Statistical Learning Theory LAB

Project Report

Maurice Wenig

Contents

1 Algorithms

1.1 Neighbourhood-Based Recommenders

1.1.1 User-Based Recommender

The user-based recommender compares users based on their rated items. An item rating from a user is then predicted based on ratings of that item from similar users. The similarity measure used to compare users is Pearson:

$$\mathrm{Sim}(u,v) = \frac{\sum\limits_{i \in I_{uv}} w_i \cdot s_{ui} s_{vi}}{\sqrt{\sum\limits_{i \in I_{uv}} w_i \cdot s_{ui}} \cdot \sqrt{\sum\limits_{i \in I_{uv}} w_i \cdot z_{vi}}}$$

where I_{uv} are the items that both users u and v have rated, $s_{ui} = r_{ui} - \mu_u$ is the centered rating from user u. The mean rating μ_u can be calculated individually for every comparison, based on only the common items, or it can be calculated once for every user, based on all their rated items. $w_i = \log \frac{\#\text{ user}}{\#\text{ users who rated } i}$ are item weights to combat the impact of the long tail in the number of item ratings. To prefer users with more common items, the similarites of users with less than β common items are made smaller.

$$\mathrm{Sim}(u,v) \leftarrow \mathrm{Sim}(u,v) \cdot \frac{\min\{|I_{uv}|, \beta\}}{\beta}$$

Based on these similarities, a peer group $P_u(i)$ of users can be determined. For this, the users are sorted by similarity. To improve efficiency in the online phase, the sorting is done in the offline phase. In the online phase, this order is simply applied. Users who have not rated the item i and users whos similarity does not exceed a certain threshold, are then removed. The top k remaining users are then the peer group $P_u(i)$.

The predicted item rating from a user is then calculated as a sum of ratings from similar users, weighted with their relative similarity. The means and variances of the user ratings are normalized to not influence the prediction.

$$\hat{r}_{ui} = \mu_u + \sigma_u \cdot \frac{\sum\limits_{v \in P_u(i)} \mathrm{Sim}(u,v) \cdot z_{vi}}{\sum\limits_{v \in P_u(i)} |\mathrm{Sim}(u,v)|}$$

where $z_{ui} = s_{ui}/\sigma_u$ is the normalized rating, σ_u the variance of the ratings of user u. If the user has only ever given one distinct rating, the variance is set to an arbitrary value of 1.

In special cases where the user or the item has never been seen before, similarities can not be calculated. In those special cases, the ratings were predicted in a different way:

- user and item unknown: predict average between minimum and maximum possible rating
- user unknown, item known: predict average rating of that item
- user known, item unknown: predict average rating of that user

1.1.2 Item-Based Recommender

The item-based recommender compares items based on their user ratings. An item rating from a user is then predicted based on ratings of similar items from that user. It functions very similar to the user-based recommender. The similarity function is adapted to still use user means μ_u instead of item means μ_i to normalize user preferences. The predicted rating is then a weighted sum of similar items, the user has rated.

$$\hat{r}_{ui} = \frac{\sum\limits_{j \in P_i(u)} \mathrm{Sim}(i,j) \cdot r_{uj}}{\sum\limits_{j \in P_i(u)} |\mathrm{Sim}(i,j)|}$$

1.1.3 Clustering

To increase efficiency, the users/items can be clustered before the creation of the similarity matrix. This way, only users/items that are in the same cluster have to be compared.

This is done using an adapted version of K-means. To account for the sparsity of the rankings matrix, the entries of the mean of a cluster are calculated using only vectors, where the entries are not missing in the respective dimensions. The distance measure is also adapted. The distance between two vectors is calculated only with the common dimensions, where ratings are not missing. The result is then divided by the number of common dimensions to get a distance measure that is independent of the number of common dimensions.

1.2 Factorization-Based Recommender

1.3 Hybrid Recommender

bruh

2 Results

2.1 Performance

2.2 Error Scores

2.3 Final Score

References