CF Assignment-2

Rhea Goel (2010068), Sahil Jain (2010071)

Note- The zip file in the email contains the .py files.

(A) Variance Weighting

(using top 5 neighbours i.e. K=5)

Fold	MAE
Fold 1	1.04217351458
Fold 2	1.08117135241
Fold 3	1.06534096967
Fold 4	1.0600830137
Fold 5	1.03951836835
Average	**1.057657443742**

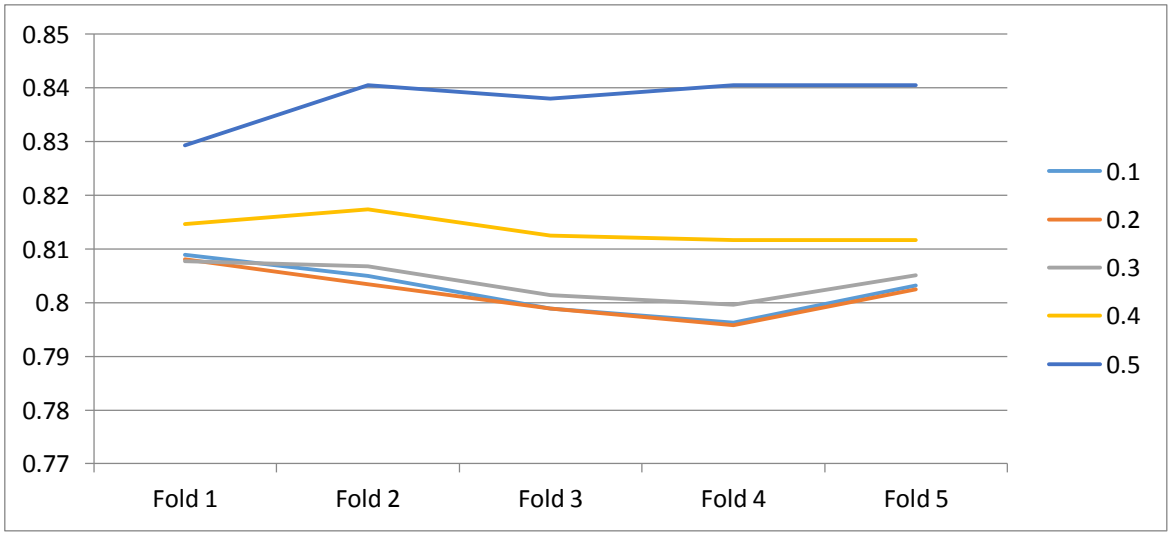
(B) Correlation Thresholding

Threshold	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
0.1	0.808881838611	0.80493561131	0.798927956935	0.796306748837	0.803251054582
0.2	0.808132760663	0.80342170248	0.798918622972	0.795789049969	0.802418362689
0.3	0.807686204947	0.806716566205	0.801353600736	0.799630466344	0.805108230909
0.4	0.814602548363	0.817409040759	0.812463472532	0.811704703282	0.811704703282
0.5	0.829301960775	0.840454183284	0.837965325091	0.840454183284	0.84045782967

Average MAE for each threshold

Threshold	0.1	0.2	0.3	0.4	0.5
Average	**0.80246**	**0.80174**	**0.8041**	**0.81358**	**0.83773**

Comparison of MAE obtained for different thresholds in Correlation Thresholding



We can see, that threshold of 0.2 gives the least Mean Absolute Error. also, note that threshold of 0.5 gives maximum error. This could be attributed to the fact that Pearson coefficient >0.5 means high similarity between the users, and there were not many so similar neighbours for any user.

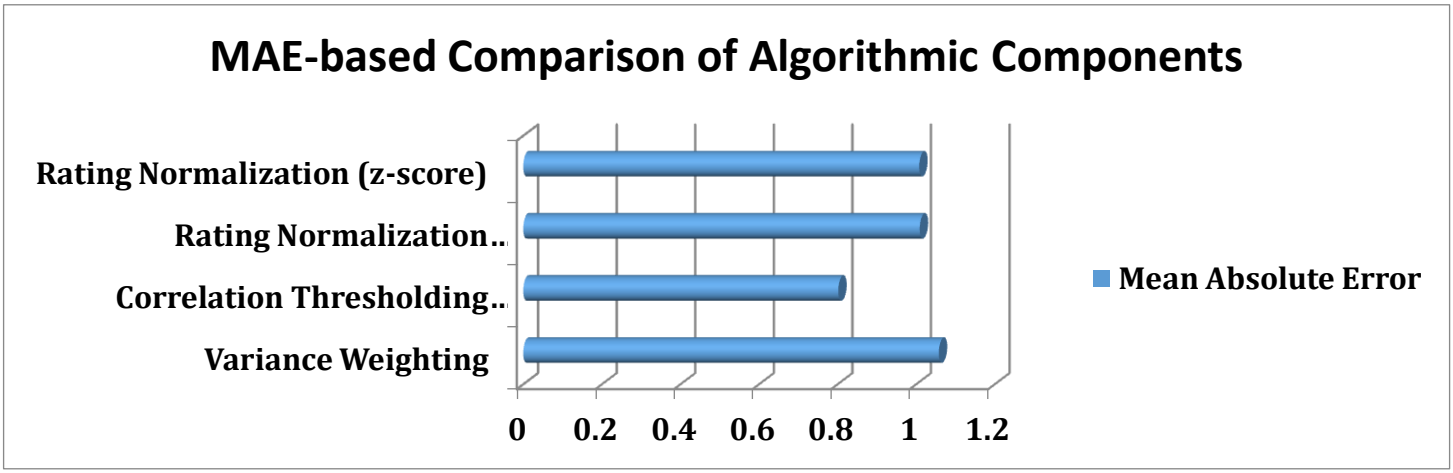
(C) Rating Normalization

(assuming 5 neighbours i.e. K=5)

Fold	Deviation-from-Mean	Z-score
Fold 1	1.01755444475	1.01578526577
Fold 2	1.00548269658	1.00151622475
Fold 3	1.01888747113	1.01770457113
Fold 4	1.00891115279	1.00961502454
Fold 5	0.995027146012	0.994189591687
Average	**1.009172582252**	**1.007762135575**

It is clear that both average-deviation-from-mean, and z-scores give more or less the same Mean Absolute Error.

- **Comparison of Variance Weighting, Correlation Thresholding and Rating normalization (both, average-deviation-from-mean and z-scores)**



(D) (BONUS) Variance Weighting & Correlation Thresholding

(assuming 5 neighbours i.e. K=5, and Correlation Threshold = 0.2)

From Correlation Thresholding of Part (B) we find that the **best threshold is 0.2**. Using this threshold on all 5 files, we get the following output:

Fold	MAE
Fold 1	0.946245403236
Fold 2	0.965797777134
Fold 3	0.959054860915
Fold 4	0.956746655691
Fold 5	0.948095208701
Average	**0.9551879811354**

(E) (BONUS) Correlation Thresholding & Rating Normalization (deviation-from-mean)

(assuming 5 neighbours i.e. K=5, and Correlation Threshold = 0.2)

From Correlation Thresholding of Part (B) we find that the **best threshold is 0.2**. Using this threshold on all 5 files, we get the following output:

Fold	MAE
Fold 1	0.757191589578
Fold 2	0.746133190497
Fold 3	0.743338084769
Fold 4	0.741210353655
Fold 5	0.740531689421
Average	**0.745680981584**

We can see that the MAE reduces significantly as we combine any two of these algorithmic concepts, and our predictions become more robust.