Data Analysis Final Project Guidelines

Project overview:

One of this course’s main goals is to prepare you to apply state-of-the-art data analysis to an application. The class's final project will be an opportunity for you to test your data analysis skills.

Project topics:

Your first task is to pick a project topic. If you aren’t sure about what topic to choose and would like help you may visit your instructor for possible ideas. Make sure to choose a topic that interests and motivates you since we will be working on this problem throughout the coming months and possibly later.

If you aren’t sure which one to choose then begin with an application project and see where the problem takes you. Many projects will combine elements of applications, algorithms and theory.

A very good project will comprise a publishable or nearly-publishable piece of work. Some students may continue working on their projects after completing this course, and submit their work to a conference or journal.

For inspiration, you might also look at some recent research papers. Two conferences of interest are ICML and NIPS. You can also find papers from recent ICML conferences online:

http://icml.cc/2014/index/article/15.htm
http://jmlr.org/proceedings/papers/v28/.
All NIPS papers are online, at http://books.nips.cc/.

Data sets:

Collecting data is often a long process which requires consistency and care. You may wish to see what data is already available. In particular, there are many free data sets online which you may use. A few of them are as follows:

2. Kaggle: www.kaggle.com
3. KDnuggets: http://www.kdnuggets.com/datasets/
5. StatLib: http://lib.stat.cmu.edu
6. TwitteR: http://cran.r-project.org/web/packages/twitteR/index.html
7. rfigshare: http://cran.r-project.org/web/packages/rfigshare/index.html
9. Wellcome Trust Case Control Consortium: http://www.wtcc.org.uk/
Project Description, Data Set and Scheduling: (5% of Project Grade)

Due: 4/6/2016

Students must form a group of exactly 3 members to work with. If you cannot find a data set before the deadline, one will be assigned to your group. If you cannot find a group, a group will be assigned to you. If you have a group of less than 3, you may be assigned an additional member or reassigned to a different group. If you have a group larger than 3, someone will be chosen at random to leave the group until you have 4 members. Your project description should contain a 4-10 sentence abstract of the project. Additionally, you should describe your data set and the source of the data. Finally, provide me with a time slot in which you would like to do your presentation of the time slots listed. First come first serve on the time slots, however, your team must be formed and a project must be decided before you can be assigned a time slot.

Presentation: (45% of Project Grade)

Presentations will be made during the final week of classes. Each group should plan on being able to present their project within 15 minutes. Also, each member of the group should participate in the presentation. Submit an electronic copy of your presentation on the due date via email to Prof. McGibney. Points will be deducted from teams which are not respectful during all presentations. In particular, teams coming in late and leaving before all teams have presented will have points deducted from their presentation.

Final Paper: (50% of Project Grade)

Your final paper should be 6 pages and written in the ICML template which can be found at https://www.sharelatex.com/templates/journals/icml. Submit the final paper on the due date as a pdf file via email to Prof. McGibney along with your latex and supporting files.

After the presentations, we will post all the final write-ups online so that you can read about each other’s work. If you do not want your write-up to be posted online, then please notify Prof. McGibney at least one week in advance of the final submission deadline.
### Presentation Agenda

<table>
<thead>
<tr>
<th>Date/Time</th>
<th>Group:</th>
</tr>
</thead>
<tbody>
<tr>
<td>April 27, 2015: 9:00-9:15 am</td>
<td></td>
</tr>
<tr>
<td>April 27, 2015: 9:15-9:30 am</td>
<td></td>
</tr>
<tr>
<td>April 27, 2015: 9:30-9:45 am</td>
<td></td>
</tr>
<tr>
<td>April 27, 2015: 11:00-11:15 am</td>
<td></td>
</tr>
<tr>
<td>April 27, 2015: 11:15-11:30 am</td>
<td></td>
</tr>
<tr>
<td>April 27, 2015: 11:30-11:45 am</td>
<td></td>
</tr>
<tr>
<td>April 29, 2015: 9:00-9:15 am</td>
<td></td>
</tr>
<tr>
<td>April 29, 2015: 9:15-9:30 am</td>
<td></td>
</tr>
<tr>
<td>April 29, 2015: 9:30-9:45 am</td>
<td></td>
</tr>
<tr>
<td>April 29, 2015: 11:00-11:15 am</td>
<td></td>
</tr>
<tr>
<td>April 29, 2015: 11:15-11:30 am</td>
<td></td>
</tr>
<tr>
<td>April 29, 2015: 11:30-11:45 am</td>
<td></td>
</tr>
<tr>
<td>May 10, 2015: 9:00-9:15 am</td>
<td></td>
</tr>
<tr>
<td>May 10, 2015: 9:15-9:30 am</td>
<td></td>
</tr>
<tr>
<td>May 10, 2015: 9:30-9:45 am</td>
<td></td>
</tr>
<tr>
<td>May 10, 2015: 10:00-10:15 am</td>
<td></td>
</tr>
<tr>
<td>May 10, 2015: 10:15-10:30 am</td>
<td></td>
</tr>
<tr>
<td>May 10, 2015: 10:30-10:45 am</td>
<td></td>
</tr>
<tr>
<td>May 10, 2015: 11:00-11:15 am</td>
<td></td>
</tr>
<tr>
<td>May 10, 2015: 11:15-11:30 am</td>
<td></td>
</tr>
<tr>
<td>May 10, 2015: 11:30-11:45 am</td>
<td></td>
</tr>
</tbody>
</table>

### Project Due Date

Each project (paper and code) is due at 11:59pm on the date of the group’s presentation (either April 27, April 29 or May 10).

### Grading Rubric for Presentation

<table>
<thead>
<tr>
<th>Category</th>
<th>Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motivation – Did students provide adequate motivation for what they are doing?</td>
<td>20</td>
</tr>
<tr>
<td>Methodology – Was the primary (best performing) method adequately explained? Is there an obvious better way to do the problem?</td>
<td>30</td>
</tr>
<tr>
<td>Results – Are the results explained adequately? Comparison?</td>
<td>10</td>
</tr>
<tr>
<td>Use of visual aids – How well are the slides prepared?</td>
<td>20</td>
</tr>
<tr>
<td>How well were questions answered?</td>
<td>10</td>
</tr>
<tr>
<td>Was the presentation handed in on time? (send me a copy of your slides at 8pm the day before your presentation)</td>
<td>10</td>
</tr>
</tbody>
</table>
### Presentation Penalties

<table>
<thead>
<tr>
<th>Category</th>
<th>Penalty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Timing – Is presentation between 12 and 15 minutes?</td>
<td>-5</td>
</tr>
<tr>
<td>Team effort – Did all members participate equally in presentation?</td>
<td>-5</td>
</tr>
<tr>
<td>Respectful to other teams – Inappropriate talking/noise while another team is presenting. Arriving/leaving midway through another group’s presentation.</td>
<td>-5</td>
</tr>
</tbody>
</table>

### General Grading Rubric for Paper

<table>
<thead>
<tr>
<th>Category</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Formatting – Is the format of the paper appealing and nice to look at?</td>
<td>5</td>
</tr>
<tr>
<td>Motivation – Did you convince the reader that you are solving an interesting problem? You should explain why your problem is important.</td>
<td>10</td>
</tr>
<tr>
<td>Features – Were your features explained well? Include equations when applicable.</td>
<td>10</td>
</tr>
<tr>
<td>Methodology – Was the primary (best performing) method adequately explained? Why does the primary method work well? What did you do aside from the standard methodology?</td>
<td>10</td>
</tr>
<tr>
<td>Quantity of Results – How many methods were implemented correctly? You will get 5 points for each method you correctly use (up to 20 points total).</td>
<td>20</td>
</tr>
<tr>
<td>Quality of Results – How well did you do? Particularly, how well did you do on Kaggle or what is your overall performance?</td>
<td>10</td>
</tr>
<tr>
<td>Use of graphs – Use at least three graphs, discuss and reference them. In place of one graph you may use a table. In place of one graph you may use an algorithm description.</td>
<td>15</td>
</tr>
<tr>
<td>Citations – Are citations made properly and referenced within the paper?</td>
<td>5</td>
</tr>
<tr>
<td>How well is your code written and documented? (don’t forget to hand a copy of your code in)</td>
<td>15</td>
</tr>
</tbody>
</table>