Homework 5

Problem 1:

1. Load the data into R

There are **807 songs** in Songs.csv, **2421 users** in Users.csv and the range of ratings goes from **1** to **3.43**. After setting the seed (set.seed(345)), we split the dataset with the dimensions:

* Training set with 84% of observations (Train)
* Validation set A for tuning the collaborative filtering model with 4% of observations (Val1)
* Validation set B to be used for blending with 4% of observations (Val2)
* Test set with 8% of observation (Test)

1. Let’s create the ratings matrix

i) For this model:

Number of parameters: 2421 + 807 = 3228

Number of observations: 2421 \* 807 = 243103

ii) We use the **biscale** standardize a matrix to have optionally row means zero and variances one, and/or column means zero and variances one.

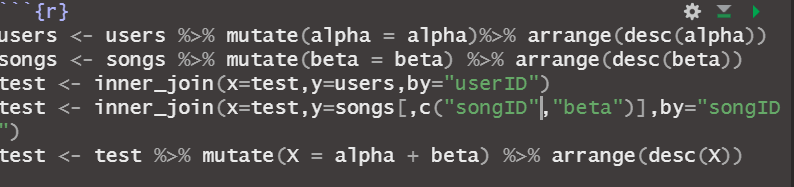
After removing the user affinity for rating songs highly (or lowly), we use **mutate** to join alpha and beta respectively to their dataset users and songs, then we use **inner\_join** to join them by userID and songID to the test set. We sum up the two new columns then we have Xij.

Three most popular songs:

We simply look for the songs with the highest Beta since it’s the rating without the bias of the user

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Rank | SongID | Song Name | Artist Name | Beta |
| 1 | 54 | You’re The One | Dwight Yoakam | 1.71 |
| 2 | 26 | Undo | Bjork | 1.69 |
| 3 | 439 | Secrets | One Republic | 1.64 |

Here is the code



iii) The three users that are most enthused about songs after removing the bias due to the effect of the popularity of songs are:

|  |  |  |
| --- | --- | --- |
| Rank | usersID | alpha |
| 1 | 1540 | 0.59 |
| 2 | 838 | 0.49 |
| 3 | 1569 | 0.47 |

iv) Performances on the test set:

We will use the metrics OSR, RMSE and MAE to assess the performances of the model on the test set

|  |  |
| --- | --- |
|  | **Collaborative Filtering** |
| **OSR** |  |
| **RMSE** |  |
| **MAE** |  |

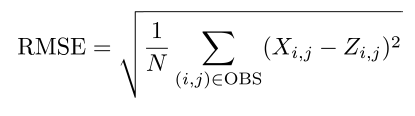
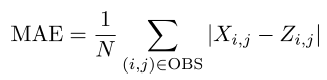
1. Let’s consider the following model

i) Number of parameters

We will train the model on the same training set, thus we still have 243103 observations.

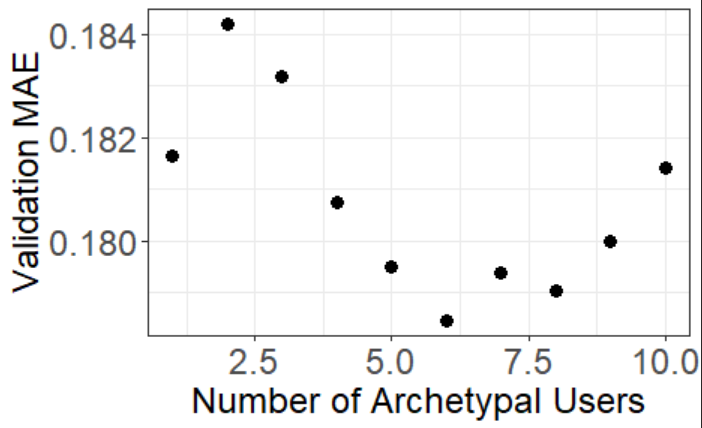
ii) Number of archetypes

The performances would be measured by the metrics



Remark: Here we normalize MAE and RMSE by the scale of the ratings which is 3.43 – 1 = 2.43

We will find the number of archetypes that minimizes the MAE



We choose k=6

iii) Final collaborative model

|  |  |  |
| --- | --- | --- |
|  | **CF** | **CF (k=6)** |
| **OSR** |  |  |
| **RMSE** |  |  |
| **MAE** |  |  |

Comments:

We see that we sligthly decrease the MAE while the coefficient of determination sligthly decrease too.

1. Add others features associated with songs

i) In this part, we will fit Random Forest and a Linear Regression model based on the independent variables: “genre” and “year”. The goal is to build a final ensemble model that will also catch the specificities of each song for the predicition.

Remark: Before trainning our algorithm we set those two independent variables as factor.

* Random Forest: Mtry = 1; num.trees = 500
* Linear Regression:

|  |  |  |
| --- | --- | --- |
|  | **Linear Regression** | **Random Forest** |
| **OSR** |  |  |
| **RMSE** |  |  |
| **MAE** |  |  |

ii) We use the validation set B to perfom blending of the collaborative filtering model, Linear Regression and Random Forest. Here are the results

|  |  |  |  |
| --- | --- | --- | --- |
|  | **CF (k=6)** | **Blending** | **Increase** |
| **OSR** |  |  | **12.6%** |
| **RMSE** |  |  | **-1%** |
| **MAE** |  |  | **-1.2%** |

Interpretation:

We observe that the MAE is almost the same despite a small increase of the OSR. Thus, blending the collaborative filtering model with other model doesn’t add a lot of predictive power on top of the collaborative model. However, we can imagine tuning the parameters (features selection for Linear Regression and cross validation for Random Forest) to sharp our added models.

