Mar 30: Mid-term report due (data description, preliminary results)
  Apr 30: Project presentation and Poster session (open to all faculty and students)
  Final report due (task and model description, major discovery, lessons learned)

  Sample projects "*Mining Social Network from Twitter*": the goal of the project is to develop graph mining algorithms for twitter data analysis. Students can easily find resources available online, including twitter API and the C++ code of graph mining. A project of this size usually consists of 2 persons. The team will work together on collecting the twitter data, examining the preliminary results, identifying one challenge in existing approaches, and discussing potential solutions.

  Grading breakdown:
  Proposal: 5%
  Mid-term report: 5%
  Final report: 5%
  Poster: 15%
  All members in one team will get the same grade

Course Readings/Class Sessions

Date	Topics	Readings	Assignment
Jan 14	Overview of class	ESL Chapter 1,	

		DM Chapter 1	
Jan 16	Review Lectures on Probability and Statistics	Handout	Screening Exam
Jan 21	Martin Luther King Day – No class		
Jan 23	Density Estimation: Parametric Approach and Nonparametric Approach	ESL 8.2, ESL 7.1-7.5, 7.7-7.8	
Jan 28		ESL 6.6	HW #1 Out
Jan 30	Linear Regression	ESL 2.3.1, 3.1-3.3, 3.4, 6.1	
Feb 4	Nonlinear regression	ESL 6.1	
Feb 6	Decision Tree and Information Theory	ESL 9.2	HW #1 Due
Feb 11	Naïve Bayes and Generative Models	ESL 4.1-4.5, 11.3	HW #2 Out
Feb 13	Support Vector Machines, Logistic regression and discriminative models	ESL 12.1-12.3	
Feb 18	Presidents' Day – No Class		
Feb 20	Clustering: K-means, Mixture of Gaussian, hierarchical Clustering	ESL 14.3, DM 7.1-7.6	
Feb 25			
Feb 27	Association Rules	DM 5.1- 5.6, ESL 14.2	HW #2 Due
Mar 4	Dimension Reduction: PCA, ICA, Manifold Learning	ESL 14.5 – 14.7	HW # 3 Out
Mar 6			Mid-term Exam
Mar 11	Time Series Analysis: ARMA, Exponential Smoothing	DM 8.2	Project Proposal Due
Mar 13	Time Series Analysis: Hidden Markov Model	Handout and DM 8.1	
Mar 18, 20	Spring Break – No Class		
Mar 25			
Mar 27	Graph Mining: Graph Search and Classification	DM 9.1	HW #3 Due
Mar 21	Graph Modeling	Handout	HW #4 Out
Mar 26	Anomaly Detection	DM 7.11	
Apr 1	Recommendation System: matrix factorization	Handout	
Apr 3	Recommendation System: latent approach	Handout	Project Mid-term report Due
Apr 8	Social Media Analysis:	DM 9.2	HW # 4 Due

	authority, centrality		
Apr 10	Social Media Analysis: topic modeling	DM 10.4	HW #5 Out
Apr 15	Massive Data Analysis: Map-reduce	Handout	
Apr 17	Massive Data Analysis: Parallel Learning Algorithms	Handout	
Apr 22	Massive Data Analysis: Online Learning	Handout	
Apr 23	Review of Classes		Quiz #3
Apr 29	Poster Presentation		
May 1	Poster Presentation		HW #5 Due, Project final report due

Late Policies

Homework is due at 12:00 pm of the indicated day (by email). Each student is allowed to miss the deadline once without penalty. The penalty of late submission is equal to no submission.

Proposal, mid-term report and final report are due at the beginning of the class on the indicated days. Each student is granted an extension of three days for either proposal and/or mid-term report without penalty. No extension will be granted for final report.
The penalty of late submission is equal to no submission (Proposal: 5%, mid-term report: 5%, final report: 5%, poster: 15%).

Statement for Students with Disabilities

Any student requesting academic accommodations based on a disability is required to register with Disability Services and Programs (DSP) each semester. A letter of verification for approved accommodations can be obtained from DSP. Please be sure the letter is delivered to me (or to TA) as early in the semester as possible. DSP is located in STU 301 and is open 8:30 a.m.–5:00 p.m., Monday through Friday. The phone number for DSP is (213) 740-0776.

Statement on Academic Integrity

USC seeks to maintain an optimal learning environment. General principles of academic honesty include the concept of respect for the intellectual property of others, the expectation that individual work will be submitted unless otherwise allowed by an instructor, and the obligations both to protect one's own academic work from misuse by others as well as to avoid using another's work as one's own. All students are expected to understand and abide by these principles. *Scampus*, the Student Guidebook, contains the Student Conduct Code in Section 11.00, while the recommended sanctions are located in Appendix A: http://www.usc.edu/dept/publications/SCAMPUS/gov/. Students will be referred to the Office of Student Judicial Affairs and Community Standards for further review, should there be any suspicion of academic dishonesty. The Review process can be found at: http://www.usc.edu/student-affairs/SJACS/.

Emergency Preparedness/Course Continuity in a Crisis

In case of a declared emergency if travel to campus is not feasible, USC executive leadership will announce an electronic way for instructors to teach students in their residence halls or homes using a combination of Blackboard, teleconferencing, and other technologies.