CSE 255

Data Mining and Predictive Analytics

Assignment 1

Assignment 1

- Open-ended
- Due **Feb 23** (four weeks from today)
- Submissions should be made electronically to Dongcai (doshen@cs.ucsd.edu)

Assignment 1

- Open-ended
- Due **Feb 23** (four weeks from today)
- Submissions should be made electronically to Dongcai (doshen@cs.ucsd.edu)

Assignment 1

Basic tasks:

1. Identify a dataset to study
2. Identify a predictive task on this dataset
3. Describe literature relevant to the task
4. Identify features that will be relevant to the prediction task at hand
5. Develop a model for the task and run experiments
6. Describe results and conclusions

Assignment 1

1. Identify a dataset to study

- Amazon data

(http://snap.stanford.edu/data/web-Amazon-links.html)

- 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)

- Reddit submissions

(http://snap.stanford.edu/data/web-Reddit.html)

Assignment 1

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.

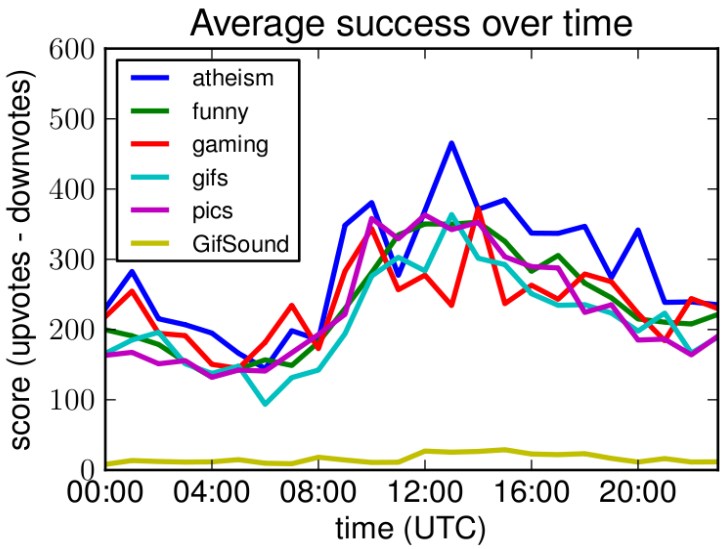
http://snap.stanford.edu/data/

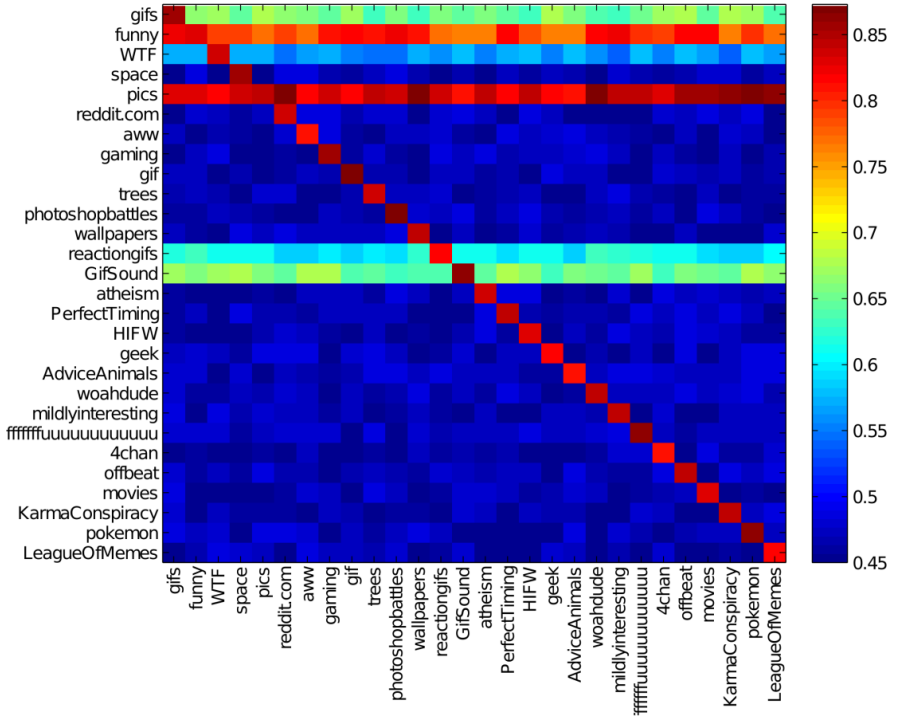
Use whatever you like, as long as it's **big** (e.g. 50,000 datapoints minimum)

Assignment 1

1b: Perform an **exploratory analysis** of this dataset to identify interesting phenomena

e.g.





Assignment 1

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

- How will you evaluate the model?
- What are the relevant baselines that can be compared?
- How will you assess the validity of your predictions and confirm that they are significant?

Assignment 1

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? Are any of them suitable to implement for comparison?

Assignment 1

4. Identify features that will be relevant to the task at hand

- Why do you expect the chosen features to be useful for prediction?
- Your exploratory analysis of the data should justify the features you have selected
- What pre-processing of the data was necessary to select or manipulate the features?

Assignment 1

5. Describe your model

- How will you optimize it?
  - What issues did you face scaling it up to the required size?
  - Any issues overfitting?
- What other models did you consider besides the one you proposed (and what were your unsuccessful attempts before arriving at the right model)?
- What are the strengths and weaknesses of the different models being compared?

Assignment 1

6. Describe results and conclusions

- How well did your model perform compared to alternatives?
- What is the significance of the results?
- Are they robust to noise in the data, mislabeled examples etc.?
- What is the interpretation of the parameters in your model? Which features ended up being predictive?
- Why did the proposed model succeed while others failed?

Assignment 1

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)

Assignment 1

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?

Assignment 1

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

Assignment 1

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.

Assignment 1

4. 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…

Assignment 1

5. Describe your model

- E.g. Adapt collaborative filtering techniques to include a geographic regularizer
- Adapt long-term forecasting techniques to make use of user and rating information
- Analyze the text of people's reviews to predict linguistic signals of popular and successful businesses

Assignment 1

6. Describe results and conclusions

- Did geographical information help? If not why not?
- Which locations are the most price sensitive?
- Do people prefer restaurants that are unlike anything in their area, or restaurants which are exactly the same as others in their area?

Assignment 1

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

Assignment 1

Evaluation

- These 6 sections will be worth (roughly) 5 out of 30 percent each
- Not all sections will be relevant for all assignments so there is some flexibility, but **be reasonable**
- Assignments can be done **individually or in pairs**, though if done in pairs the expected contribution should be larger

- Length is not strict, but I'd expect a report of about 6-10 pages (more like 6 for individuals, more like 10 for pairs)
- This probably adds up to 3-5k words (plus figures tables, equations etc.)