Assignment 2

• Open-ended
• Due **June 2** (four weeks from two days ago)
• Submissions should be made electronically to Long Jin (longjin@cs.ucsd.edu)
Basic tasks:

1. Identify a dataset to study and describe its basic properties
2. Identify a predictive task on this dataset and describe the features that will be relevant to it
3. Describe literature & research relevant to the dataset and task
4. Describe and analyze results
Evaluation

• E.g. about this much:

(acm proceedings format)

https://www.acm.org/sigs/publications/proceedings-templates
Assignment 2

Teams of four
1. **Identify a dataset to study**

   - Beer data
     (http://snap.stanford.edu/data/Ratebeer.txt.gz
     http://snap.stanford.edu/data/Beeradvocate.txt.gz)

   - Wine data
     (http://snap.stanford.edu/data/cellartracker.txt.gz)

   - Google Local (Maps & Restaurants)
     (http://jmcauley.ucsd.edu/data/googlelocal.tar.gz - warning: kind of huge)

   - Sensor data
     (https://github.com/rpasricha/MetroInsightDataset)
1. Identify a dataset to study
   
   • Reddit submissions
     (http://snap.stanford.edu/data/web-Reddit.html)
   
   • Facebook/twitter/Google+ communities
     (http://snap.stanford.edu/data/egonets-Facebook.html
     http://snap.stanford.edu/data/egonets-Gplus.html
     http://snap.stanford.edu/data/egonets-Twitter.html)
   
   • Many many more from other sources, e.g.
     http://snap.stanford.edu/data/

   Use whatever you like, as long as it’s **big**
   (e.g. 50,000 datapoints minimum)
1b: Perform an **exploratory analysis** on this dataset to identify interesting phenomena

- Start with basic results, e.g. for a recommender systems type task, how many users/items/entries are there, what is the overall distribution of ratings, what time period does the dataset cover etc.
1b: Perform an **exploratory analysis** of this dataset to identify interesting phenomena.
2. Identify a **predictive task** on this dataset

- How will you evaluate the model? Which models from class are relevant to your predictive task, and why are other models inappropriate?
- What are the relevant baselines that can be compared?
- How will you assess the validity of your predictions and confirm that they are significant?
- It’s totally fine here to implement a model that we covered in class, e.g. for a classification task you could implement svms+logistic regression+naïve Bayes
- You should also compare the results of different feature representations to identify which ones are effective
- Did you have to do pre-processing of your data in order to obtain useful features?
- How do the results of your exploratory analysis justify the features you have chosen?
3. Describe related literature

- If you used an existing dataset, where did it come from and how was it used there?
- What other similar datasets have been used in the past and how?
- What are the state-of-the-art methods for the prediction task you are considering? Were you able to borrow any ideas from these works for your model? What features did they use and are you able to use the same ones?
- What were the main conclusions from the literature and how do they differ from/compare to your own findings?
4. Describe your results

- If you used a complex model, how did you optimize it?
- What issues did you face scaling it up to the required size?
- Any issues overfitting?
- Any issues due to noise/missing data etc.?
- Of the different models you considered, which of them worked and which of them did not?
- What is the interpretation of the parameters in your model? Which features ended up being predictive? Can you draw any interesting conclusions from the fitted parameters?
Example
Maybe I want to use restaurant data to build a model of people’s tastes in different locations
(http://jmcauley.ucsd.edu/data/googlelocal.tar.gz)
1. Perform an exploratory analysis of this dataset to identify interesting phenomena

• How many users/items/ratings are there? Which are the most/least popular items and categories?
• What is the geographical spread of users, items, and ratings?
• Do people give higher/lower ratings to more expensive items, or items in certain countries/locations?
2. Identify a **predictive task** on this dataset

- Predict what rating a person will give to a business based on the time of year, the past ratings of the user, and the geographical coordinates of the business.
- Predict which businesses will succeed or fail based on its geographical location, or based on its early reviews.
- What model/s and tools from class will be appropriate for this task or suitable for comparison? Are there any other tools *not* covered in class that may be appropriate?
2b. Identify features that will be relevant to the task at hand

- Ratings, users, geolocations, time
- Ratings as a function of price
- Ratings as a function of location
  - How to represent location in a model? Just using a linear predictor of latitude/longitude isn’t going to work...
3. Describe related literature

- Relevant literature or predicting ratings
- Literature on using geographical features for various predictive tasks
- Literature on predicting long-term outcomes from time series data
- Literature on predicting future ratings from early reviews, herding etc.
4. Describe results and conclusions

• Did features based on geographical information help? If not why not?
• Which locations are the most price sensitive according to your predictor?
• Do people prefer restaurants that are unlike anything in their area, or restaurants which are exactly the same as others in their area?
Example 2
Maybe I want to use reddit data to see what makes submissions successful
(http://snap.stanford.edu/data/web-Reddit.html)
1. Perform an **exploratory analysis** of this dataset to identify interesting phenomena

   - How many users/submissions are there? How does activity differ across subreddits?
   - What times of day are submissions most commented on or most rated?
   - Do people give more/fewer votes to submissions that have long/short titles, or which use certain words?
Assignment 2

2. Identify a **predictive task** on this dataset

- Predict whether a post will have a large number of comments or a high rating
- Predict whether there will be a large *discrepancy* between the number of comments and the positivity of ratings a post receives
- What model/s and tools from class will be appropriate for this task or suitable for comparison? Are there any other tools *not* covered in class that may be appropriate?
2b. Identify features that will be relevant to the task at hand

- Votes, users, subreddits, time
- Resubmissions of the same content & the success or failure of previous submissions
- Text of the post title
3. Describe related literature

- Relevant literature or predicting votes on Reddit
- Literature on virality in social media
- Literature on using text for predictive tasks
- Literature on temporal forecasting or user preference modeling
4. Describe results and conclusions

• What features helped you to predict whether content would be controversial or not?
• Does the text of the title help to predict whether a submission will be controversial or get many comments but a low vote?
• Which subreddits generate more controversial content than others?
More examples

A similar type of project from Stanford’s "Social and Information Network Analysis" course:

http://snap.stanford.edu/class/cs224w-2013/projects.html
More examples

Last quarter’s graduate course (cse 255)

http://cseweb.ucsd.edu/~jmcauley/cse255/projects/
Evaluation

- These 4 sections will be worth (roughly) 5 marks each (for a total of 20% of your grade)
- Assignments can be done in groups of up to 3. The marking scheme is the same regardless of group size.
- Length is not strict, but should be about 4 pages in small-font double-column format.

- Groups of 4?
Assignment 2

Evaluation

- E.g. about this much:

(acm proceedings format)

https://www.acm.org/sigs/publications/proceedings-templates
CSE 190
Data Mining and Predictive Analytics

Assignment 2 – examples of assignments from 255
Assignement 2

Raw rating data

binned regression

dual regression

"inflection" point

Andrew Prudhomme – “Finding the Optimal Age of Wine”
Assignment 2

ratings vs. time

ratings vs. review length
Assignment 2

positive words in wine reviews
negative words in wine reviews

positive words in beer reviews
negative words in wine reviews

Ben Braun & Robert Timpe – “Text-based rating predictions from beer and wine reviews”
Assignment 2

Figure 3: Restaurant Ratings

Figure 6: Average rating per location

ratings per location

k-means of ratings per location

4.00

3.52

Diego Cedillo & Idan Izhaki – “User Score for Restaurants Recommendation System”
Assignment 2

Long Jin & Xinchi Gu – “Rating Prediction for Google Local Data”

\[ \hat{r}_{ui} = \mu + b_u + b_i + (q_i + \frac{1}{|M(i)|} \sum_{n \in M(i)} |s_n|) T_p_u \]

set of geographic neighbours  impact of neighbours
Topic model from Google Local business reviews
Assignement 2

Wikispeededia navigation traces:

Figure 5: Graph of a complete path

<table>
<thead>
<tr>
<th></th>
<th>Average Click</th>
<th>Average Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Finish Path</td>
<td>4.72</td>
<td>158.27</td>
</tr>
<tr>
<td>Finished Path Back</td>
<td>6.75</td>
<td>158.31</td>
</tr>
<tr>
<td>Unfinished Path</td>
<td>2.97</td>
<td>835.29</td>
</tr>
<tr>
<td>Unfinished Path Back</td>
<td>5.2</td>
<td>836.00</td>
</tr>
</tbody>
</table>
Images from Chictopia

Power laws!

Wei-Tang Liao & Jong-Chyi Su – “Image Popularity Prediction on Social Networks”
Questions?