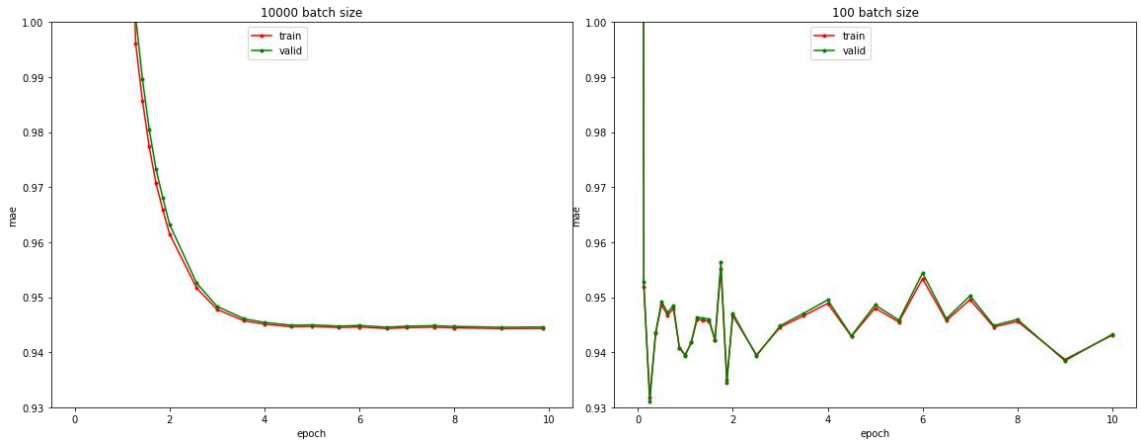
Project C Report

Problem 1

1a:

Figure 1a: Model1 mean absolute error vs. epoch



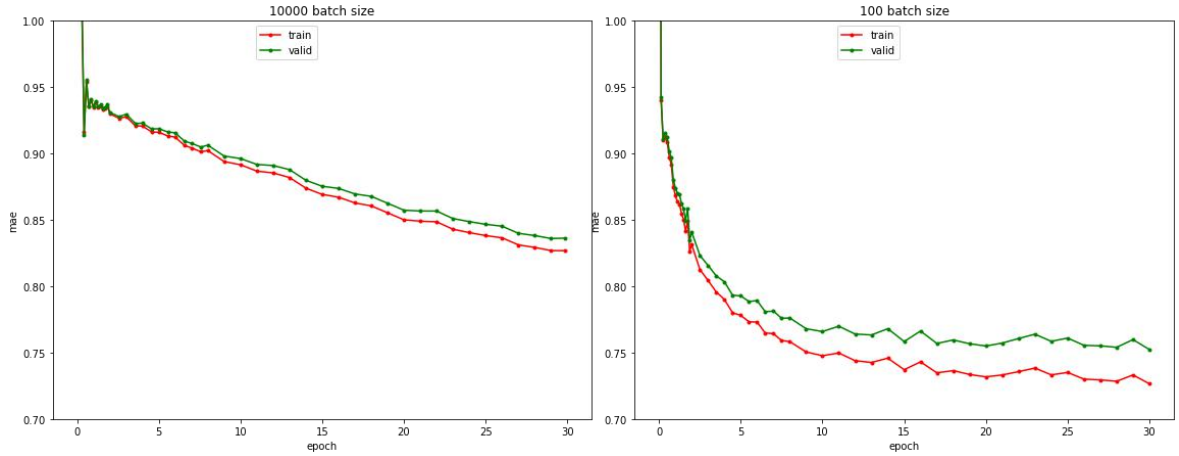
1b:

I would chose to compute the mean rating of training data set. The computed mu value is 3.530, and compare to the model's mu value 3.58, they are very close, so the result agrees with the SGD solution.

Problem 2

2a:

Figure 2a: Model2 mean absolute error vs. epoch



2b:

	title	adjustment
0	Toy Story (1995)	0.509517
1	Lion King, The (1994)	0.384573
2	Snow White and the Seven Dwarfs (1937)	0.351903
3	Wizard of Oz, The (1939)	0.694477
4	Sound of Music, The (1965)	0.358339
5	Star Wars (1977)	1.002044
6	Empire Strikes Back, The (1980)	0.823100
7	Return of the Jedi (1983)	0.679109
8	Jurassic Park (1993)	0.309447
9	Lost World: Jurassic Park, The (1997)	-0.447043
10	Raiders of the Lost Ark (1981)	0.960315
11	Indiana Jones and the Last Crusade (1989)	0.586573
12	While You Were Sleeping (1995)	0.169043
13	Sleepless in Seattle (1993)	0.144257
14	My Best Friend's Wedding (1997)	-0.033909
15	Nightmare Before Christmas, The (1993)	0.243246
16	Shining, The (1980)	0.394687
17	Nightmare on Elm Street, A (1984)	-0.090654
18	Scream (1996)	0.048227
19	Scream 2 (1997)	-0.280708

Figure 2b: Movie title and adjustment, we can see Star Wars Series movies have very large positive bias, it seems sci-fi and adventure movie have large positive bias. And it is obvious that subsequent movie have large negative bias like Lost World: Jurassic Park and Scream 2. large c it means movie has higher rating overall and is very popular, large negative means movie has lower rating overall so it may not be a very good movie.

Problem 3

3a:

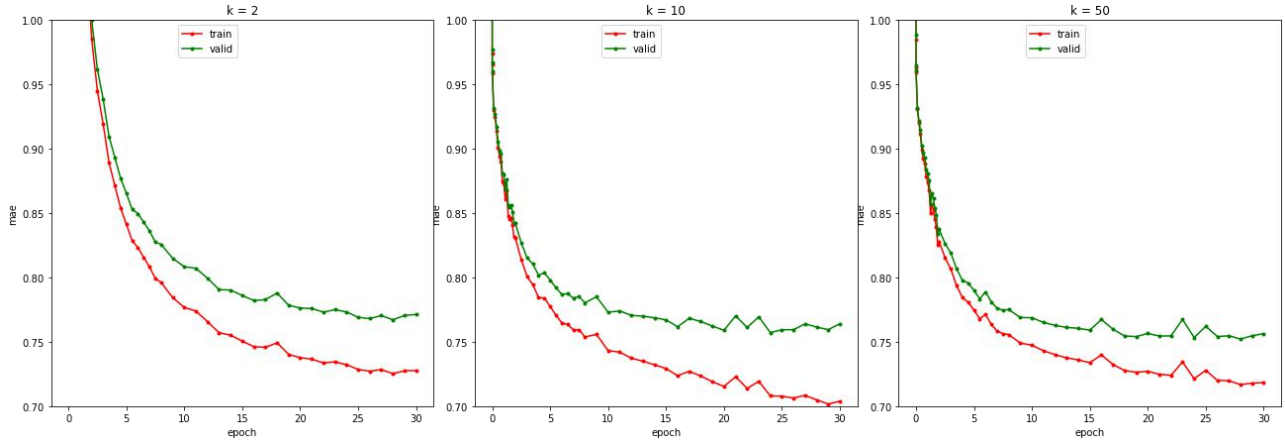


Figure 3a: There is sign of overfitting as we can see the training and validation MAE curves are clearly different and the training MAE gives much better performance. When k increase, the performance of the model increase

3b:

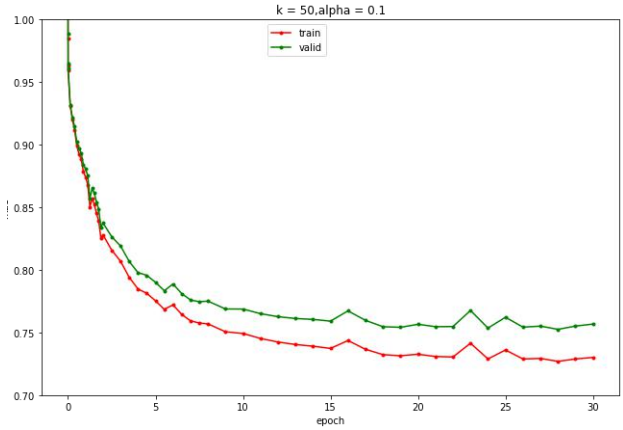


Figure 3b: I will select alpha = 0.1, batch size is 1000 and step size is 0.3, the heldout error is very close to 3a with alpha>0, it is just slightly better, 0.756 compared to 0.752.

3c:

Model	valid MAE set	test set MAE
M1	0.93844	0.94325
M2	0.75234	0.75399
M3(K=2)	0.76713	0.76803
M3(K=10)	0.75945	0.76144
M3(K=50)	0.75283	0.75479

Table 3c: MAE on validation set and test set for the "best version" of each model

I decide the best version of each model by choosing the model with best validation MAE, and see the hyperparameter of that model.

I recommend K = 50 for M3, I think we should not try more than 50 factor, because it does not increase the performance significantly but the training time become much longer, it is not a very good trade-off.

M3 (K=50) is the best overall because the complexity of the model is the highest among the 5 models, so it can better fit the features and give more precise prediction that other models.

3d:

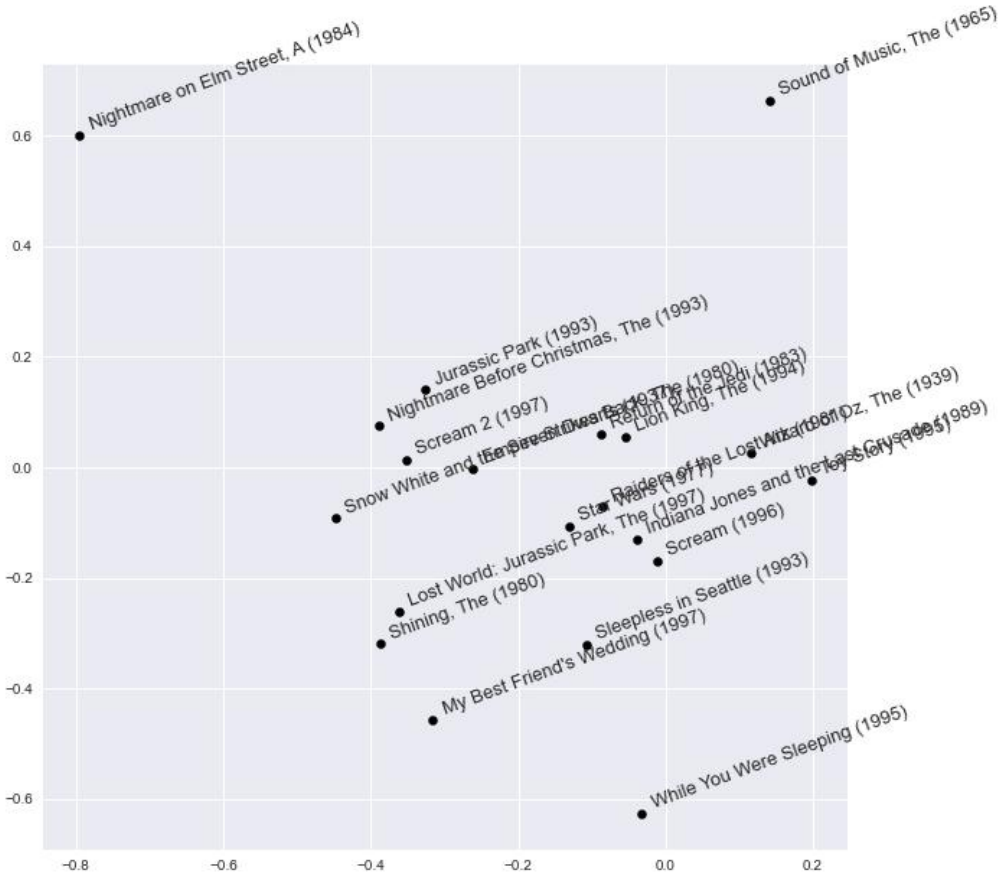


Figure 3d: Scatter plot for embedding vector V

It seemed horror movie are on the left part where V[0] <0, and non-horror movie are on the right part where V[0] >0. And movie on the top has early filmed year.

Problem 4

4a:

First, I remove mu from the parameter set, because I think it is unreliable to have

an exactly same baseline for all movie, and this could limit the performance of the model.

Then I add 2 other parameter: user age and user gender, I use the information from user_info.csv and use them as additional information in training to find out the potential correlation of rating and age or gender. Because we know different age group and gender group have different taste of movie so based on that we can use these information to improve the model. Hyper parameter selection is done by grid search. And I use early stop to get a better model since overfitting is observed after too many epochs.

4b:

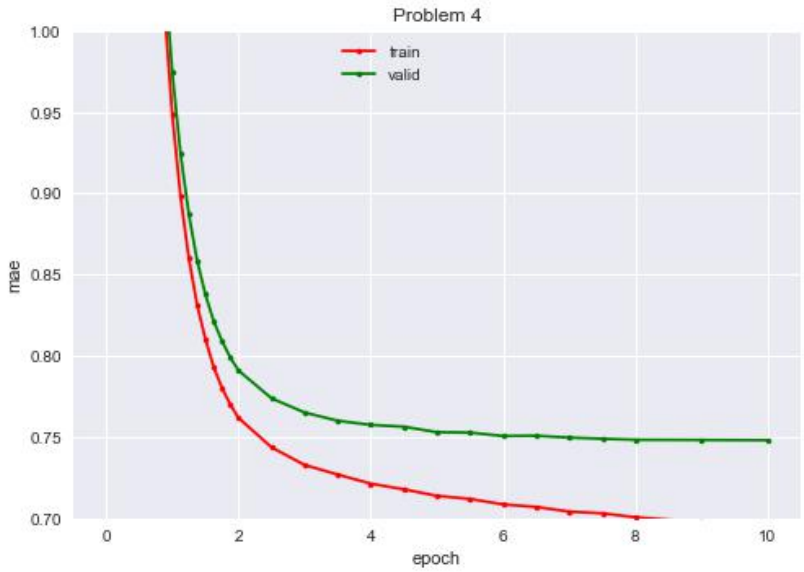


Figure 4b: Model with age and gender information added to training and mu removed from training, hyperparameter: n_epochs=10, batch_size=100, step_size=0.5, n_factors=10, alpha=0.1

4c:

Model	valid MAE set	test set MAE
M4	0.746	0.742

Figure 4c: ultimate(test) performance MAE and validation MAE

4d:

The leader board MAE is 0.742, compared to the validation MAE 0.746, they are consistent in general.

Compared to the best M3, this one perform better than M3, it is 0.1 lower in MAE.

4e:

The limitation is that the model can not find deeper relationship between user

rating and movie, all the learned parameter can only explain part of the correlation. So for future work I consider using deep neural network for the model and find some high level features that we may not observe or understand.

Problem 5

5a:

I split the data using k-fold split and use folds to do cross validation. I use MLP classifier because MLP can establish relation between each feature of the learned U vector, and can effectively find out which features have strong connection with gender. I tune hyperparameter with GridSearchCV.

5b:

```
[ 34  53]
[ 63 164]
```

Figure 5b: confusion matrix of gender

I get an 0.556 balanced accuracy for gender prediction, so it is not as accurate as the chance of the data set.

5c:

Predicting user's gender may not always be harmless. This is about user privacy and we should not offend user by prediction their gender, even if we can make very accurate prediction. And beside there are more than 2 kinds of gender so the ethic issue will be very troublesome.

A responsible AI should make all data anonymous and just try to learn some pattern from the data, and keep the privacy of customer.

To decide if the tool should be released, we should ask: will this offend people's privacy, will this cause trouble or harm for people, is this tool against social ethics, will this tool be used in a bad way.