Data Mining and Analytics

Data Mining & Analytics (INFO 254/190)
Spring 2018
University of California, Berkeley
School of Information

Instructor: Prof. Zachary Peddies
zepeddies@berkeley.edu
South Hall 202 / Talmie Hall 4641
Office Hours: After Tuesday classes in Moffitt
Note: My office hours for April 10th have been moved to
Friday April 6th 11a-12p in Talmie 4641.

GSE(s): Ali Madani
alimadan@berkeley.edu
Office Hours; 9am on Mondays in Stanley Hall Lower 1130. 15-minute slots by appointment via email.

Sayan Sanay
sayan@berkeley.edu
Office Hours; 10:30am on Wednesdays in the Co-Lab in South Hall, 15-minute slots by appointment.

Class Time: Tuesdays and Thursdays, 2:00 - 3:30 PM
Location: 145 Moffit Library

Course Description
Data Mining and Analytics introduces students to the practical fundamentals and emerging paradigms of data mining and machine learning with enough theory to add intuition building. The course is project-oriented, with a project beginning in class every Thursday and to be completed outside of class by the following week, or two for longer assignments. The in-class portion of the project is meant to be collaborative and a time for the instructor and GSI's to work closely with project groups to understand the objectives, help work through software logistics, and connect project work to lecture. Tuesdays lectures introducing theories, concepts, contexts, and algorithms. Students should expect to leave the class with hands-on, contemporary data mining skills they can confidently apply in research and industry. The course will be written midterm test and a final project report and presentation. Experience with Python is required.

Course Objectives
- Foster critical thinking about real-world actionability from analytics.
- Develop intuition in various machine learning classification algorithms (e.g., decision trees, feed-forward neural networks, recurrent neural networks, support vector machines) and clustering techniques (e.g., k-means, spectral, spectral, k-means).
- Conduct manual feature engineering (from domain knowledge) vs. machine induced featureization (representation learning).
- Provide an overview of issues in research and practice that will affect the practice of data science in a variety of domains.

Grading
Homework/Labs: 30%
Midterm: 25%
Final Project: 35%
Quizzes: 10%

Late Policy: Late submissions will be penalized 20% up to one week after the original due date. Another 10% penalty will be added for each subsequent week. UPDATE: Each student’s most severe late penalty will be forgiven at the end of the course.

Texts

Special Needs/Accommodations
I am committed to creating a learning environment welcoming of all students. If you have any special needs, please notify me as soon as possible so that appropriate accommodations can be made.

Agenda and Assignments
Typically, Tuesdays will be lecture days and Thursdays will be lab and quiz days. Labs are due one week after they are assigned (at 5pm) unless otherwise specified. Readings (recommended, not required) are associated with the day they are listed. Preparation materials for the labs will be posted on Course. Readings refer to the textbook unless otherwise specified.

Note on collaboration/group work: You may share strategies, ideas and help debug others assignments but all assignments are to be turned in individually and completed individually otherwise stated in the assignment. Please do not copy and paste code without referring the source material in your comments or write up. The final project will be a team project.

Slides and lab assignment documents: We will provide a link to the lecture slides dock at the start of each lecture. Lab assignments and their respective tutorial helper pages will be available at the start of class on the Thursday of the start of the lab.

<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
<th>Readings</th>
<th>Extra (optional) Material</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tuesday, 16</td>
<td>Course Introduction</td>
<td></td>
<td></td>
</tr>
<tr>
<td>January</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thursday, 18</td>
<td>In class office hours</td>
<td></td>
<td></td>
</tr>
<tr>
<td>January</td>
<td>Data Preprocessing lecture</td>
<td>Section 2.1</td>
<td></td>
</tr>
<tr>
<td>Tuesday, 20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>January</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Course assignments are not weighted.
<table>
<thead>
<tr>
<th>Date</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan 24, 2018</td>
<td>Beginning of class survey</td>
</tr>
<tr>
<td>Jan 29, 2018</td>
<td>Quiz 1 - Data Preprocessing</td>
</tr>
<tr>
<td>Feb 5, 2018</td>
<td>Quiz 2 - Clustering</td>
</tr>
<tr>
<td>Feb 12, 2018</td>
<td>Quiz 3 - Decision Trees</td>
</tr>
<tr>
<td>Feb 19, 2018</td>
<td>Quiz 4 - Neural Networks and SVM</td>
</tr>
<tr>
<td>Feb 26, 2018</td>
<td>Quiz 5 - Ensemble Models</td>
</tr>
<tr>
<td>Mar 5, 2018</td>
<td>Quiz 6 - Cross-validation and Error Metrics</td>
</tr>
<tr>
<td>Mar 6, 2018</td>
<td>Lab 5 - Kaggle Competition/Weather Data</td>
</tr>
</tbody>
</table>

**Course Summary:**

- **Data Pre-processing Lab**
  - Sections 10.1, 10.2, 10.6
  - Notebook

- **Clustering Lab**

- **Decision Trees Lab**
  - Sections 8.1 - 8.2.3
  - Softmax Regression

- **Neural Networks Lab**
  - Sections 9.2 (skim) and 9.3
  - Notebook

- **Kaggle Competition**
  - Notebook

- **Cross-validation and Error Metrics Lecture**
  - Section 8.5

- **Continue Kaggle Competition**

- **Midterm Review**

- **Midterm**

- **Final Exam Lecture**

- **Competition mini report-outs, Skip-gram Lab**
  - The amazing power of word vectors [repo](#)
  - Academic paper: [on word2vec](#)

- **RNNs, Recommendation & Education**
  - Outlier pitch slide due, Midterm handed back

- **Dataset 1 slide presentations for final project**

- **Spring Break, No Class**

- **Spring Break, No Class**

- **Dimensionality Reduction & Data Visualization**
  - Reducing the Dimensionality of Data with NN [repo](#)
  - Notebook

- **Quiz (word2vec, finish Dimensionality Reduction & Data Visualization Lab**

- **TFIDF + final project work**

- **Quiz (RNN - dimensionality reductions, Deep Learning Lab (RNN) - extra credit + final project work**
  - Final project due: APPROVAL due @ 11:59am

- **Advanced Clustering Lecture**
  - Self-tuning spectral clustering [repo](#)

- **Final project group work**

- **Final project presentations**

- **Final Project Presentations**

- **Final Project Presentations**

- **Final Project Presentations**

- **Final Project Presentations**

- **Final Project Presentations**

- **Final Project Presentations**
<table>
<thead>
<tr>
<th>Date</th>
<th>Activity</th>
<th>Due By</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thu Mar 10, 2016</td>
<td>Final Project Dataset Pitch</td>
<td>1:59 pm</td>
</tr>
<tr>
<td></td>
<td>Lab 6a - Word2Vec</td>
<td>1:59 pm</td>
</tr>
<tr>
<td>Thu Apr 5, 2016</td>
<td>Quiz 7 - Sklearn/numpy</td>
<td>2:20 pm</td>
</tr>
<tr>
<td></td>
<td>Lab 6b - Dimensionality Reduction and Visualization</td>
<td>2:20 pm</td>
</tr>
<tr>
<td>Thu Apr 12, 2016</td>
<td>Quiz 8 - RNN/Dimensionality reduction</td>
<td>2:20 pm</td>
</tr>
<tr>
<td></td>
<td>Final Project descriptions</td>
<td>15:59 pm</td>
</tr>
<tr>
<td>Thu Apr 19, 2016</td>
<td>Lab 7 - Neural RNN</td>
<td>2:20 pm</td>
</tr>
<tr>
<td>Tue May 3, 2016</td>
<td>Final Project Presentation (grading rubric)</td>
<td>16:59 pm</td>
</tr>
<tr>
<td></td>
<td>Final Project Write-up (submission)</td>
<td>23:59 pm</td>
</tr>
</tbody>
</table>

- Final Project Presentation sign-up for (Thursday, May 3rd)
- Final Project Presentation sign-up for (Tuesday, May 1st)