Project4
Collaborative Filtering

PRESENTED BY Group 7

Ding, Xueying,
Fan, Xiaochen,
Guo, Tao,
Jiang, Chenfei,
Yu, Linna

Contents

DATA

MS

V1	V2	V3
C	10010	10010
V	1010	1
V	1000	1
V	1011	1
V	1012	1
V	1013	1
V	1014	1

Movie

	Movie	User	Score
1	1	1	4
2	2	1	4
3	17	1	5
4	18	1	2
5	25	1	5
6	31	1	2
7	32	1	4

Evaluation

- For MS:

Rank Score

$$R_a = \sum_j \frac{\max(v_{a,j} - d, 0)}{2^{(j-1)/(\alpha-1)}}$$

$$R = 100 \frac{\sum_a R_a}{\sum_a R_a^{max}}$$

- For Movie:

MAE: mean absolute error

ROC: true positive rate ~ false positive rate

Memory-based Algorithm

Similarity Weight

Pearson Correlation ——

$$w(a,i) = \frac{\sum_j (v_{a,j} - \overline{v}_a)(v_{i,j} - \overline{v}_i)}{\sqrt{\sum_j (v_{a,j} - \overline{v}_a)^2 \sum_j (v_{i,j} - \overline{v}_i)^2}}$$

Vector Similarity ——

$$w(a,i) = \sum_j \frac{v_{a,j}}{\sqrt{\sum_{k \in I_a} v_{a,k}^2}} \frac{v_{i,j}}{\sqrt{\sum_{k \in I_i} v_{i,k}^2}}$$

SimRank (only for Movie data set) ——

$$s(A,B) = \frac{C_1}{|O(A)||O(B)|} \sum_{i=1}^{|O(A)|} \sum_{j=1}^{|O(B)|} s(O_i(A), O_j(B))$$

$$s(c,d) = \frac{C_2}{|I(c)||I(d)|} \sum_{i=1}^{|I(c)|} \sum_{j=1}^{|I(d)|} s(I_i(c), I_j(d))$$

Selecting Neighbors & Prediction

- Selecting Neighbors
- Weight Threshold:-- parameter: p
- Best-n-estimator-- parameter: n
- Combined-- parameter: n,p

Prediction :

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

Model-based Algorithm

- Cluster Model (only for MS data set)

$$\Pr(C = c, v_1, \ldots, v_n) = \Pr(C = c) \prod_{i=1}^{n} \Pr(v_i | C = c)$$

- EM algorithm

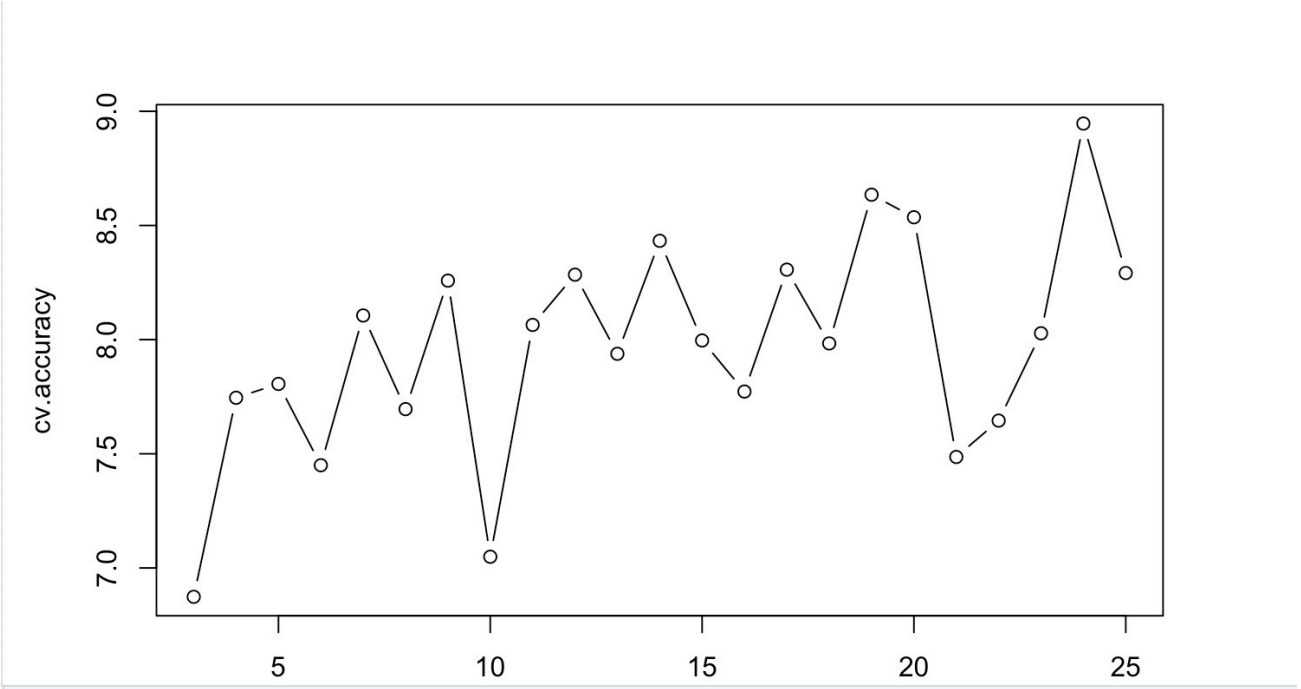
E-step: guess labels

M-step: estimate parameters

2 steps are iterated repeatedly

Cross-Validation

Choosing C



- All but 1: In validation data, we withhold a single randomly selected vote for each user and try to predict its value given all other votes the user has voted on.
- So, based on our output , C=24 has the highest ranking score.

MS data results

Ranking Score

Variance Weighting = F		Selecting Neighbors		
		Weight Threshold	Best-n-estimator	Combined
Similarity Weight	Pearson Correlation	7.57351	7.94754	7.96865
	Vector Similarity	7.55442	7.95246	7.98182
Cluster (C = 24)		**39.33172**		
Cluster (C = 19)		39.11221		
Cluster (C = 14)		38.61513		

Based on Rank Score: (We choose d = 0)
Cluster performs the best

Movie data results

Based on MAE: Pearson + Best-n-estimator

Variance Weighting = F		Selecting Neighbors		
		Weight Threshold	Best-n-estimator	Combined
Similarity Weight	Pearson Correlation	1.181029	0.9605008	1.443416
	Vector Similarity	1.154001	1.13852	1.156417
	SimRank	**1.097066**	**1.090524**	**1.097066**

Based on ROC: Vector Similarity + Best-n-estimator

Variance Weighting = F		Selecting Neighbors		
		Weight Threshold	Best-n-estimator	Combined
Similarity Weight	**Pearson Correlation**	**0.6708**	**0.6711**	**0.6683**
	Vector Similarity	0.6639	0.6699	0.6619
	SimRank	0.6461	0.6453	0.6461