CIS520 Final Project
2014

Lyle Ungar
Data Set: Real Estate Ads

• Ads from 7 cities
  • (Boston, Chicago, LA, Miami, NYC, Philly, Vegas)

399000 stunning skyline views like something from a postcard are yours with this large 2 bedroom, 2 bath loft in Dearborn tower! Detailed hardwood floors throughout the unit complement an open kitchen and spacious living room and dining room. Huge walk-in closet, steam shower and marble entry. Parking available.

250000 large corner lot features landscaping & nice pool w/ waterfall. Large kitchen w/ island, granite counter tops, custom cabinets, ceramic tile flooring opens to family room w/ fp. Oversized master suite w/ 3way fireplace, sitting room & balcony. Potshelves, central vac, 1st floor office/den - could be bedroom.
The task and data

• Predict the price from the words in the ad and the city
  • Performance assessed by Root Mean Squared Error (RMSE)

• We will provide
  • the original ads (tokenized plain text as shown above) with mappings from the words and bigrams to integers
  • Four files
    — Prices (n*1) an integer = log(price)
    — Cities (n*7) a unit vector (exactly one 1, the rest zeros)
    — Words (n*v) a sparse matrix indicating if each of the v most frequent words is present in the given ad
    — Bigams (n*v) a sparse matrix indicating if each of the v most frequent bigrams is present in the given ad

• \( N = 20,311 \) \( v = 5,000 \)
Why is this interesting?

• Finding properties of people or products from words is very popular now
  • And words are nice features so you can look at them and see if they make sense
• Small enough to no require too much computation
• but p > n so it is hard
• And includes different kinds of features
  • But not so many as to take too much time.
• Regression setting is good for regularization/feature selection
  • Allows for some creativity and exploration
Leaderboard will look similar to:

<table>
<thead>
<tr>
<th>Group Name</th>
<th>Time Submitted</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unconscious Bayes</td>
<td>2012-11-30T23:26:56.246532</td>
<td>78.46%</td>
</tr>
<tr>
<td>musiclab</td>
<td>2012-11-30T23:08:02.232505</td>
<td>75.81%</td>
</tr>
<tr>
<td>Music Up</td>
<td>2012-11-30T23:05:02.659313</td>
<td>75.39%</td>
</tr>
<tr>
<td>pythia</td>
<td>2012-11-30T22:57:12.343108</td>
<td>75.39%</td>
</tr>
<tr>
<td>J.R.S. (Decisions Of Care Taken Over by Robotic Surrogates)</td>
<td>2012-11-30T22:33:51.776995</td>
<td>75.39%</td>
</tr>
<tr>
<td>Pandora's Black Box</td>
<td>2012-11-30T22:46:43.721001</td>
<td>74.81%</td>
</tr>
<tr>
<td>Genre Safety</td>
<td>2012-11-30T22:05:19.538428</td>
<td>73.49%</td>
</tr>
<tr>
<td>3dumplings</td>
<td>2012-11-30T19:37:53.039127</td>
<td>73.41%</td>
</tr>
<tr>
<td>Good Luck</td>
<td>2012-11-30T23:04:38.103531</td>
<td>72.66%</td>
</tr>
<tr>
<td>The Aggressive Regressors</td>
<td>2012-11-30T23:43:16.954605</td>
<td>74.32%</td>
</tr>
</tbody>
</table>

Figure: Fall 2012 leaderboard
Requirements

- **Must work in groups of 2-3**
- **Must beat threshold (ridge regression)**
  - On test data, of course!
- **Implement 4 different learners from among:**
  - generative method
  - discriminative method
  - instance based method
  - your own kernel
  - your own regularization method
  - Semi-supervised dimensionality reduction
- **Write a final report (2-5) pages**
  - results and analysis for all methods tried
  - data visualization
- **Extra credit and prizes for top 10 groups**
Submission details

- **Group name submission:**
  ```
  turnin -c cis520 -p proj_groups group.txt
  ```

- **Quiz (Leaderboard) submission:**
  ```
  turnin -c cis520 -p leaderboard submit.txt
  ```

- **Final submission:**
  ```
  turnin -c cis520 -p project <list of files including make_final_prediction.m>
  ```
Deadlines

- **Nov. 20** Submit group.txt – group name (1%)
- **Nov. 22** Beat 1st baseline RMSE=0.8540 (9%)
- **Dec. 3** Beat 2nd baseline TBA (20%)
- **Dec. 6** Submit final prediction code (50%)
- **Dec. 11** Submit final report (20%)
Rules and Restrictions

• Groups must be of size 2 or 3!
• Do not obtain additional data! Harsh penalty!
  • i.e. no scraping more real estate ads
• < 50 Mb code submission size
• 64-bit linux Matlab R2014a compatible
• < 10 minute run time to test 5,000 test samples
• Leaderboard updated every 5 hours
• You are allowed to use third party matlab code
  • But cite it
Other info

- Special prizes to the most creative names

- We will post more info on the wiki on Monday

- Leaderboard will be at
  http://www.seas.upenn.edu/~cis520/fall14/leaderboard.html
GOOD
LUCK