CSCI-599 Advanced Big Data Analytics

1. Basic Information

Course:	**Advanced Data Analytics**, CSCI-599
Place and time:	TBA, Wed 2:00-4:40pm/ Fall
Instructor:	Yan Liu
	Assistant Professor of Computer Science
Office:	Powell Hall (PHE 336)
Telephone:	(213)740-4371
Email:	yanliu.cs@usc.edu
Office hours:	Wed 5:00pm – 6:00pm
Class web page:	http://www-bcf.usc.edu/~liu32/fall2011.htm

Recommended preparation: CS567 Machine Learning or CS573 Advanced Artificial Intelligence or EE559 Mathematical Pattern Recognition, or any other graduate level classes that provide the foundation of machine learning, or permission by instructor.

2. Classroom policy

Electronic communication devices (phones, blackberries, and similar) must be turned off or placed away during lectures and laboratories. You can check them at the break. Likewise, you should not use instant messenger or similar chat programs during lectures.

3. Course Goals, Learning Objectives, and relationship to CS Program Outcomes

3.1. Goals: We are currently in an era of data deluge. In many areas and domains, data are generated at a phenomenal speed that we have never experienced before. Given the large amount of data, one fundamental scientific challenge is how to develop efficient and effective computational tools to analyze the data, revealing insight and make predictions. Data analytics is the science of achieving these goals. It is an inter disciplines of machine learning, data mining, statistics, and so on.

This class aims to provide an overview of advanced machine learning, data mining and statistical techniques that arise in data analytic applications. In this class, you will learn and practice advanced data analytic techniques, including: parallel algorithms, online algorithm, locality sensitive hashing, topic modeling, structure learning, and time-series analysis. One or more applications associated with each technique will also be discussed. Notice that the class is an advanced class build on top of the knowledge learned in CS567 Machine Learning or CS573 Advanced Artificial Intelligence.

The class consists of lectures by the instructor, student presentations and discussions on recent papers in this area (which will change gradually as the area evolves), course projects, and invited talks by top researchers in this field.

3.2 Learning objectives and relationship to CS program outcomes[i]*:* After successfully completing this course, you should be able to:

- Describe the basic idea of graphical model (outcome a, b);
- Explain the basic inference and learning algorithms of graphical models (outcome a, b);
- Analyze the applications and design graphical model solutions to the problems (outcome c, i, j);
- Explain the basic idea of advanced topic modeling techniques (outcome a, b);
- Analyze the text contents in real applications and design topic modeling models to mining and learning text contents (outcome c, i, j);
- Explain the basic idea of graph structure learning techniques (outcome a, b);
- Analyze the relational data in real applications and design graph structure learning techniques to learn the independence structures from the data (outcome c, i, j);
- Explain the basic idea of the graph mining and graph modeling techniques (outcome a, b);
- Analyze the graph data in real applications and design graph mining techniques to reveal frequent graph patterns in the data (outcome c, i, j);
- Explain the basic idea of the time series and spatial time series techniques (outcome a, b);
- Analyze the time series and spatial time series data in real applications, and design statistical models to predict and modeling (outcome c, i, j);
- Explain the basic idea of different techniques in learning with less supervision (outcome a, b);
- Analyze the classification and regression problems in the real applications and design machine learning techniques to resolve the issue that no sufficient training data are available (outcome b, c, j);
- Implement, test and troubleshoot massive data analytics algorithms using existing parallel program environment (outcome c, i, j);
- Work as a team to practice statistical learning and data mining techniques and complete the process of data analytics (outcome d, g, h);
- Prepare written reports and technical illustrations summarizing procedures, technical results and interpretation of experiment results and projects (outcome i, j, k);
- Prepare presentation slides, and lead discussions in the class (outcome e, f);
- Supplement through independent study of the course textbook, readings, and component data sheets the presentations of the course material given in class (outcome g, h);

CSCI-686 strongly contributes to CS program outcomes a, b, c, i, and j, and moderately to CS program outcomes d, e, f, g and k.

4. Course Plan

The course plan detailed below reflects the course goal and the learning objectives. Review of papers, student presentations and a final project are planned to sharpen these skills and extend them to real applications. The class material is covered in the following tentative order:

Week 1: Introduction and review lectures on probability and statistics
Week 2: Review of graphical models 1: directed and undirected graphical models
Week 3: Review of graphical models 2: inference and learning
Week 4: Parallel programming environment
Week 5: Topic modeling 1
Week 6: Topic modeling 2
Week 7: Graph structure learning 1
Week 8: Graph structure learning 2
Week 9: Graph mining
Week 10: Graph modeling
Week 11: Time series analysis
Week 12: Spatial time series analysis
Week 13: Massive Data Analytics: parallel algorithms
Week 14: Massive Data Analytics: online learning algorithms
Week 15: Massive Data Analytics: locality sensitive hashing
Week 16: Course Project Presentation

Corresponding reading assignments are listed at the end of the syllabus.

5. Teaching Team

Instructor:
Yan Liu, PhD, Assistant Professor in Computer Science Department
Office: PHE 336
Tel: (213)740-4371; Email: yanliu.cs@usc.edu

Teaching Assistant:
Taha Bahadori; Email: mohammab@usc.edu

6. Source Material

- Textbook: There are no required textbooks. Students may find the following books useful:
  C. Bishop, Pattern Recognition and Machine Learning, Springer 2007.
  All of statistics: a concise course in statistical inference. Larry Wasserman. Springer, 2004.
  Trevor Hastie, Robert Tibshirani, Jerome. H. Friedman. The elements of statistical learning: data mining, inference and prediction. Springer, 2009 (Short name as ESL). Freely available at: http://www-stat-class.stanford.edu/~tibs/ElemStatLearn/

- Reading materials: electronic version of the required research papers will be available on course websites. The set of papers will change over the time with the development of the research area.

7. Assessment

Learners are assessed based on their grades five paper reviews, three quizzes, in-class presentations, and course projects. The following schedule and percentages are used:

Assessment Procedure	Date	Proportion
Quiz #1	Sep 5	5%
Quiz #2	Oct 17	5%
Quiz #3	Nov 14	5%
Paper Reviews	Every other week (6 times)	30% (5% each review)
Course Project	Dec 5	30%
In-class Presentations	Two papers per student	25%

7.1. Quiz: The three quizzes are closed book tests for which you are only required to bring a calculator and a pen. Each quiz will have 4 problems solving questions, which cover the all contents taught in the class up to the day of the quiz. A sample-quiz will be provided one week before each scheduled quiz to familiarize the students with the material and format of the upcoming quiz.

Students who are not able to attend a quiz (medical or other emergency) must notify the instructor before the test (email at yanliu.cs@usc.edu).

7.2 Paper reviews: the purpose of the reviews is for you to read the papers and required reading before the class, so that we can have effective learning and discussions in the class. The reviews should cover three parts for each paper, including (1) summary of the basic idea of the paper; (2) discuss pros and cons of the proposed methodology; (3) if you're the author of the paper, what improvement you will make for the paper. The reviews are due in the class of the indicated day.

7.2 Course project: the purpose of the class project is for you to learn hands-on experience of solving data analytic problems. Students are encouraged to identify new applications, but sample topics will be provided to 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:

Sep 5 – Sep 19: Identifying team members and project topics

Sep 19: Proposal due (team member, topics and milestone)
Oct 24: Mid-term report due (data description, preliminary results)
Dec 5: Project presentation and Poster session (open to all faculty and students)
Dec 5: Final report due (task and model description, major discovery, lessons learned)

Sample projects "*Topic-modeling for analyzing twitter data*": the goal of the project is to develop a topic model for twitter data. Students can easily find resources available online, including twitter API and the C++ topic modeling code (e.g. LDA model). 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 current topic models for twitter application, and providing a reasonable solution.

Grading breakdown of the course project:
Proposal: 5%
Mid-term report: 5%
Final report: 5%
Presentation: 10%
Poster: 5%

All members in one team will get the same grade

7.4 In-class Presentations: the purpose of the in-class presentation is for you to lead the discussions on research papers and understand in-depth of the techniques in the paper. Each student will present 1-2 papers, which is determined based on your research interest. In the presentation, you should prepare presentation slides of 10-12 pages for each paper, which should cover the following contents: (1) the motivation of the paper; (2) the proposed techniques; (3) the experiment results; (4) pros and cons of the proposed techniques.

8.5. Late policies: Reviews are due in class of the indicated day. Each student is allowed to miss the deadline once by 7 days 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 indicated days. Each student is allowed a total of three days of extension for either proposal and/or mid-term report without penalty. No extension is granted for final report. The penalty of late submission is equal to no submission.
(Proposal: 5%, mid-term report: 5%, final report: 5%, presentation: 10%)

The in-class presentations will not allow any extensions.

8.6. Course grade: The course grade is computed based on the individual assessment grades using the indicated percentages. The letter grade is assigned on a straight scale: 85% and above leading to A, 70%-85% leading to B, etc. Pluses and minuses are

assigned by dividing each range in corresponding halves (A, A-) or thirds (B+, B, B-, C+, ...).

9. Policy against Cheating

All USC students are responsible for reading and following the Student Conduct Code, which appears in the Scampus and at http://www.usc.edu/dept/publications/SCAMPUS/goverance.

The USC Student Conduct Code prohibits plagiarism. Some examples of what is not allowed by the conduct code: copying all or part of someone else's work (by hand or by looking at others' files, either secretly or if shown), and submitting it as your own; giving another student in the class a copy of your assignment solution; consulting with another student during an exam; modifying a graded assignment before asking for re-grading, letting your lab partner prepare the report and expect a grade for their work. If you have questions about what is allowed, please discuss it with the instructor.

The policy regarding incidental cheating for this course is the following: students found cheating on a homework or laboratory assignment will not receive a grade on that assignment. Instead, the points corresponding to the assignment will be reassigned to final exam. Repeat offenders will expose themselves to the general University policy. Students who violate University standards of academic integrity are subject to disciplinary sanctions, including failure in the course and suspension from the University. Since dishonesty in any form harms the individual, other students, and the University, policies on academic integrity will be strictly enforced. We expect you to familiarize yourself with the Academic Integrity guidelines found in the current Scampus. Violations of the Student Conduct Code will be filed with the Office of Student Conduct, and appropriate sanctions will be given.

This policy does not apply to discussion, exchange of information, working together, etc. On the contrary, we encourage that you consult with classmates regarding learning material and homework assignments. Team projects require that you work with your team and assist your partner as much as he or she assists you. However, for individual marks, it is required that you prepare the final product by yourself and to the best of your possibilities; for group marks, it is required that you bring in to the group as much as you take from the group.

10. Disability Policy Statement:

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.

11. Resources

11.1. Web page: A class website will be setup on Backboard containing information about the course: syllabus, laboratory handouts, grades, miscellaneous information about weekly class activities, solution to the homework sets, and an email directory of all people in the class. Use it as much as you find it useful. The web page can be accessed through the main stem https://Blackboard.usc.edu.

11.2 Office Hours: The teaching assistants and I will hold office hours every week. This is for your benefit and you should feel welcome to the office hours as much as you need assistance. Time and location for my office hours are at the beginning of the syllabus. Those of the teaching assistant will be decided with you in class. Both of us are available by email to help you as much as you need.

12. Weekly Readings

To maximize the benefit of attending class, you must read the selected pages listed below before coming to class. You should also look at the corresponding section of the learning guide.

Date	Topics	Readings (see reference below)	Assignment
Aug 29	(1) Overview of class (2) Review Lectures on Probability and Statistics	Introduction, agenda and course project; Basic concepts on probability, statistic, linear algebra, and calculus	
Sep 5	Graphical Models (Part 1)	Textbook: page 359-390	Sample Quiz
Sep 12	Graphical Models (Part 2)	Textbook: page 391-418	
Sep 19	Parallel programming environment	TBA	In-class Quiz #1
Sep 26	Topic modeling	Latent semantic indexing [Deerwester et al, 1990], probabilistic latent semantic indexing [Hofmann 1999], Latent dirichlet allocation (LDA) [Blei et al, 2003]	Project Proposal Due
Oct 3		Gibbs sampling for LDA [Griffiths et al, 2004], author	Review #1 Due

		topic model [Rosen-Zvi et al, 2004], Author-receipt-topic models [McCallum et al, 2005], and topic link LDA models [Liu et al, 2009]	
Oct 10	Graph structure learning	Constraint-based and score-based algorithms [Heckerman, 1999: Section 7-12], L1-based structure learning algorithm [Meinshausen et al, 2006, Friedman et al, 2008]	
Oct 17		Logistic regression-based algorithm [Wainwright et al, 2006], microarray data analysis [Friedman, 2004], 8nomaly detection [Schmit et al, 2008], theoretical analysis [Bento & Montanari, 2011]	Review #2 Due
Oct 24	Graph Mining	Graph analysis [Leskovec et al, 2005] and Invited Talk	In-class Quiz #2; Review #3 Due
Oct 31	Graph Modeling	[Airoldi et al, 2008]	Mid-term proposal Due
Nov 7	Time-series and spatial data analysis	Handout provided in the previous class	Review #4 Due
Nov 14		Non-linear time-series analysis [Tong 2002], spatial analysis	Review #5 Due In-class Quiz #3

		[Finkelstein, 1984], spatial-temporal models [Smith, 2003, Yin et al, 2009]	
Nov 28	Massive data analytics	Map-reduce for machine learning [Chu et al, 2006], Nearest-neighbor classifier [Liu et al, 2004], Multi-task learning [Weinberger et al, 2009] ,Topic model [Newman et al, 2007], [Porteous et al, 2008]	Review #6 Due
Dec 5	Project presentation and Poster Sessions	10 min presentation	Project Final Report Due

Reference:

[Airoldi et al, 2008] Airoldi, E.M., Blei, D.M., Fienberg, S.E., & Xing, E.P. (2008). Mixed-membership stochastic blockmodels. Journal of Machine Learning Research, 9, 1981-2014.

[Deerwester et al, 1990] S. Deerwester, S. T. Dumais, G. W. Furnas, T.K. Landauer, and R. Harshman. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 1990.

[Hofmann 1999] T. Hofmann. Probabilistic latent semantic indexing. In Proceedings of SIGIR, 1999.

[Blei et al, 2003] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. Journal of Machine Learning Research, 3:993–1022, January 2003.

[Griffiths et al, 2004] T. L. Griffiths and M. Steyvers. Finding scientific topics. In Proceedings of the National Academy of Sciences, 101, 5228-5235, 2004.

[Rosen-Zvi et al, 2004] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The Author-Topic Model for authors and documents. In Proceedings of UAI, 2004.

[McCallum et al, 2005] A. McCallum, A. Corrada-Emmanuel and X. Wang. Topic and Role Discovery in Social Networks. In Proceedings of IJCAI, 2005

[Liu et al, 2009] Y. Liu, A. Niculescu-Mizil, and W. Gryc. Topic-link LDA: joint models of topic and author community. In Proceedings of ICML, 2009.

[Heckerman, 1999] D. Heckerman. A Tutorial on Learning with Bayesian Networks. In Learning in Graphical Models, M. Jordan, ed.. MIT Press, Cambridge, MA, 1999.

[Meinshausen et al, 2006] N. Meinshausen, P. BÃ¼hlmann, and E. ZÃ¼rich. High dimensional graphs and variable selection with the Lasso. Annals of Statistics, 2006.

[Schmidt et al, 2008] M Schmidt, K Murphy, G Fung. Structure learning in random fields for heart motion abnormality detection, In Proceedings of CVPR, 2008.

[Wainwright et al, 2006] M. J. Wainwright, P. Ravikumar, and J. D. Lafferty. High-Dimensional Graphical Model Selection Using L1-Regularized Logistic Regression. In NIPS 2006.

[Bento & Montanari, 2011] J. Bento and A. Montanari. Which graphical models are difficult to learn? In NIPS 2011.

[Friedman et al, 2008] J. Friedman, T. Hastie, R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. Biostatistics, Jul;9(3):432-41, 2008.

[Friedman, 2004] N. Friedman. Inferring cellular networks using probabilistic graphical models. Science. 303:799-805, 2004.

[Leskovec et al, 2005] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In Proceedings of SIGKDD, 2005.

[Finkelstein, 1984] Peter L. Finkelstein. The Spatial Analysis of Acid Precipitation Data. Journal of Climate and Applied Meteorology 23:1, 52-62, 1984.

[Tong 2002] Howell Tong. Nonlinear time series analysis since 1990: some personal reflections, 2002.

[Smith, 2003] Richard L. Smith. Spatio-temporal model.
http://www.stat.unc.edu/faculty/rs/s321/spatemp.pdf

[Yin et al, 2009] J. Yin, D. H.Hu and Q. Yang. Spatio-Temporal Event Detection Using Dynamic Conditional Random Fields. In IJCAI 2009.

[Blitzer et al, 2006] J. Blitzer, R. McDonald, and F. Pereira. Domain adaptation with structural correspondence learning. In Proceedings of EMNLP, 2006.

[Druck et al, 2008] Druck, G., Mann, G., & McCallum, A. Learning from labeled features using generalized expectation criteria. In Proceedings of SIGIR, 2008.

[Sheng et al, 2008] S. Sheng, F. Provost and P. Ipeirotis. Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers. In Proceedings of SIGKDD, 2008.

[Lafferty et al, 2001] J. Lafferty, A. McCallum, F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceeding of ICML 2001.

[Liu et al, 2004] T.Liu, A. W. Moore, A. Gray, and K. Yang. An Investigation of Practical Approximate Nearest Neighbor Algorithms. In Proceedings of NIPS, 2004.

[Aggarwal, 2003] C. Aggarwal, J. Han, J. Wang, and P.S. Yu. A framework for clustering evolving data streams. In Proceedings of VLDB, 2003.

[Newman et al, 2007] D. Newman, A. Asuncion, P. Smyth, and M. Welling. Distributed Inference for Latent Dirichlet Allocation. In Proccedings of NIPS, 2007.

[Porteous et al, 2008] I. Porteous, D. Newman, A. Ihler, A. Asuncion, P. Smyth, and M. Welling. Fast collapsed gibbs sampling for latent dirichlet allocation. Proceeding of SIGKDD, 2008.

[Chu et al, 2006] C. T. Chu, S. K. Kim, Y. A. Lin, Y. Yu, G. R. Bradski, A. Y. Ng, and K. Olukotun. Map-Reduce for Machine Learning on Multicore. In Proceedings of NIPS, 2006.

[Weinberger et al, 2009] K. Weinberger, A. Dasgupta, J. Attenberg, J. Langford, A. Smola. Feature Hashing for Large Scale Multitask Learning. In Proceedings of ICML, 2009.

---

[i] COMPUTER SCIENCE PROGRAM OUTCOMES

[i] Students completing the computer science program should have

(a) An ability to apply knowledge of computing and mathematics appropriate to the discipline;

(b) An ability to analyze a problem, and identify and define the computing requirements appropriate to its solution;

(c) An ability to design, implement and evaluate a computer-based system, process, component, or program to meet desired needs;

(d) An ability to function effectively on teams to accomplish a common goal;

(e) An understanding of professional, ethical, legal, security, and social issues and responsibilities;

(f) An ability to communicate effectively with a range of audiences;

(g) An ability to analyze the local and global impact of computing on individuals, organizations and society;

(h) Recognition of the need for, and an ability to engage in, continuing professional development;

(i) a recognition of the need for, and an ability to engage in life-long learning

(j) An ability to apply mathematical foundations, algorithmic principles, and computer science theory in the modeling and design of computer-based systems in a way that demonstrates comprehension of the tradeoffs involved in design choices;

(k) An ability to apply design and development principles in the construction of software systems of varying complexity.