Project proposal

Team name: M2 Robo

Lee Chun Yin	Chiu Yu Ying	Chan Kwan Yin
3035469140	3035477630	3035466978
		Team leader

1. Data set to analyze

We will analyze the "Netflix Prize" dataset.

2. Proposed methodology

We are going to construct the following three models to analzye the data set and compare their performance.

2.1. KNN

We will include item-based k-nearest neighbours (KNN) as the first model. It is a kind of neighbourhood-based approaches, which are popular collaborative filtering methods in previous competitions (Progress Prize 2008's KNNMovie). The item-item correlation ρ_{ij} between item i, j will be calculated based on the ratings of users who rated both movies using different statistics such as Pearson and Spearman correlation. It is a robust classifier with a simple implementation and involves few parameters to tune (e.g. distance metric and k). However, it is sensitive to outliers during the choice of neighbour process based on distance criteria. Furthermore, KNN does not output rating values, reducing the interpretability of the recommendation results and hinders the potential to compare with other data directly.

2.2. Simple SVD

Our second model will use the "SVD" model, proposed by Yehuda Koren [2]. The model predicts the rating for movie i by user u in the form

$$\hat{r}_{ui} = \mu + b_u + b_i + \mathbf{q}_i \cdot \mathbf{p}_u$$

where the parameters to train are

- μ is the mean rating
- $b_u(t)$ is a time-sensitive user-specific bias predicted by parameter binning
- $b_i(t)$ is a time-sensitive item-specific bias predicted by parameter binning
- $\mathbf{q}_i$ and $\mathbf{p}_u$ are latent factors produced from matrix factorization with simple stochastic gradient descent after the ratings are offset by the bias.

Compared to KNN, SVD provides better serendipity since it computes all movies globally and provides a single objective metric, unlike KNN, which largely depends on the user's similarity to other users. However, this method requires more training time since it involves a collection of different parameters to tune.

2.3. SVD++

The third model is a modification to the SVD model by introducing time effects. We will bin the training data over time intervals and train the parameters $b_u, b_i, \mathbf{p}_u$ as functions in terms of rating date t, where each user-bin is considered as a separate u in the factorization. This idea is a simplification based on the "SVD++" model proposed by BellKor in Progress Prize 2008 [1].

Since this model additionally considers time effect, it is speculated that it performs no worse than the simple SVD model. The temporal effects of item bias, user bias and user-preference bias are thoroughly explained in BellKor's paper [1]. However, since recommendation systems are often only useful for predicting future events, the time-dependent functions may be somehow extrapolated to the future, while extrapolation is known to be unreliable. Nevertheless, this is helpful in offseting rating bias during training.

References

[1] Robert M Bell, Yehuda Koren, and Chris Volinsky. The bellkor 2008 solution to the netflix prize. *Statistics Research Department at AT&T Research*, 1(1), 2008. 2

[2] Yehuda Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 426–434, 2008. 2