Stats 170A/B, Project in Data Science

Schedule for Spring Quarter

Michael Carey and Padhraic Smyth
Department of Computer Science
Bren School of Information and Computer Sciences
University of California, Irvine
Outline

• Weekly Schedule

• Guest speakers

• Grading
  – Attendance
  – Weekly updates
  – Presentations
  – Final report

• General guidelines on projects and presentations
<table>
<thead>
<tr>
<th>Week</th>
<th>Monday</th>
<th>Wednesday</th>
</tr>
</thead>
<tbody>
<tr>
<td>1, April 2</td>
<td>Short intro lecture</td>
<td>Joint office hours (review of projects/presentations), DBH 2091</td>
</tr>
<tr>
<td>2, April 9</td>
<td>Guest speaker #1</td>
<td>Student progress presentations</td>
</tr>
<tr>
<td>3, April 16</td>
<td>Smyth office hours, DBH 4016</td>
<td>Carey office hours, DBH 2091</td>
</tr>
<tr>
<td>4, April 23</td>
<td>Guest speaker #2</td>
<td>Student progress presentations</td>
</tr>
<tr>
<td>5, April 30</td>
<td>No office hours or class</td>
<td>Carey office hours, DBH 2091</td>
</tr>
<tr>
<td>6, May 7</td>
<td>Guest speaker #3</td>
<td>Student progress presentations</td>
</tr>
<tr>
<td>7, May 14</td>
<td>Smyth office hours, DBH 4016</td>
<td>Carey office hours, DBH 2091</td>
</tr>
<tr>
<td>8, May 21</td>
<td>Guest speaker #4</td>
<td>Student progress presentations</td>
</tr>
<tr>
<td>9, May 28</td>
<td>Memorial Day, no class</td>
<td>Smyth office hours, DBH 4016</td>
</tr>
<tr>
<td>10, June 4</td>
<td>Guest speaker #5</td>
<td>Final student presentations</td>
</tr>
</tbody>
</table>
Guest Speakers

• 5 guest speakers
  – From local companies in the data science area
  – CS PhD students who have worked for companies on data science projects

• Format of talks
  – Case studies: how data science is practiced in the real world
  – 30 minute presentation
  – Remaining time for questions
  – (these talks may be attended by other data science majors)

• Speakers and dates being finalized...will be announced shortly
## Grading

<table>
<thead>
<tr>
<th>Activity</th>
<th>Points</th>
<th>Deadline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attendance</td>
<td>10 (1 point x 10 weeks)</td>
<td>Weekly</td>
</tr>
<tr>
<td>Google doc update</td>
<td>18 (2 points x 9 weeks)</td>
<td>Weekly, due midnight Sunday</td>
</tr>
<tr>
<td>Presentations</td>
<td>40 (8 points x 5 weeks)</td>
<td>Every 2(^{nd}) Wednesday</td>
</tr>
<tr>
<td>Final Report</td>
<td>32</td>
<td>Due to EEE by midnight Sunday June 9th</td>
</tr>
</tbody>
</table>
More on Grading

• Attendance
  – Attend all 5 talks (5 points)
  – Attend at least 1 office hour meeting per week (for 5 weeks with office hours)

• Weekly summary in Google docs
  – Bulleted list of what you did that week (cumulative each week)
  – Due every week for first 9 weeks of class. At least 3 bullets per week.
  – You will need to send us a link to your Google doc (week 1, via Piazza)

• In-class presentations
  – 6 minute presentation + 4 minute discussion and questions
  – Wednesdays, every 2\textsuperscript{nd} week, starting week 2
  – Must do at least 2 slide presentations and 2 Jupyter notebook presentations

• Final report
  – Written report, 5 to 10 pages, format and instructions will be announced later
Grading across Winter and Spring quarters

• 50% of your final grade will come from each quarter

• Winter quarter:
  – Homework: 40%
  – Project proposal: 50%
  – Class participation: 10%

• We will let you know your grades for Winter later this week
Example of a Weekly Summary

• Week 1, April 2
  – Downloaded a 10Gbyte data set from Yelp
  – Wrote Python code to parse JSON files for reviews, businesses, users
  – Imported data into Pandas and did initial exploration
    • Summary: found 300k users, 1.5 million reviews, 60k business, etc
  – Designed initial database schema to store data
  – Next week: implement and test database

• Week 2, April 9
  – Implemented database tables in PostgreSQL
  – Loaded JSON Yelp data into Python and inserted into table in my database
  – Tested database-Python connectivity with sample queries
    • Example query:
    • Result:
  – Read 3 research papers/articles on information extraction from text
  – Next week: design and begin implementing of text analysis pipeline
Suggestions for your Project Presentations (every 2 weeks)

• The next few slides discuss the type of information we would like to see presented each week
  – You don’t need to follow this precise format, but you should think about covering most or all of these items if you can

• “A picture is worth a thousand words”
  – Avoid having too many “text-heavy slides” (like this one!)

• You don’t need to give us all the details
  – You will only have 6 minutes
  – Make sure you communicate the main points clearly

• Preparing for your presentation
  – Revise, revise, revise
  – Practice your talk: does it flow? Are some slides too complex? Is the overall presentation too long? Have you covered what we need to know?
Weekly Presentation: Project Overview

• An initial slide that states clearly the major aspects of your project
  – What problem you are solving
  – Data sets
  – Technical approaches
  – Evaluation methods
  (This is to remind everyone what your project is about, in 1 slide)

• Figures are very helpful
  – You may want to have one major figure that you start with each week to remind the audience of what your project is about
  – Can then point to parts/components you are going to discuss that day
Weekly Presentation: Timetable

• Provide a brief timetable of milestones (both achieved and planned)

• For example, a list for every 2 weeks
  – *Weeks 1-2*
  – *Weeks 3-4*
  – *Weeks 5-6*
  – *Weeks 7-8*
  – *Weeks 9-10*

• This can be similar for every presentation
  – Updated every 2 weeks with the main items you have accomplished
  – Future plans might also be updated if there is a change in plans
Weekly Presentation: Status from Past 2 Weeks

- A status update: what you have done since your last presentation
  - This is the main part of your presentation (particularly in weeks 4, 6, 8)

- This should be a 1 or 2 slide summary
  - Should be consistent with your weekly Google doc summary
  - You do not need to go into every detail of what you did
  - Provide examples, figures, where appropriate

- One possible approach
  - 1st slide provides an overall summary of what you worked on
  - 2nd slide discusses some aspect in more detail (e.g., a graph or figure)
Weekly Presentation: Challenges and Next Steps

• Challenges
  – List any aspects of your project that are causing unexpected problems or issues
  – Feel free to pose questions to the group:
    • “I need to find open-source software to do X – ideas?”
    • “How might I evaluate algorithm Y?”

• Finish by briefly summarizing what you plan to do in the next 2 weeks
  – Should be consistent with what you have listed in your milestones
<table>
<thead>
<tr>
<th>Week</th>
<th>Monday</th>
<th>Wednesday</th>
</tr>
</thead>
<tbody>
<tr>
<td>1, April 2</td>
<td>Short intro lecture</td>
<td>Joint office hours (review of projects/presentations), DBH 2091</td>
</tr>
<tr>
<td>2, April 9</td>
<td>Guest speaker #1</td>
<td>Student progress presentations</td>
</tr>
<tr>
<td>3, April 16</td>
<td>Smyth office hours, DBH 4016</td>
<td>Carey office hours, DBH 2091</td>
</tr>
<tr>
<td>4, April 23</td>
<td>Guest speaker #2</td>
<td>Student progress presentations</td>
</tr>
<tr>
<td>5, April 30</td>
<td>No office hours or class</td>
<td>Carey office hours, DBH 2091</td>
</tr>
<tr>
<td>6, May 7</td>
<td>Guest speaker #3</td>
<td>Student progress presentations</td>
</tr>
<tr>
<td>7, May 14</td>
<td>Smyth office hours, DBH 4016</td>
<td>Carey office hours, DBH 2091</td>
</tr>
<tr>
<td>8, May 21</td>
<td>Guest speaker #4</td>
<td>Student progress presentations</td>
</tr>
<tr>
<td>9, May 28</td>
<td>Memorial Day, no class</td>
<td>Smyth office hours, DBH 4016</td>
</tr>
<tr>
<td>10, June 4</td>
<td>Guest speaker #5</td>
<td>Final student presentations</td>
</tr>
</tbody>
</table>