SYLLABUS OF SELECTED TOPICS IN DATA ANALYTICS

Heading and Administrative Matters
• *Course ID and title*: Selected Topic in Data Analytics
• *Recommended preparation*: CS542 or CS567 or CS573, or any other graduate level classes that provide the foundation of machine learning, or permission by instructor.
• *Semester and day/time/location*: Fall / Tue and Thu 12:30pm – 1:50pm/ THH 106
• Professor(s) and how to contact
  o *Instructor*: Yan Liu
  o *Office and office hours*: 5:00pm – 6:00pm Tue
  o *Phone and email*: TBA
  o *Blackboard address, homepage*: TBA

Introduction and Purposes
Data analytics is the science of analyzing the data, generating insights, and making predictions. It easily finds applications in social media analysis, computational biology, climate modeling, health care, traffic monitoring and so on.

This class aims to provide an overview of advanced machine learning, data mining and statistic techniques that arise in real data analytic applications. Selected topics include topic modeling, structure learning, time-series analysis, learning with less supervision, and massive-scale data analytics. One or more applications associated with each technique will also be discussed. The class consists of two lectures by the instructor on overview of the course, student presentations of papers and course projects, and two invited talks by the speakers from academia and industries.

Course Requirements and Grades
• *Required text/readings and supplementary instructional materials* (students need to know what they need to buy, if anything)
  There are no required textbooks. Students may find the following books useful:

  2. C. Bishop, Pattern Recognition and Machine Learning, Springer 2007

• *Grading breakdown*
  o *Statement of percentages of grade associated with any graded assignment or exam.*
    The grade will be derived from four parts:
    (1) Write reviews (two papers) every week (30%)
    (2) Present one paper or two papers and lead the discussion after his/her presentation (25%)
    (3) Complete a research project (40%)
    (4) Class participation (5%)
• **Course project:** the purpose of the class project is for the students 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 1-3 persons, but a team of 3 persons need to provide justified reasons for a bigger team in the proposal.

Timeline:
Aug 24 – Sep 12: Identifying team members and project topics
Sep 14: Proposal due (team member, topics and milestone)
Oct 21: Mid-term report due (data description, preliminary results)
Nov 30: Project presentation
Dec 2: Poster (open to all faculty and students) and final report (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:
Proposal: 5%
Mid-term report: 5%
Final report: 10%
Presentation: 10%
Poster session: 10%
All members in one team will get the same grade

**Course Readings/Class Sessions**
• Students should have a detailed course calendar, including dates of exams and assignments, with a weekly breakout of topics and assignments including:
  o Required reading or other assignments per class session, including pages
  o Expectations about use of reading for the class session, and role of required and any supplementary reading.
  o Changes to the syllabus regarding course requirements (such as a change of topic in a graduate seminar) are communicated clearly to students.

<table>
<thead>
<tr>
<th>Date</th>
<th>Topics</th>
<th>Readings</th>
<th>Assignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aug 24</td>
<td>Overview of class</td>
<td>Introduction, agenda and course project</td>
<td></td>
</tr>
<tr>
<td>Aug 26</td>
<td>Review of basics</td>
<td>basic concepts on probability, statistic, linear algebra, and</td>
<td>In-class quiz (WILL NOT be counted for</td>
</tr>
<tr>
<td>Date</td>
<td>Topic</td>
<td>Lecture Content</td>
<td>Notes</td>
</tr>
<tr>
<td>---------</td>
<td>--------------------------------------------</td>
<td>----------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>--------------------------------</td>
</tr>
<tr>
<td>Sep 2</td>
<td></td>
<td>Latent dirichlet allocation (LDA) [Blei et al, 2003]</td>
<td></td>
</tr>
<tr>
<td>Sep 14</td>
<td>Graph structure learning</td>
<td>Traditional structure learning techniques: constraint-based and score-based algorithms [Heckerman, 1999: Section 7-12]</td>
<td>Review due</td>
</tr>
<tr>
<td>Sep 16</td>
<td></td>
<td>L1-based structure learning algorithm [Meinshausen et al, 2006; Friedman et al, 2008]</td>
<td></td>
</tr>
<tr>
<td>Sep 21</td>
<td></td>
<td>Structure learning with priors [Gu, 2007]</td>
<td>Review due</td>
</tr>
<tr>
<td>Sep 23</td>
<td></td>
<td>Applications: microarray data analysis [Friedman, 2004]</td>
<td></td>
</tr>
<tr>
<td>Sep 28</td>
<td>Graph Mining</td>
<td>Graph analysis [Leskovec et al, 2005]</td>
<td>Review due</td>
</tr>
<tr>
<td>Sep 30</td>
<td></td>
<td>Invited Talk: managing and mining graphs by Xifeng Yan (UCSB)</td>
<td></td>
</tr>
<tr>
<td>Oct 7</td>
<td></td>
<td>Non-stationary time-series analysis [Tong 2001, 2002]</td>
<td></td>
</tr>
<tr>
<td>Oct 14</td>
<td></td>
<td>Spatial-temporal model [Smith, 2003]</td>
<td></td>
</tr>
<tr>
<td>Oct 19</td>
<td>Learning with less supervision</td>
<td>Learning from data in other domain [Blitzer et al, 2006]</td>
<td>Review due</td>
</tr>
<tr>
<td>Oct 21</td>
<td></td>
<td>Measurement models [Liang et al, 2009]</td>
<td></td>
</tr>
<tr>
<td>Oct 26</td>
<td></td>
<td>Learning with labeled predicates [Druck et al, 2008]</td>
<td>Review due</td>
</tr>
<tr>
<td>Date</td>
<td>Activity</td>
<td>Details</td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>---------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Oct 28</td>
<td>Learning with structured-label constraints [Lafferty et al, 2001]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nov 4</td>
<td>Invited talk</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nov 9</td>
<td>Massive data analytics</td>
<td>Map-reduce for machine learning [Chu et al, 2006]</td>
<td>Review due</td>
</tr>
<tr>
<td>Nov 11</td>
<td>Nearest-neighbor classifier [Liu et al, 2004]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nov 16</td>
<td>Clustering stream data</td>
<td>Multi-task learning [Weinberger et al, 2009]</td>
<td>Review due</td>
</tr>
<tr>
<td>Nov 18</td>
<td>Nearest-neighbor classifier [Liu et al, 2004]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nov 25</td>
<td>Thanksgiving - No class</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nov 30</td>
<td>Project presentation</td>
<td>10 min presentation</td>
<td></td>
</tr>
<tr>
<td>Dec 2</td>
<td>Project presentation</td>
<td>Poster session (open to all)</td>
<td>Project final report due</td>
</tr>
</tbody>
</table>

- **Policies related to late or make-up work, if relevant**
  
  Reviews are due at 11:00 am of the indicated day (by email)  
  Each student is allowed to miss the deadline twice without penalty  
  The penalty of late submission is equal to no submission (30% * 1/12 = 2.5% each time)

  Proposal, mid-term report and final report are due at the beginning of the class on indicated days (printout only)  
  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: 10%, presentation: 10%, poster session: 10%)

**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/](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/](http://www.usc.edu/student-affairs/SJACS/).

**Reference:**


