Project Report 3

Recommendation system using Collaborative Filtering

COMP 135 Intro to Machine Learning

Jiacheng Qu

1234031

Friday, April 20, 2018

1 State the learning algorithms you have used and the parameter setting of your algorithm.

The learning algorithm I have chosen is SVD from scikit-surprise.

$$\sum_{r_{ui} \in R_{train}} \left(r_{ui} - \hat{r_{ui}}\right)^2 + \lambda \left(b_i^2 + b_u^2 + ||q_i||^2 + ||p_u||^2\right)$$

prediction: $\widehat{r_{ui}}$ is calculated by the following equation as an unbiased (biased= False) version of SVD algorithm

$$\hat{r_{ui}} = q_i^T p_u$$

where user vectors U is p_u and item vectors V is q_i

```
U_vectors, V_vectors = svd_algo.pu, svd_algo.qi
```

The parameter setting that I modified in this project are:

```
biased= False, n_epochs=500, lr_all=0.01,
n_factors=[?], reg_all =[?]
```

As stated above, biased, number of epochs and learning rate have been kept unchanged. The number of factors k and regulation term λ are the main focus here.

1.1 State the model selection procedure you used to decide model hyper-parameters.

To figure out a optimal setting of hyper-parameters for the SVD algorithm, I used GridSearchCV with the estimation of root-mean-square error (RMSE)

```
from surprise.model_selection import GridSearchCV
```

I selected and stored the setting with the best rmse performance into a python dictionary for further use.

The one I saved for the exported predictions is stated as the following:
>>> best rmse is: 0.9292760536313887
>>> best setting is: {'n_epochs': 100, 'biased': False, 'reg_all': 0.11, 'lr_all': 0.01, 'n_factors': 40}

2 Report your analysis of learned vectors. Include the plots required in Section 3 in your report.

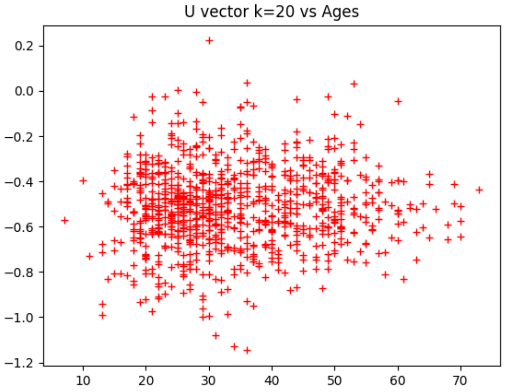
RMSE	'n_epochs'	'lr_all	'reg_all'	'n_factors'		Age vs U vector	Gender vs U vector	Release Year vs V vector
0.925	50	0.01	0.11	40		0.0722	0.0405	0.0858
0.931	100	0.01	0.1	30		0.0876	0.0500	0.0817
0.948	20	0.005	0.02	15		0.0548	-0.0588	0.0345
0.952	20	0.005	0.03	20		0.0468	0.0548	-0.0626
0.950	20	0.005	0.03	15		0.0325	0.0310	0.0625
0.9506	20	0.005	0.01	10		0.0318	-0.0197	0.0409
0.9457	20	0.012	0.09	30		0.0524	0.0587	0.0552

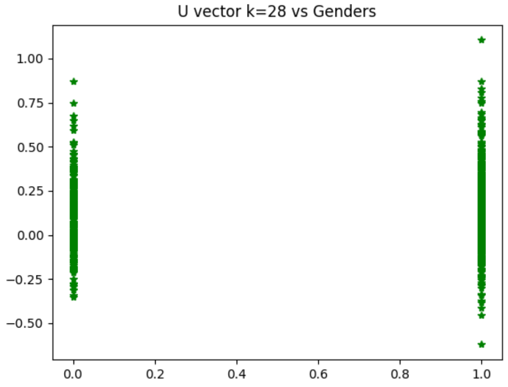
As the table shown above, SVD provides the best performance when the setting is as the following:

'n_epochs'	'lr_all	'reg_all'	'n_factors'
50	0.01	0.11	40

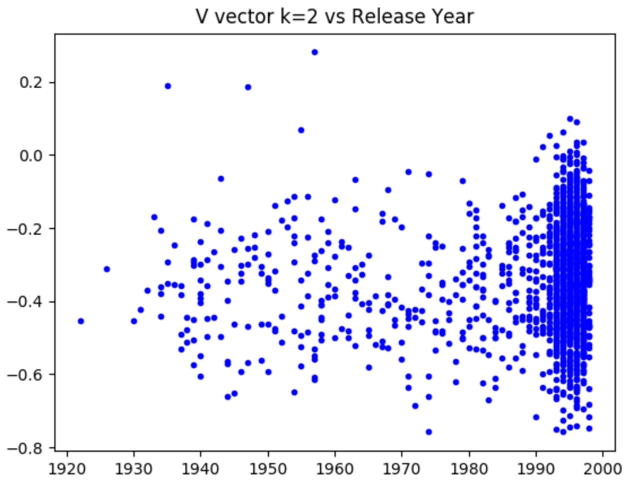
However, none of the three correlation coefficients is significate enough to provide a relationship between the user/item information with user/item vectors.

There could be inferred from the following plots as well:





The only significate info that could be concluded by plotting the correlation coefficient is that release year are condensed at around the 90s.



3 Write a short paragraph to discuss how to incorporate user information and movie information into the recommendation model.

Since the correlation coefficients are always below 0.1, the data hardly implies any relation between user vectors and user information, neither does user vectors vs user information either. To improve the recommendation model, there is almost nothing could be imported from the user/item information sets.