Phan Anh Nguyen
Comp135

Project 3 Report

Problem 1a:

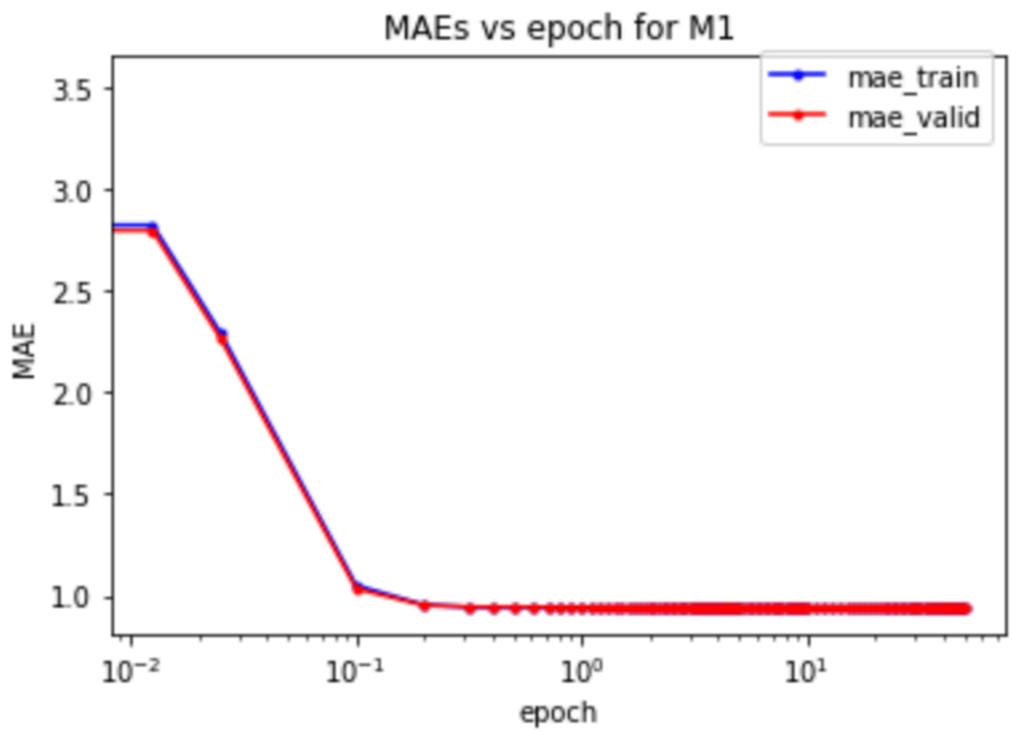

Fig 1. Plot of MAE over epoch for training and validation data using movie dataset of size 90000 (80000 training samples and 10000 validation samples) for Baseline Matrix Factorization model that exhibits performs mean-guessing behavior using log-scale for epochs.

Problem 1b:

Addition of regularization **will not** noticeably improve the performance with this model. This is because although SGD eventually finds the mean, there is no per-user or per-item vectors to be learned and therefore no weights to be assigned. Thus, adding regularization will not have any effect on the performance.

Problem 1c:

The closed-form operation we can apply is to simply compute the average ratings across all training data. Using this approach, we can obtain μ = 3.5367. Computed optimal μ is also 3.5367 which means that the two values are similar.

Phan Anh Nguyen
Comp135

Problem 2a:

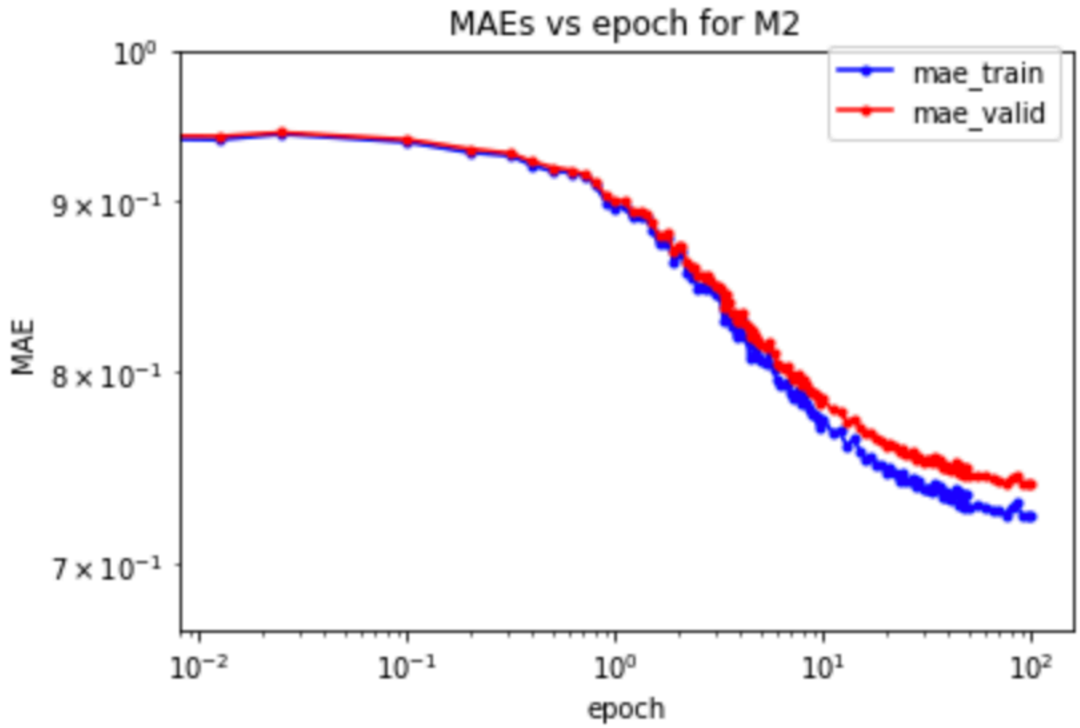

Fig 2a. Plot of MAE over epoch for training and validation data using movie dataset of size 90000 (80000 training samples and 10000 validation samples) for One-Scalar-Per-Item Matrix Factorization model that takes into consideration scalar biases for user and movie terms alongside mean. Figure uses log-scale for MAE and epochs.

Problem 2b:

Based on the results obtained, M2 has better predictive performance than M1 based on validation set MAE. M1 has final **MAE = 0.94551** at epoch 50 while M2 has **MAE = 0.73980** at epoch 50.

Phan Anh Nguyen
Comp135

Problem 2c:

Movie Name	Per-item Bias
Toy Story (1995)	0.5696288563914853
Lion King, The (1994)	0.43008068317211057
Snow White and the Seven Dwarfs (1937)	0.39482389631418874
Wizard of Oz, The (1939)	0.7479098158729092
Sound of Music, The (1965)	0.3721750666204736
Star Wars (1977)	1.0432849567315792
Empire Strikes Back, The (1980)	0.8962017259244739
Return of the Jedi (1983)	0.7005972130496689
Jurassic Park (1993)	0.3650406152514954
Lost World: Jurassic Park, The (1997)	-0.4614203735750662
Raiders of the Lost Ark (1981)	1.026563709855685
Indiana Jones and the Last Crusade (1989)	0.6269169756655962
While You Were Sleeping (1995)	0.20928101743795524
Sleepless in Seattle (1993)	0.18277341180020273
My Best Friend's Wedding (1997)	0.013032219029298355
Nightmare Before Christmas, The (1993)	0.2541997341728302
Shining, The (1980)	0.44802717139104475
Nightmare on Elm Street, A (1984)	-0.05150762575229475
Scream (1996)	0.09690445333126127
Scream 2 (1997)	-0.25499149835075524

Figure 2b: List of movie name with respective per-item biases. Examples were obtained from a list of selected movies and per-item biases were obtained via training of the One-Scalar-Per-Item model.

As there are no item and user feature vectors, the optimal biases for each item/user directly reflect how 'good' the movie as compared to the mean. A high positive bias means that the movie is rated higher on average, while a lower negative bias means that the movie is rated lower on average.

The first noticeable trend is that more popular franchises such as Star Wars, or Indiana Jones tend to be better liked (have higher bias terms). It is possible that those who see these movies are already fans and would tend to rate the movies higher. Within the franchise, it seems that sequels tend to be rated lower than first movies (The original Star Wars and Indiana Jones were all rated higher than sequels).

The horror movie genre seems to be disliked as movies in the genre (Scream movies, Nightmare on Elm Street), tend to have close to 0 or negative biases.

Phan Anh Nguyen
Comp135

Problem 3a:

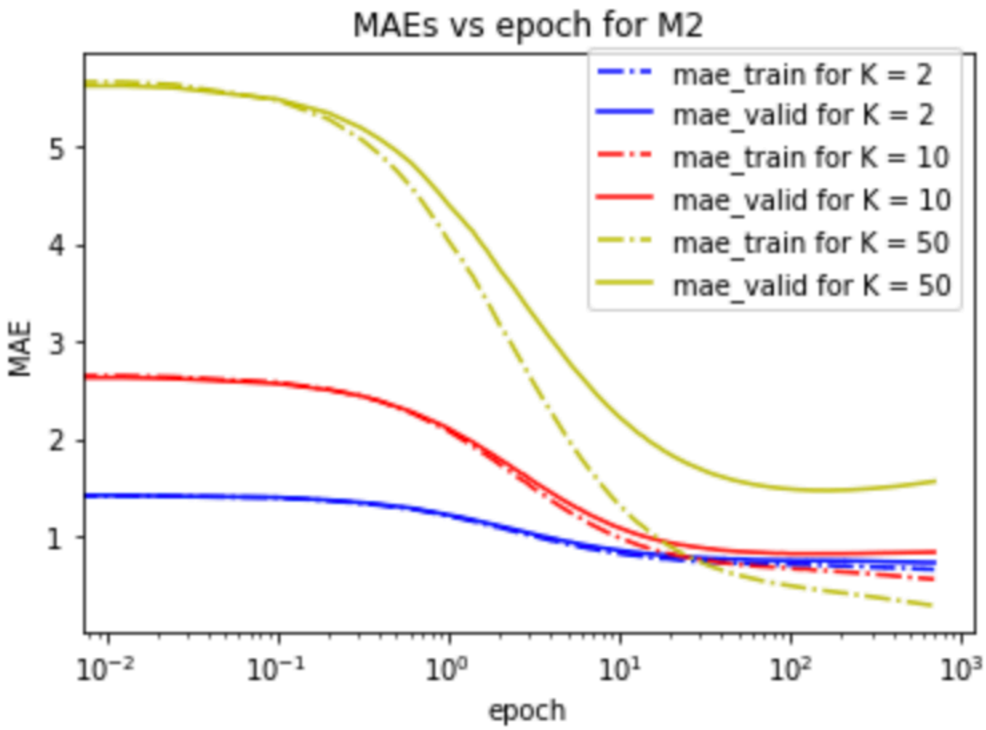

Fig 3a. Plot of MAE over epoch for training and validation data using movie dataset of size 90000 (80000 training samples and 10000 validation samples) Full Matrix Factorization model that takes into consideration scalar biases for user and movie terms alongside mean and user, item feature vectors. Regularization was set at $\alpha = 0$. Figure uses log-scale for epochs. ***Notice that overfitting occurs on validation data for K = 50.***

Phan Anh Nguyen
Comp135

Problem 3b:

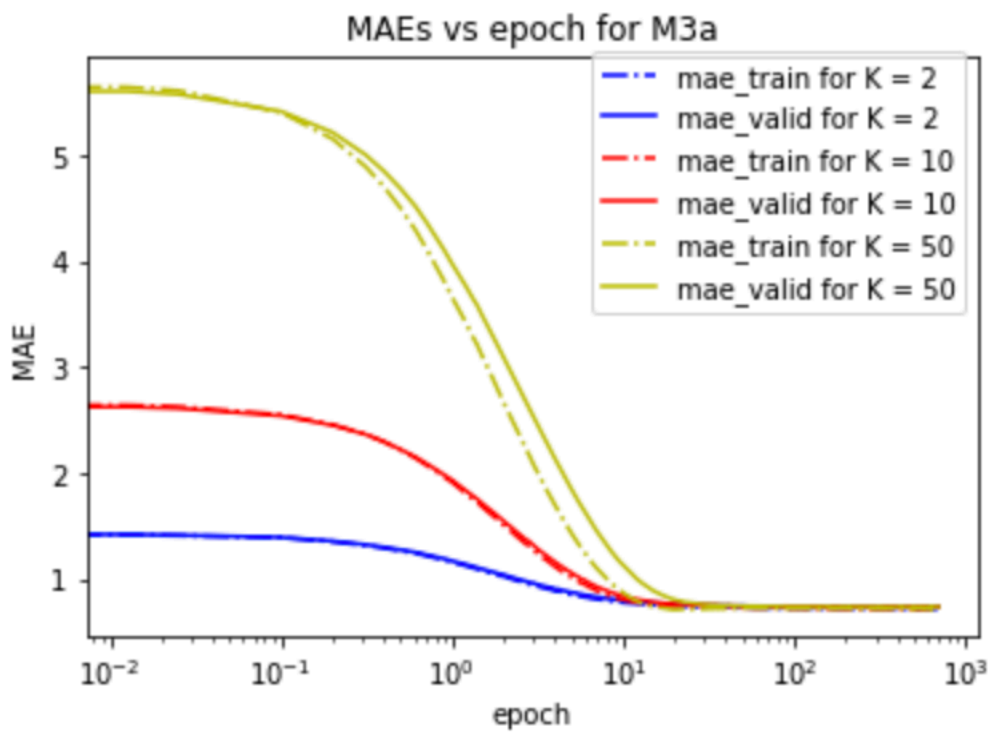

Fig 3b. Plot of MAE over epoch for training and validation data using movie dataset of size 90000 (80000 training samples and 10000 validation samples) Full Matrix Factorization model that takes into consideration scalar biases for user and movie terms alongside mean and user, item feature vectors. Regularization was set at $\alpha = 0.75$. Figure uses log-scale for epochs. ***Notice that addition of regularization = 0.75 has prevented the overfitting that previously occurs on validation data for K = 50.***

Problem 3c:

Other than L2/L1 regularization, we could ideally increase the number of data samples via Bootstrap Sampling. We could also decrease the learning rate to prevent the model from overfitting early on.

Problem 3d:

Based on the results obtained, M3 with $\alpha = 0$ and K = 2 has better predictive performance than M3 based on validation set MAE. M2 has final **MAE = 0.73980** while M2 has final **MAE = 0.73537**.

Phan Anh Nguyen
Comp135

Based on the results obtained, the K = 2 model yields the best results. Performance is observed to decrease as the number of K factors increases. Hence, we would not even need to test for K values larger than 50.

Problem 3e:

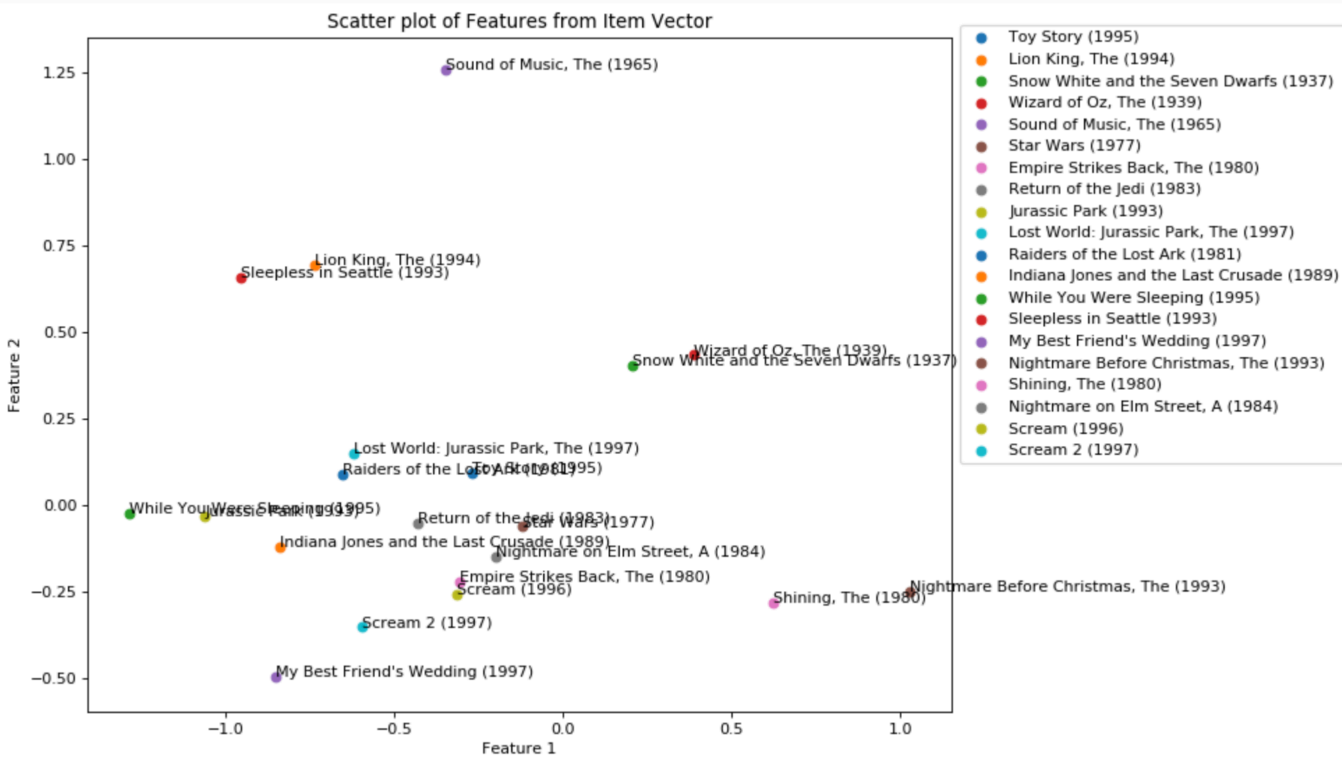

Fig 3c. Annotated scatter plot of feature 1 vs feature 2 from Item feature vector for sample movies obtained viz Matrix Factorization using Autograd with Regularization set at $\alpha = 0$.

Based on the annotated scatterplot, it is observed that general ‘blockbuster’ movies such as the Jurassic Park series, Star Wars series, and Indiana Jones series tend to be grouped together closer to the middle.

Mythical/Fantasy movies such as The Wizard of Oz and Snow White are grouped together.

Lastly, it is observed that movies close to each other tend to be released in the same time period (within 5 years of each other). Such examples are: Lion King – Sleepless in Seattle, Wizard of oz – Snow White, Lost World, Jurassic Park – Raiders of the lost Arc.

Phan Anh Nguyen
Comp135

Problem 4a:

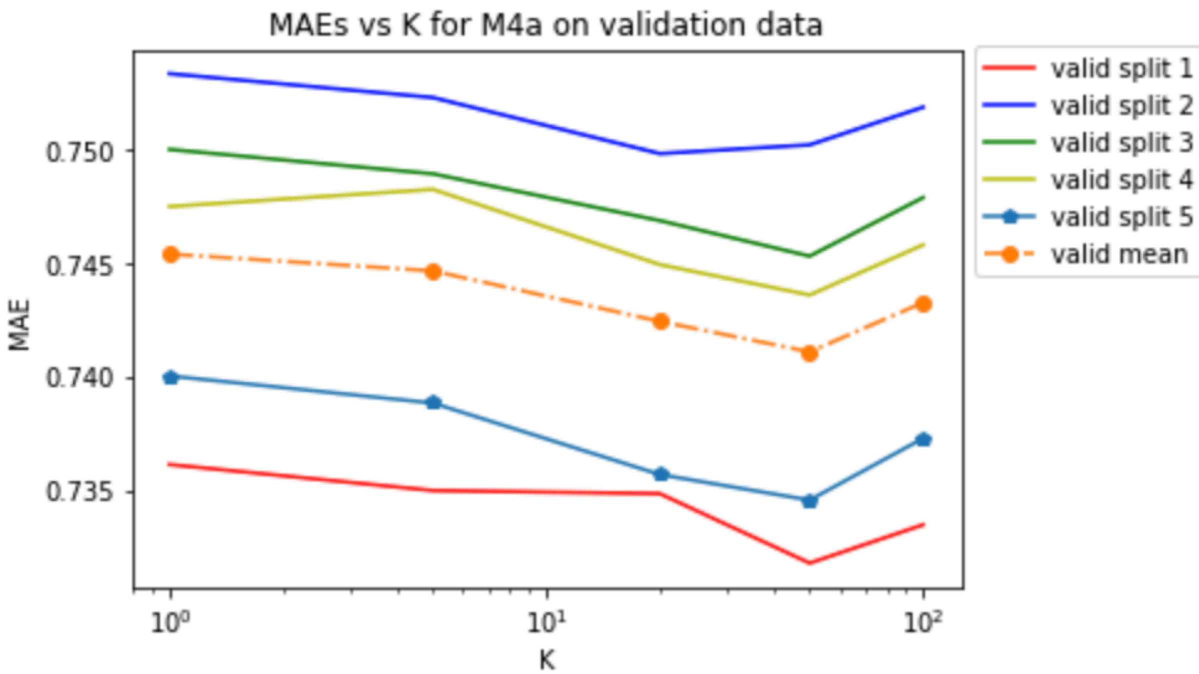

Fig 4a. Plot of Validation MAE over K for 5-fold cross validation using movie dataset of size 90000 with Surprise SVD. ***K = [1,5,20,50,100].***

Based on the figure obtained and mean MAE across 5 folds, it seemed that K value of 50 resulted in the best predictive performance. At K values lower than 50, underfitting occurs while at K values higher than 50, we see that overfitting occurs as MAE increases again.

Phan Anh Nguyen
Comp135

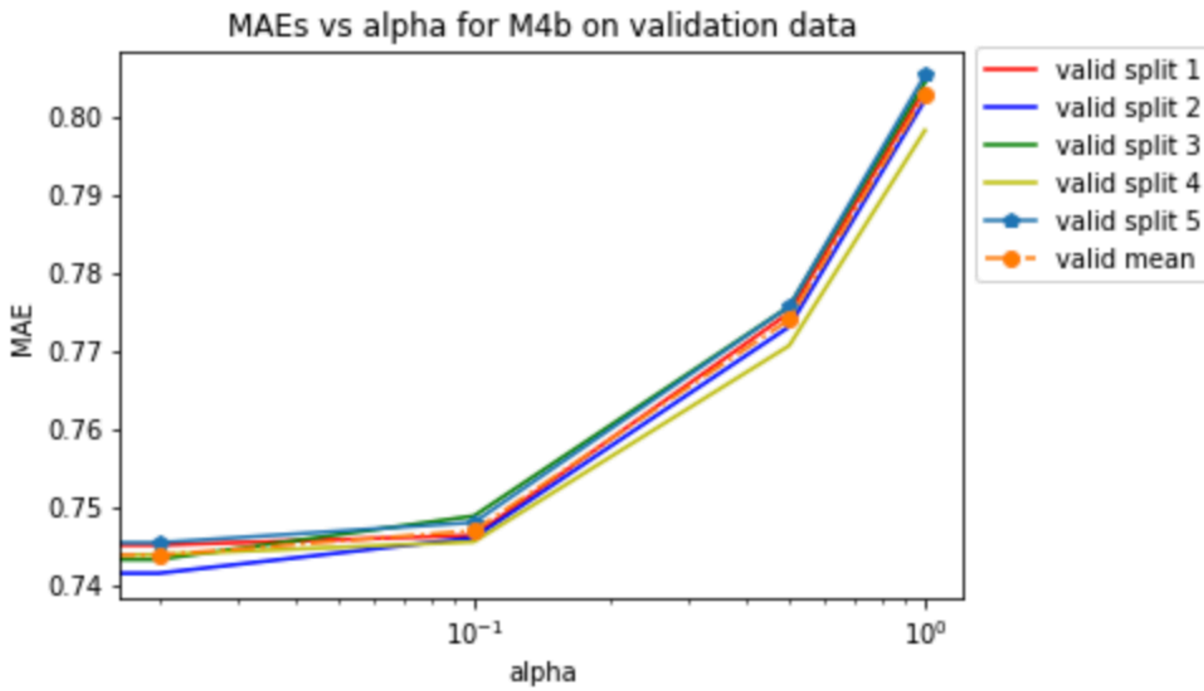

Fig 4b. Plot of Validation MAE over alpha for 5-fold cross validation using movie dataset of size 90000 with Surprise SVD. ***Alpha = [0,0.02,0.1,0.5,1]***

Based on the figure obtained and mean MAE across 5 folds, it seemed that alpha value of 0.02 resulted in the best predictive performance. At alpha values lower than 0.02, underfitting occurs while at K values higher than 0.02, we see that overfitting occurs as MAE increases again.

Problem 4b:

For this problem, the Surprised SVD with K = 50 and alpha = 0.02 was trained using 5-fold cross validation data. Final MAE obtained was **0.74109739.**

When compared to M3 (MAE = 0.73537), we see that **Surprised SVD performs worse**. One possible difference may be due to the gradient descent algorithm between the implemented M3 and M4's Surprise which allows M3 to reach a lower MAE.

A potential improvement to determine whether this is true is to run the models for more epochs (i.e: 2000+) and remeasure MAE at epoch 2000.

Phan Anh Nguyen
Comp135

<u>Problem 5a:</u>

For this problem, I chose to use and tune another algorithm within the Surprise package for Matrix Factorization called "SVDpp". Compared to the base SVD algorithm, this algorithm takes into consideration implicit ratings. Implicit ratings take into account other factors such as the fact that a use chose to watch certain types of movies and 'rate' certain types of movies (i.e: If they feel strongly about a certain type of movie). This may allow higher-level relationships between user and movies that were previously not captured in the base SVD algorithm to be determined, leading to more accurate predictions.

The figures below show the optimization process for K and alpha.

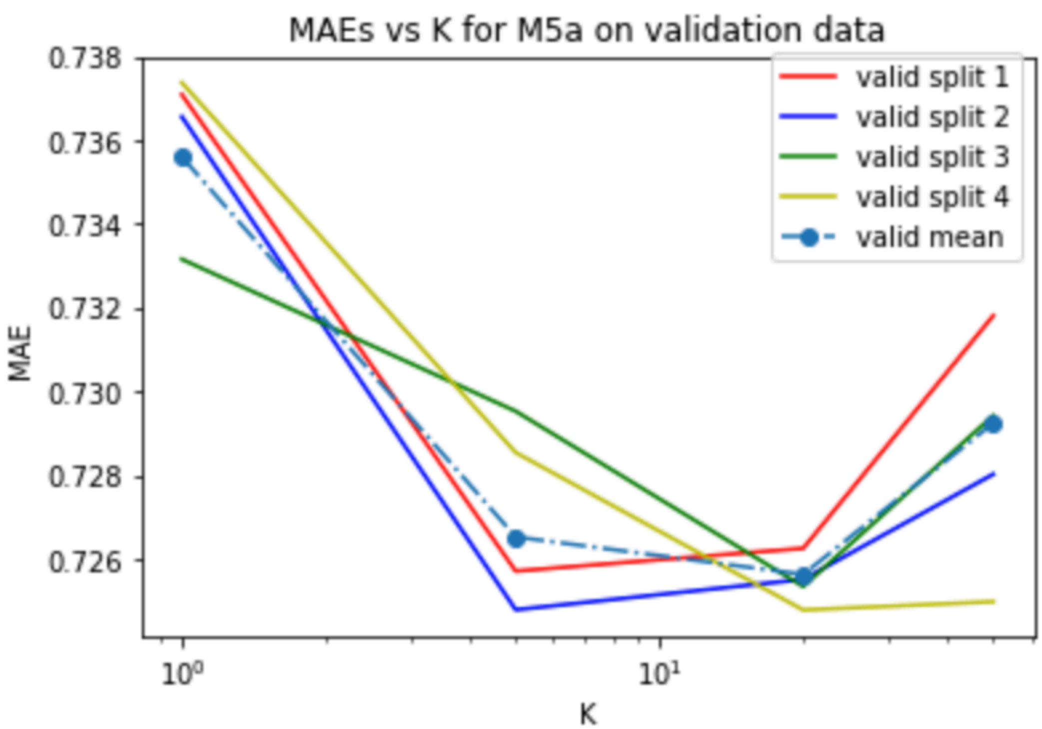

Fig 5a. Plot of Validation MAE over K for 5-fold cross validation using movie dataset of size 90000 with Surprise SVD. ***K = [1,5,20,50].***

Based on the figure obtained and the mean MAE across 5 folds, it seemed that K value of 20 resulted in the best predictive performance. At K values lower than 20, underfitting occurs while at K values higher than 20, we see that overfitting occurs as MAE increases again.

Phan Anh Nguyen
Comp135

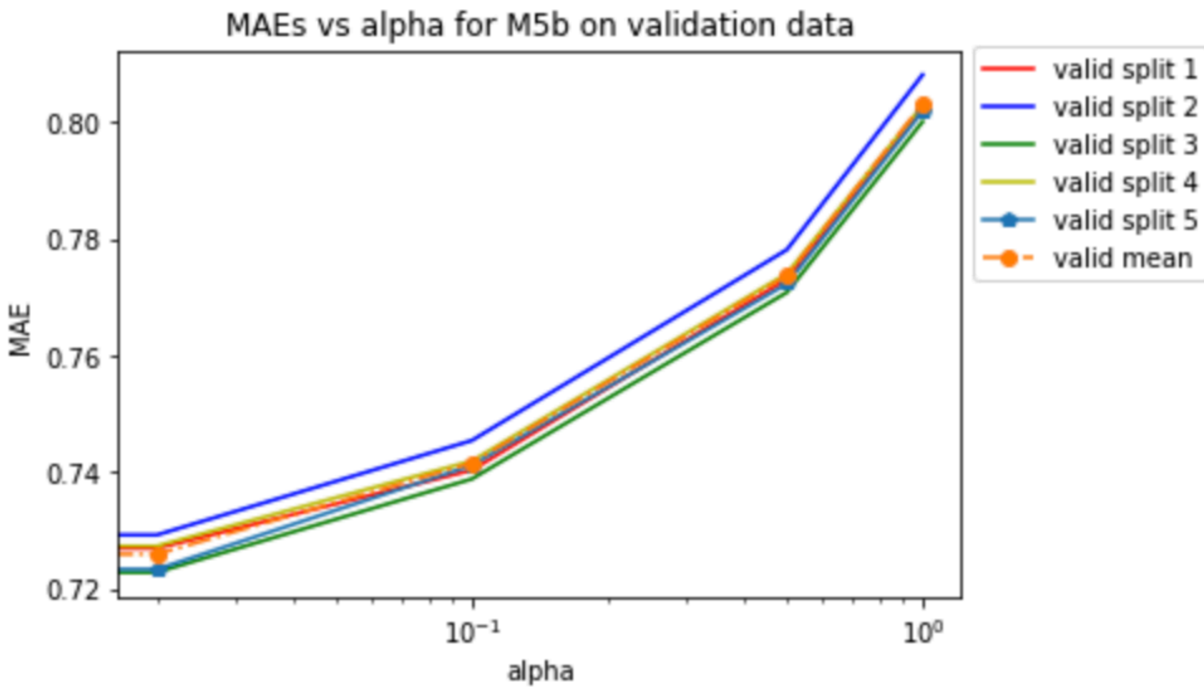

Fig 5b. Plot of Validation MAE over alpha for 5-fold cross validation using movie dataset of size 90000 with Surprise SVD. ***Alpha = [0,0.02,0.1,0.5,1]***

Based on the figure obtained and mean MAE across 5 folds, it seemed that alpha value of 0.02 resulted in the best predictive performance. At alpha values lower than 0.02, underfitting occurs while at K values higher than 0.02, we see that overfitting occurs as MAE increases again.

Ultimately, the Surprised SVD++ with K = 20 and alpha = 0.02 was trained using 5-fold cross validation data. Final MAE obtained was **0.72468302**

When compared to M3 (MAE = 0.73537) and M4 (MAE = 0.74109739), we see that **Surprised SVD++ is the best performing model.** Once again, the possible difference may be due to the ability of SVD++ being able to capture implicit ratings leading to more accurate predictions.

On leaderboard test set, my observed **MAE was 0.7151**. This is better than on validation set which suggests that overfitting was not observed.