Empirical Analysis of Predictive Algorithms for Recommender System

Ruoheng Li
Rutgers University
Ruoheng@eden.rutgers.edu
Introduction

Incredibly large amount of information out there on the internet

Create the technologies helping us to find out which is most valuable to us

Recommender System

Apply knowledge discovery techniques to the problem of making personalized recommendations

Collaborative or Social Filtering

Based on previous examples of user’s likes and dislikes
Collaborative Filtering Algorithm

A rating matrix, including \( n \) users and \( m \) items

- Memory-based algorithm – operate over the entire user database
- Model-based algorithm – use the user database to learn a model
Memory-Based Algorithms

Utilize the entire user database to make predictions

\[ p_{a,j} = \bar{v}_a + \kappa \sum_{i}^{n} w(a,i)(v_{i,j} - \bar{v}_i) \]

Correlation method (Pearson correlation coefficients)

\[ w(u_{new}, u_i) = \frac{\sum_{j} (v_{new,j} - \bar{v}_{new})(v_{i,j} - \bar{v}_i)}{\sqrt{\sum_{j} (v_{new,j} - \bar{v}_{new})^2 (v_{i,j} - \bar{v}_i)^2}} \]

Prediction for new user is a weighted sum

\[ p_{new,j} = \bar{v}_{new} + \frac{\sum_{i} w(u_{new}, u_i)(v_{i,j} - \bar{v}_i)}{\sum_{i} w(u_{new}, u_i)} \]

Cold start problem !!!
**Model based Algorithm**

- Provide item recommendation by developing a model of user ratings
- A probabilistic approach – computing the expected value of a user prediction

\[ p_{a,j} = E(v_{a,j}) = \sum_{i=0}^{m} \Pr(v_{a,j} = i \mid v_{a,k} \in I_a) i \]

**Bayesian Model**

\[ \Pr(C = c, v_1, \ldots, v_n) = \Pr(C = c) \prod_{i=1}^{n} \Pr(v_i \mid C = c) \]

- Given the class, the ratings are independent
- Probabilities of class variable C are based on user covariates

**Provide solution to the cold start problem!!!**
Empirical Analysis

Dataset - beverage ratings
- 210 users rated how well they liked 16 different drinks

Evaluation Criteria
- Brier Score
  \[ B_{\text{new}} = \frac{1}{m} \sum_{j=1}^{m} \left( v_{\text{new},j} - p_{\text{new},j} \right)^2 \]
  - Accuracy percentage

Experiment Results
compare prediction based on
- Grand Mean
- Group Mean
- Correlation Method
- Bayesian model
Experiment Results

![Graph showing experiment results with different lines representing Grand Mean Score, Group Mean Score, Correlation, and Bayesian Model. The x-axis represents the number of rated beverages ranging from 0.5 to 15.5, and the y-axis shows the scores ranging from 0.60 to 0.70.]
## Experiment Results

<table>
<thead>
<tr>
<th>Method</th>
<th>0</th>
<th>2</th>
<th>5</th>
<th>10</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Past User’s Mean</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.37125</td>
</tr>
<tr>
<td>Past Users’ Group Mean</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.34635</td>
</tr>
<tr>
<td>Correlation</td>
<td>NA</td>
<td>0.34821</td>
<td>0.35714</td>
<td>0.32725</td>
<td>0.28217</td>
</tr>
<tr>
<td>Bayesian Model</td>
<td>0.36875</td>
<td>0.33376</td>
<td>0.31297</td>
<td>0.30181</td>
<td>0.29164</td>
</tr>
</tbody>
</table>

## Conclusion
- Group mean outperforms Grand mean
- Bayesian model performs better than correlation method when only partial information are observed
- Bayesian model solved the cold start problem
- Given large dataset, correlation method is likely to be competitive with the Bayesian model
The End
Thanks!