Predicting Business Ratings on Yelp

Travis Gingerich, Yevhen Bochkov

{travisg, ebochkov}@stanford.edu

Motivation

- Recommender systems are vital for many modern services
- Good personalized recommendation add another dimension to the user experience, enhancing user satisfactory and loyalty.
- A common strategy is to recommend item based on item rating prediction.
- We would like to explore how modern recommender techniques can be applied to business ratings in the Yelp dataset

Problem Formulation

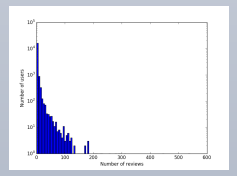
- Individual users' ratings of a business can be expressed in matrix form
- The rating matrix $R \in \mathbb{R}^{m \times n}$ consists of known ratings , $r_{ui}, (u, i) \in K$.
- The goal is to give an estimation, $\hat{r}_{ui}$, of unknown ratings, such that $(u, i) \notin K$.

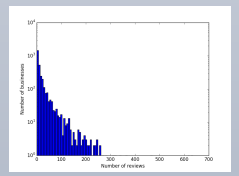
Dataset Description

- The Yelp dataset contain a wealth of information about users, businesses, and ratings, including:
  - Business reviews
    - Rating (on a scale from 1-5)
    - Review text
  - Business information
    - Business location
    - Business categories (business type, cuisine type, etc.)
    - Various business attributes (location, name, WiFi availability, ambience, etc.)
  - User information
    - User friendship connections
    - Length of membership]
- Data is provided in JSON format. An example of the review schema is below:

```
{ 'type': 'review',
 'business_id': (encrypted business id),
 'user_id': (encrypted user id),
 'stars': (star rating, rounded to half-stars),
 'text': (review text),
 'date': (date, formatted like '2012-03-14'), '
votes': {(vote type): (count)}}
```

- A key feature of the data is sparseness of reviews:



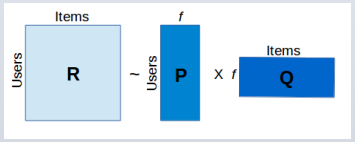


Data Selection

- We used a subset of the Yelp Dataset Challenge data to perform our analysis to allow quicker training of the models and allow easier development and experimentation
- In order to ensure reasonable overlap between users and businesses, we chose a subset of reviews by geographic location, taking the set of all reviews in the state of Pennsylvania

Methods

- Baseline implementation: $\hat{r}_{ui} = \mu + b_u + b_i$
  - Simply predicts ratings by adding the average overall rating and accounting for biases present in the ratings of both individual users and individual items
- Matrix factorization is a technique used by many recommender systems; the user-item rating matrix is approximated by the product of two lower-rank matrices representing latent factors corresponding to individual users and items



- Basic matrix factorization predicts the rating via the formula $\hat{r}_{ui} = \mu + b_u + b_i + p_u^T q_i$
- Data sparseness makes matrix factorization prone to overfitting; regularization is often used to mitigate this
- In addition to only including latent factors, additional information can be incorporated into the matrix factorization model:

$$\hat{r}_{ui} = \mu + (\sum_j b_j^{(g)} \gamma_j + \sum_j b_j^{(u)} \alpha_j + \sum_j b_j^{(i)} \beta_j) + (\sum_j p_j \alpha_j)^T (\sum_j q_j \beta_j)$$

- The b terms represent bias, and γ, α, and β terms represent global, user, and business features, respectively.
- This model can implement basic matrix factorization, as well as more complex models including additional features.
- The objective is to minimize mean squared error. To solve the minimization problem, stochastic gradient descent is used.
- Standard matrix factorization as well as matrix factorization incorporating additional information about the business' categories is used.

Evaluation

- Root mean squared error is used to identify the minimization objective as well as to evaluate results
- RMSE was evaluated separately for groups of reviews, depending on the number of known ratings by the user and for the business for which the rating is being predicted
- The training/test set was split in an 80%/20% fashion on a per-user basis. This ensures that evaluation of the classifier on users and businesses with low numbers of reviews is valid.

Results

- The baseline model does surprisingly well; bias accounts for a large portion of ratings, and therefore must be correctly modeled in most predictors
- Without care, matrix factorization methods seem to be very prone to over-fitting the data; regularization and limiting iterations prevents this. Adding business features was very prone to introducing overfitting.

Future Work

- One potential source of information we would like to include in our model is information regarding social connections between users
- "Social Regularization" imposes a penalty on the deviation of scores between friends; alternatively, shared user features could be added to the current representation to describe friendship connections
- Further investigation into the effectiveness of including additional business features in the model should be performed.