CF : Assignment 1

Anshuman Suri
2014021

Note: Plots and values are attached in *results.pdf*

1. For user-user model, the Pearson correlation coefficient (with values substituted appropriately) :

$$\kappa_{x,y} = \mathrm{sim}\,(u_x, u_y) = \frac{\sum_{h=1}^{n'} \left(r_{u_x,i_h} - \overline{r_{u_x}}\right) - \left(r_{u_y,i_h} - \overline{r_{u_y}}\right)}{\sqrt{\sum_{h=1}^{n'} \left(r_{u_x,i_h} - \overline{r_{u_x}}\right)^2}\sqrt{\sum_{h=1}^{n'} \left(r_{u_y,i_h} - \overline{r_{u_y}}\right)^2}}.$$

, is used for explicit rating. Per user rating prediction is then taken as the rounded off value (clipped between minimum and highest possible values) as :

$$\mathrm{CF}_{\mathrm{UB-ER}} = p_{u_\mathrm{a},i_\mathrm{a}} = \bar{r}_{u_\mathrm{a}} + \frac{\sum_{h=1}^{m'} k_{\mathrm{a},h}(r_{u_h,i_\mathrm{a}} - \bar{r}_{u_h})}{\sum_{h=1}^{m'} \left|k_{\mathrm{a},h}\right|}.$$

2. For item-item model, the Pearson correlation coefficient (with values substituted appropriately) :

$$\lambda_{x,y} = \mathrm{sim}\,(u_x, u_y) = \frac{\sum_{h=1}^{p} \left(r'_{u_x,c_h} - \overline{r'_{u_x}}\right) - \left(r'_{u_y,c_h} - \overline{r'_{u_y}}\right)}{\sqrt{\sum_{h=1}^{p} \left(r'_{u_x,c_h} - \overline{r'_{u_x}}\right)^2}\sqrt{\sum_{h=1}^{p} \left(r'_{u_y,c_h} - \overline{r'_{u_y}}\right)^2}}$$

, is used for explicit rating. Per user rating prediction is then taken as the rounded off value (clipped between minimum and highest possible values) as :

$$\mathrm{CF}_{\mathrm{IB-ER}} = p_{u_a,i_a} = \bar{r}_{i_a} + \frac{\sum_{h=1}^{n'} \mu_{a,h}(r_{u_a,i_h} - \bar{r}_{u_a})}{\sum_{h=1}^{n'} |\mu_{a,h}|}.$$

3. Significance weighting masks the similarity weighting function :

$$w_{a,u} = \frac{\sum_{i=1}^{m} (rank_{a,i} - \overline{rank}_a) * (rank_{u,i} - \overline{rank}_u)}{\sigma_a * \sigma_u}$$

   with function *f* such that *f(x,n) = { x; n >= threshold, x * n/threshold, otherwise*
   The value of threshold here is varied and the most optimal one used to get lowest NMAE. Using these weights, predictions from users are considered (weighted mean) to perform a prediction. The rank here us calculated using Spearman coefficient.

4. Variance weighting uses the same underlying algorithm, but uses a different weighting function:

$$w_{a,u} = \frac{\sum_{i=1}^{m} v_i * z_{a,i} * z_{u,i}}{\sum_{i=1}^{m} v_i}$$

5. Similar to the third and fourth parts, except that instead of considering all ratings, we only pick the top *K* (*K* empirically decided). This, as claimed in the paper given for reference, results in better predictions, as it filters out ratings which are not so significant, thus reducing noise. For this assignment, the weighting function used in the third question is used for this part.