Chp 2: Neighborhood Based Collaborative Filtering

§ 2.1 Introduction

- Types of collaborative filtering

  (a) User based: ratings provided by similar users to user A are used to recommend to the A

  (b) Item based: To make recommendation for target item B, you determine a set of similar items to B, rated by user A, and predict the rating of item B

- Rmk: user based – uses neighbors ratings, similarity among rows of R

  item based – uses your own ratings, similarity among columns of R.

- $R \in \mathbb{R}^{m \times n}$ m – users n – items

- Problems: (a) predicting missing values in R

  (b) Determing the top k-items/users of unknown ratings

§ 2.2 Properties of R

- Entries in R can be supported on really any subset of $\mathbb{R}$
- Typically the # ratings is a skewed dist. so if $M \in \mathbb{R}^{m \times n}$ s.t. $m_{ij} = \begin{cases} 1 & R_{ij} \neq \text{missing} \\ 0 & \text{o.w} \end{cases}$

  then $\text{colsums}(M) = I^T M$ is a skewed dist.

- Implications: over recommendation of popular items, few often rated products make neighborhood methods unstable.

§ 2.3 Predicting Ratings with Neighborhood – Based Methods

- Basic idea: similar users rate items similarly

  similar items are rated similarly by the same user.

- Think about this like a N.N. classification problem

§ 2.3.1 User-Based Neighborhood Models

- Create neighboords for each user

- Let $I_u = \{$ indicies from which ratings have been given by user $u\}$

- Measures for similarity $Sim(u,v)$

  - Pearson Corr. for ratings in $I_u \cap I_v$

$$Sim(u,v) = \frac{\sum_{k \in I_u \cap I_v} (R_{uk} - \mu_u)(R_{vk} - \mu_u)}{\sqrt{\sum_{k \in I_u \cap I_v} (R_{uk} - \mu_u)^2} \sqrt{\sum_{k \in I_u \cap I_v} (R_{vk} - \mu_v)^2}}$$

  where $\mu_u = (|I_u|)^{-1} \sum_{k \in I_u} R_{uk}$.

  Rmk: Not quite Pearson because μ_u, μ_v is computed over I_u, I_v, respectively, not $I_u \cap I_v$. But you get some computational speed up this way.

  - Define a users neighborhood on an item-by-item basis so that your peer group all have ratings for the item in question

  - The estimated rating is then provided by the weighted average

$$\hat{R}_{ui} = \frac{\sum_{v \in P_u(i)} sim(u,v) \cdot R_{vi}}{\sum_{v \in P_u(i)} sim(u,v)} \quad : \quad P_u(i) = \{\text{peers most similar to } u \text{ who rated } i\}$$

- Issues: Ratings on different scales. (general neg./pos. raters)

- Sol: Rowise centering - before defining estimate.

- Let $s_{uj} = R_{uj} - \mu_u$ be the user-centered rating

- So the estimated centered rating is then added to the user's rating to provide the estimate

$$\hat{R}_{uj} = \mu_u + \frac{\sum_{v \in P_u(j)} sim(u,v) \cdot s_{vj}}{\sum_{v \in P_u(j)} sim(u,v)}$$

- Extensions are abound with modifications to $Sim(u,v)$ and $P_u(j)$.

- Other similarity functions may include