Web Mining and Recommender Systems

Assignment 2
Assignment 2

- Open-ended
- Due Dec 7 (Monday)
- Submissions should be made via gradescope
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 what model/s you will use to solve this task
4. Describe literature & research relevant to the dataset and task
5. Describe and analyze results
Teams of one to four
1. Identify a dataset to study

   • My own repository of Recommender Systems datasets:
   • [https://cseweb.ucsd.edu/~jmcauley/datasets.html](https://cseweb.ucsd.edu/~jmcauley/datasets.html)
1. **Identify a dataset to study**
   - Beer data
     - [Ratebeer.txt.gz](http://snap.stanford.edu/data/Ratebeer.txt.gz)
     - [Beeradvocate.txt.gz](http://snap.stanford.edu/data/Beeradvocate.txt.gz)
   - Wine data
     - [Cellartracker.txt.gz](http://snap.stanford.edu/data/cellartracker.txt.gz)
   - Exercise data
     - [Fitrec-project/home](https://sites.google.com/eng.ucsd.edu/fitrec-project/home)
   - Steam purchases
     - [Steam_data](https://cseweb.ucsd.edu/~jmcauley/datasets.html#steam_data)
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.
Assignment 2

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

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

- How will you assess the validity of your predictions and confirm that they are significant?
- 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. Select/design an appropriate model

- How will you evaluate the model? Which models from class are relevant to your predictive task, and why are other models inappropriate?
- 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
- What are the relevant baselines that can be compared?
- 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.?
4. 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?
5. Describe your results

• 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
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. Select an appropriate model

• Some kind of latent-factor model

• How to incorporate the geographical term? Should we cluster locations? Use the location as a regularizer? (etc.)

• How can we optimize this (presumably complicated) model?
4. 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.
5. 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?
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. Select an appropriate model

- Some kind of regression
- Need to use gradient descent or is there a closed-form solution?
- What are the hyperparameters and how do we regularize?
- How can you incorporate the temporal terms?
4. 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
5. 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?
Evaluation

- These 5 sections will be worth (roughly) 5 marks each (for a total of 25% of your grade)
- Assignments can be done in groups of up to 4. 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.
Assignment 2

Evaluation

• E.g. about this much:

(acm proceedings format)
https://www.acm.org/sigs/publications/proceedings-templates
Web Mining and Recommender Systems

Examples of previous assignments
Supervised funniness detection in the New Yorker cartoon caption contest

- Predict whether a caption will be scored as “funny” by human judges
- 65 images, 320k captions
- Scores from 1.0 – 2.75

- BoW methods w/ and w/o TF-IDF
- Dimensionality-reduction-based feature representations
Predicting Vegetation Changes as Responses to Forest Fires

- Geological data from LANDFIRE program and FRAP (Fire and Resource Assessment Program), 1992-2012
- Estimate changes as a result of forest fires

\[ y = x_{2012\text{ vegetation}} = x_{2014\text{ vegetation}} \quad \forall x \in X \]
AirBnB Price Per Night Prediction

- AirBnB Paris data
- Predict listing price given various features
Anonymized Uber Movement data from 7 cities

Trip time given source, destination, and hour

Weekday travel times in two cities

<table>
<thead>
<tr>
<th>Feature Representation</th>
<th>Week Category</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>hod, source ID, dest ID</td>
<td>Weekday</td>
<td>26.544%</td>
</tr>
<tr>
<td></td>
<td>Weekend</td>
<td>29.247%</td>
</tr>
<tr>
<td>hod mean, source ID mean, dest ID mean</td>
<td>Weekday</td>
<td>26.788%</td>
</tr>
<tr>
<td></td>
<td>Weekend</td>
<td>29.113%</td>
</tr>
<tr>
<td>hod, source ID, dest ID, hod mean, source ID mean, dest ID mean, combined source ID-dest ID mean</td>
<td>Weekday</td>
<td>21.318%</td>
</tr>
<tr>
<td></td>
<td>Weekend</td>
<td>25.024%</td>
</tr>
<tr>
<td>hod, combined source ID-dest ID mean</td>
<td>Weekday</td>
<td>79.218% / 79.975%*</td>
</tr>
<tr>
<td></td>
<td>Weekend</td>
<td>87.041% / 87.146%*</td>
</tr>
</tbody>
</table>

SVM, Random Forest, MLP

Tynan Dewes, David Thomson
Predicting the Accepted Answer for StackOverflow Questions

- Large dataset of StackOverflow posts
- Predict which answer is marked as “accepted” (classification)
Does historical Bitcoin data contain enough information to predict its future value ("autoregression"-like task)
Wine demand appears to exhibit seasonal variability. Can this be predicted?

Consumption of “high quality” wine is seasonal.

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>random selection</td>
<td>0.25</td>
</tr>
<tr>
<td>pick most popular</td>
<td>0.714</td>
</tr>
<tr>
<td>k-nearest neighbor</td>
<td>0.786</td>
</tr>
</tbody>
</table>
## Duplicate Question Detection on Quora

<table>
<thead>
<tr>
<th>Question1</th>
<th>Question2</th>
<th>label</th>
</tr>
</thead>
<tbody>
<tr>
<td>What can make Physics easy to learn?</td>
<td>How can you make physics easy to learn?</td>
<td>1</td>
</tr>
<tr>
<td>What’s causing someone to be jealous?</td>
<td>What can I do to avoid being jealous of someone?</td>
<td>0</td>
</tr>
</tbody>
</table>

### Results

<table>
<thead>
<tr>
<th>Type</th>
<th>Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cosine</td>
<td>Cosine TF-IDF</td>
<td>0.6400</td>
</tr>
<tr>
<td></td>
<td>Cosine topic vector</td>
<td>0.5926</td>
</tr>
<tr>
<td>Traditional</td>
<td>LR</td>
<td>0.6405</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.6887</td>
</tr>
<tr>
<td></td>
<td>Decision Tree</td>
<td>0.6828</td>
</tr>
<tr>
<td></td>
<td>KNN</td>
<td>0.6769</td>
</tr>
<tr>
<td>Ensemble</td>
<td>RF</td>
<td>0.7032</td>
</tr>
<tr>
<td></td>
<td>GBDT</td>
<td>0.7015</td>
</tr>
<tr>
<td></td>
<td>Adaboost</td>
<td>0.6861</td>
</tr>
<tr>
<td>Deep model</td>
<td>Siamese LSTM</td>
<td><strong>0.7754</strong></td>
</tr>
</tbody>
</table>

Figure 5: LSTM-based feature extractor followed by handcrafted feature extraction

Table 2: Comparative evaluation of all models

<table>
<thead>
<tr>
<th>Model</th>
<th>Log-Loss</th>
<th>Accuracy(%)</th>
<th>auc</th>
<th>AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF-IDF + Cosine Distance</td>
<td>NA</td>
<td>62.9</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>TF-IDF + XGBoost</td>
<td>0.48</td>
<td>73.66</td>
<td>0.78</td>
<td>0.69</td>
</tr>
<tr>
<td>LSTM + DNN</td>
<td>0.39</td>
<td>83.6</td>
<td>0.89</td>
<td>0.83</td>
</tr>
<tr>
<td>LSTM + XGBoost</td>
<td>0.38</td>
<td>84.15</td>
<td>0.90</td>
<td>0.83</td>
</tr>
<tr>
<td>LSTM + Handcrafted features</td>
<td>0.46</td>
<td>79.0</td>
<td>0.84</td>
<td>0.82</td>
</tr>
<tr>
<td>Ensemble</td>
<td>0.37</td>
<td><strong>84.73</strong></td>
<td><strong>0.90</strong></td>
<td><strong>0.82</strong></td>
</tr>
</tbody>
</table>

Yi Luo, Jingtao Song, Haoting Chen

Vaibhav Gandhi, Akshaya Purohit, Aditya Verma
NYC Taxi Demand Prediction

Feature Importance

- Temperature
- Hour
- Income
- Population

Siyu Jiang, Simran Kapur, Siddharth Dinesh
NYC Bike Trip Duration Prediction

<table>
<thead>
<tr>
<th>Variate</th>
<th>Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trip Duration</td>
<td>in seconds format</td>
</tr>
<tr>
<td>Start Time and Date</td>
<td>Timestamp</td>
</tr>
<tr>
<td>Stop Time and Date</td>
<td>Timestamp</td>
</tr>
<tr>
<td>Start Station Name</td>
<td>String</td>
</tr>
<tr>
<td>End Station Name</td>
<td>String</td>
</tr>
<tr>
<td>Station ID</td>
<td>Number</td>
</tr>
<tr>
<td>Station Lat/Long</td>
<td>Number</td>
</tr>
<tr>
<td>Bike ID</td>
<td>Number</td>
</tr>
<tr>
<td>User Type</td>
<td>Customer or Subscriber</td>
</tr>
<tr>
<td>Gender</td>
<td>Number</td>
</tr>
<tr>
<td>Year of Birth</td>
<td>Number</td>
</tr>
</tbody>
</table>

Bike Share Growth in the US

<table>
<thead>
<tr>
<th>Model</th>
<th>FVU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>1.000006</td>
</tr>
<tr>
<td>Linear Regression</td>
<td>0.211735</td>
</tr>
<tr>
<td>Ridge Regression</td>
<td>0.211591</td>
</tr>
<tr>
<td>Random Forest Regressor</td>
<td>0.205021</td>
</tr>
<tr>
<td>XGBoost Regressor</td>
<td>0.195970</td>
</tr>
<tr>
<td>Ensemble of Random Forest and XGBoost</td>
<td>0.200575</td>
</tr>
</tbody>
</table>

subscriber vs. customer
duration vs. gender

Zhuo Cheng, Tianran Zhang, Jiamin He
Airline Delay Prediction

- **KNN, SVM, Softmax regression**

<table>
<thead>
<tr>
<th>Methods</th>
<th>AUC scores</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 score</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.298</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>0.629</td>
<td>0.3049</td>
<td>0.4044</td>
<td>0.3467</td>
<td>0.692</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>0.6492</td>
<td>0.3478</td>
<td>0.34</td>
<td>0.3367</td>
<td>0.7345</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.6129</td>
<td>0.2441</td>
<td>0.0074</td>
<td>0.0140</td>
<td>0.7975</td>
</tr>
<tr>
<td>Neural Network</td>
<td>0.6404</td>
<td>0.5218</td>
<td>0.0677</td>
<td>0.1150</td>
<td>0.7946</td>
</tr>
</tbody>
</table>

- **Delay vs. Origin**
- **Delay vs. Route**
- **Delay vs. Departure Time**
Questions?