Project 4

Collaborative Filtering

Group 8: Chen, Mengqi Huang, Yuexuan
Li, Xueyao Zheng, Jia

Datasets

Index	DataSet	Total Users	Total Titles	Vote Range	Vote Type
1	MSWEB	4151	269	0 1	implicit
2	Eachmovie	5055	1619	1 2 3 4 5 6	explicit

We sampled 5,000 users from the full data to alleviate computation burden (download link)

Memory-based Algorithm

- **Similarity Weight**
  - Pearson Correlation 1,2
  - Spearman Correlation 1,2
  - Mean-squared-difference 1,2
  - SimRank 1
- **Variance Weighting**
  - No
  - Yes
- **Selecting Neighbor**
  - Best-n estimator
    - n = 20
    - n = 50

Prediction

$$p_{a,i} = \bar{r}_a + \frac{\sum_{u=1}^{n}\left(r_{u,i} - \bar{r}_u\right) * \omega_{a,u}}{\sum_{u=1}^{n} \omega_{a,u}}$$

Model-based Algorithm

Cluster Models

- **EM algorithm**
  - Maximum Iteration $\mathrm{T} = 30$
  - Convergence Condition

    $\mathrm{SSE}(\mu_t - \mu_{t-1}) < 10^{-8}$ & $\mathrm{SSE}(\gamma_t - \gamma_{t-1}) < 10^{-2}$
- **5-Fold Cross Validation**
  - Best cluster size $\mathrm{C} = 6$
- **Evaluation on Test Set**
  - $\mathrm{MAE} = 0.994$

C	MAE
3	1.0225
4	1.0173
5	1.0121
6	1.0044
7	1.0153
8	1.0154

Ranked Scoring for MSWEB dataset

Variance Weighting	Best-n Neighbor	Pearson	Spearman	Mean-Squared-Difference	SimRank
No	n=20	32.49216	32.50296	32.54723	31.36264
	n=50	35.2103	35.20447	35.39599	33.7571
Yes	n=20	34.01533	34.01533	34.23291	33.35221
	n=50	36.67433	36.67433	37.44365	36.35416

MAE for Eachmovie Dataset

Memory-based					Model-based
Variance Weighting	Best-n Neighbor	Pearson	Spearman	Mean-Squared-Difference	Cluster Models
No	n=20	1.13742	1.13742	1.084893	0.9940024
	n=50	1.13742	1.13742	1.074721	
Yes	n=20	1.13742	1.13742	1.077093	
	n=50	1.13742	1.13742	1.058211	