

main

Group4

4/9/2018

Step 0: Load the packages

```
if (!require("kableExtra")) install.packages("kableExtra")
```

```
## Loading required package: kableExtra
```

```
## Warning: package 'kableExtra' was built under R version 3.4.4
```

```
#default the wd to the fold this rmd file exists
```

Step 1: Load and process the data

```
MS_train    <- read.csv("../data/data_sample/MS_sample/data_train.csv")
MS_test     <- read.csv("../data/data_sample/MS_sample/data_test.csv")
movie_train <- read.csv("../data/data_sample/eachmovie_sample/data_train.csv")
movie_test  <- read.csv("../data/data_sample/eachmovie_sample/data_test.csv")
```

Step 2 : Transformation

Convert the original dataset to a matrix which rows represents users and columns represents items For dataset 1 (MS), we assign 0 to those items which users never visited. For dataset 2 (Movie), we assign NA to those items which users never rated.

```
source("../lib/MemoryBased.R")
```

```
MS_train <- Transform_ms(MS_train)
```

```
MS_test  <- Transform_ms(MS_test)
```

```
# save(MS_train, file = "../output/MS_train.RData")
```

```
# save(MS_test, file = "../output/MS_test.RData")
```

```
movie_train <- Transform_m(movie_train)
```

```
movie_test  <- Transform_m(movie_test)
```

```
# save(movie_train, file = "../output/movie_train.RData")
```

```
# save(movie_test, file = "../output/movie_test.RData")
```

Memory-based Algorithm

Step 3 : Similarity Weight

Pearson Correlation & Mean-square-difference & SimRank

```
load("../output/MS_train.RData")
load("../output/MS_test.RData")
load("../output/movie_train.RData")
load("../output/movie_test.RData")

##For dataset 1 (MS)

##Pearson Correlation
# ms_pc <- pearson_corr(MS_train)
# save(ms_pc, file = "../out/put/ms_pc.RData")

##Mean-square-difference
# ms_msd <- MSD_Weight(MS_train)
# save(ms_msd, file = "../output/ms_msd.RData")

##SimRank
# ms_sr <- simrank(MS_train)
# save(ms_sr, file = "../output/simrank_MS_train.RData")

##For dataset 2 (Movie)

##Pearson Correlation
# movie_pc <- pearson_corr(movie_train)
# save(movie_pc, file = "../output/movie_pc.RData")

##Mean-square-difference
# movie_msd <- MSD_Weight(movie_train)
# save(movie_msd, file = "../output/movie_msd.RData")
```

No Variance Weighting

Step 4: Selecting Neighbors

```
## Implementation on Dataset 1

## MSD + WT
# ms_msd_wt_0.05 <- corr_thresh(ms_msd, 0.05)
# save(ms_msd_wt_0.05, file = "../output/Selecting_Neighbors_Results/ms_msd_wt_0.05.RData")
```

```

# ms_msd_wt_0.2 <- corr_thresh(ms_msd, 0.2)
# save(ms_msd_wt_0.2, file = "../output/Selecting_Neighbors_Results/ms_msd_wt_0.2.RData")

## MSD + BNN
# ms_msd_bnn_20 <- Select_BNN(ms_msd, 20)
# save(ms_msd_bnn_20, file = "../output/Selecting_Neighbors_Results/ms_msd_bnn_20.RData")
# ms_msd_bnn_40 <- Select_BNN(ms_msd, 40)
# save(ms_msd_bnn_40, file = "../output/Selecting_Neighbors_Results/ms_msd_bnn_40.RData")

## MSD + combine
# ms_msd_combine_0.05_40 <- combine(ms_msd, 0.05, 40)
# save(ms_msd_combine, file = "../output/Selecting_Neighbors_Results/ms_msd_combine_0.05_40.RData")

## PC + WT
# ms_pc_wt_0.05 <- corr_thresh(ms_pc, 0.05)
# save(ms_pc_wt_0.05, file = "../output/Selecting_Neighbors_Results/ms_pc_wt_0.05.RData")
# ms_pc_wt_0.2 <- corr_thresh(ms_pc, 0.2)
# save(ms_pc_wt_0.2, file = "../output/Selecting_Neighbors_Results/ms_pc_wt_0.2.RData")

## PC + BNN
# ms_pc_bnn_20 <- Select_BNN(ms_pc, 20)
# save(ms_pc_bnn_20, file = "../output/Selecting_Neighbors_Results/ms_pc_bnn_20.RData")
# ms_pc_bnn_40 <- Select_BNN(ms_pc, 40)
# save(ms_pc_bnn_40, file = "../output/Selecting_Neighbors_Results/ms_pc_bnn_40.RData")

## PC + combine
# ms_pc_combine_0.05_40 <- combine(ms_pc, 0.005, 40)
# save(ms_pc_combine_0.05_40, file = "../output/Selecting_Neighbors_Results/ms_pc_combine_0.05_40.RData")

## Simrank + WT
# ms_sr_wt_0.05 <- corr_thresh(ms_sr, 0.05)
# save(ms_sr_wt_0.05, file = "../output/Selecting_Neighbors_Results/ms_pc_wt_0.05.RData")
# ms_sr_wt_0.2 <- corr_thresh(ms_sr, 0.2)
# save(ms_sr_wt_0.2, file = "../output/Selecting_Neighbors_Results/ms_pc_wt_0.2.RData")

```

```

## Simrank + BNN
# ms_sr_bnn_20 <- Select_BNN(ms_sr, 20)
# save(ms_sr_bnn_20, file = "../output/Selecting_Neighbors_Results/ms_sr_bnn_20.RData")
# ms_sr_bnn_40 <- Select_BNN(ms_sr, 40)
# save(ms_sr_bnn_40, file = "../output/Selecting_Neighbors_Results/ms_sr_bnn_40.RData")

## Simrank + combine
# ms_sr_combine_0.05_40 <- combine(ms_sr, 0.005, 40)
# save(ms_sr_combine_0.05_40, file = "../output/Selecting_Neighbors_Results/ms_sr_combine_0.05_40.RData")

## Implementation on Dataset 2

## MSD + WT
# movie_msd_wt_0.05 <- corr_thresh(movie_msd, 0.05)
# save(movie_msd_wt_0.05, file = "../output/Selecting_Neighbors_Results/movie_msd_wt_0.05.RData")
# movie_msd_wt_0.2 <- corr_thresh(movie_msd, 0.2)
# save(movie_msd_wt_0.2, file = "../output/Selecting_Neighbors_Results/movie_msd_wt_0.2.RData")

## MSD + BNN
# movie_msd_bnn_20 <- Select_BNN(movie_msd, 20)
# save(movie_msd_bnn_20, file = "../output/Selecting_Neighbors_Results/movie_msd_bnn_20.RData")
# movie_msd_bnn_40 <- Select_BNN(movie_msd, 40)
# save(movie_msd_bnn_40, file = "../output/Selecting_Neighbors_Results/movie_msd_bnn_40.RData")

## MSD + combine
# movie_msd_combine_0.05_40 <- combine(movie_msd, 0.005, 40)
# save(movie_msd_combine_0.05_40, file = "../output/Selecting_Neighbors_Results/movie_msd_combine_0.05_40.RData")

## PC + WT
# movie_pc_wt_0.05 <- corr_thresh(movie_pc, 0.05)
# save(movie_pc_wt_0.05, file = "../outputSelecting_Neighbors_Results//movie_pc_wt_0.05.RData")
# movie_pc_wt_0.2 <- corr_thresh(movie_pc, 0.2)
# save(movie_pc_wt_0.2, file = "../outputSelecting_Neighbors_Results//movie_pc_wt_0.2.RData")

```

```
## PC + BNN
# movie_pc_bnn_20 <- Select_BNN(movie_pc, 20)
# save(movie_pc_bnn_20, file = "../output/Selecting_Neighbors_Results/movie_pc_bnn_20.RData")
# movie_pc_bnn_40 <- Select_BNN(movie_pc, 40)
# save(movie_pc_bnn_40, file = "../output/Selecting_Neighbors_Results/movie_pc_bnn_40.RData")

# PC + combine
# movie_pc_combine_0.05_40 <- combine(movie_pc, 0.005, 40)
# save(movie_pc_combine_0.05_40, file = "../output/Selecting_Neighbors_Results/movie_pc_combine_0.05_40.RData")
```

Step 5 : Prediction

```
## Implementation on Dataset 1

## MSD + WT
# pred_ms_msd_wt_0.05 <- avg_dev_pred(MS_train,MS_test,ms_msd, ms_msd_wt_0.05)
# save(pred_ms_msd_wt_0.05, "../output/Prediction_Results/pred_ms_msd_wt_0.05.RData")
# pred_ms_msd_wt_0.2 <- avg_dev_pred(MS_train,MS_test,ms_msd, ms_msd_wt_0.2)
# save(pred_ms_msd_wt_0.2, "../output/Prediction_Results/pred_ms_msd_wt_0.2.RData")

## MSD + BNN
# ZScore_ms_msd_bnn_20 <- ZScore_Mat(ms_msd, ms_msd_bnn_20, MS_train, MS_test)
# save(ZScore_ms_msd_bnn_20, "../output/Prediction_Results/ZScore_ms_msd_bnn_20.RData")
# ZScore_ms_msd_bnn_40 <- ZScore_Mat(ms_msd, ms_msd_bnn_40, MS_train, MS_test)
# save(ZScore_ms_msd_bnn_40, "../output/Prediction_Results/ZScore_ms_msd_bnn_40.RData")

## MSD + combine
# pred_ms_msd_combine_0.05_40 <- avg_dev_pred(MS_train,MS_test,ms_msd, ms_msd_combine_0.05_40)
# save(pred_ms_msd_combine_0.05_40, "../output/Prediction_Results/pred_ms_msd_combine_0.05_40.RData")

## PC + WT
# pred_ms_pc_wt_0.05 <- avg_dev_pred(MS_train,MS_test,ms_pc, ms_pc_wt_0,05)
# save(pred_ms_pc_wt_0.05, "../output/Prediction_Results/pred_ms_pc_wt_0.05.RData")
# pred_ms_pc_wt_0.2 <- avg_dev_pred(MS_train,MS_test,ms_pc, ms_pc_wt_0,2)
# save(pred_ms_pc_wt_0.2, "../output/Prediction_Results/pred_ms_pc_wt_0.2.RData")

## PC + BNN
```

```

# ZScore_ms_pc_bnn_20<- ZScore_Mat(ms_pc, ms_pc_bnn_20, MS_train, MS_test)
# save(ZScore_ms_pc_bnn_20, "../output/Prediction_Results/ZScore_ms_pc_bnn_20.RData")
# ZScore_ms_pc_bnn_40 <- ZScore_Mat(ms_pc, ms_pc_bnn_40, MS_train, MS_test)
# save(ZScore_ms_pc_bnn_40, "../output/Prediction_Results/ZScore_ms_pc_bnn_40.RData")

## PC + combine
# pred_ms_pc_combine_0.05_40 <- avg_dev_pred(MS_train,MS_test,ms_pc, ms_pc_combine_0.
05_40)
# save(pred_ms_pc_combine_0.05_40, "../output/Prediction_Results/pred_ms_pc_combine_0
.05_40.RData")

## Simrank + WT
# pred_ms_sr_wt_0.05 <- avg_dev_pred(MS_train,MS_test,ms_sr, ms_sr_wt_0.05)
# save(pred_ms_sr_wt_0.05, "../output/Prediction_Results/pred_ms_sr_wt_0.05.RData")
# pred_ms_sr_wt_0.2 <- avg_dev_pred(MS_train,MS_test,ms_sr, ms_sr_wt_0.2)
# save(pred_ms_sr_wt_0.2, "../output/Prediction_Results/pred_ms_sr_wt_0.2.RData")

## Simrank + BNN
# ZScore_ms_sr_bnn_20<- ZScore_Mat(ms_sr, ms_sr_bnn_20, MS_train, MS_test)
# save(pred_ms_sr_bnn_20, "../output/Prediction_Results/pred_ms_sr_bnn_20.RData")
# ZScore_ms_sr_bnn_40<- ZScore_Mat(ms_sr, ms_sr_bnn_40, MS_train, MS_test)
# save(pred_ms_sr_bnn_40, "../output/Prediction_Results/pred_ms_sr_bnn_40.RData")

## Simrank + combine
# pred_ms_sr_combine_0.05_40 <- avg_dev_pred(MS_train,MS_test, ms_sr, ms_sr_combine_0
.05_40)
# save(pred_ms_sr_combine_0.05_40, "../output/Prediction_Results/pred_ms_sr_combine_0
.05_40.RData")

## Implementation on Dataset 2

## MSD + WT
# pred_movie_msd_wt_0.05 <- avg_dev_pred(movie_train,movie_test,movie_msd, movie_msd_
wt_0.05)
# save(pred_movie_msd_wt_0.05, "../output/Prediction_Results/pred_movie_msd_wt_0.05.R
Data")
# pred_movie_msd_wt_0.2 <- avg_dev_pred(movie_train,movie_test,movie_msd, movie_msd_w
t_0.2)
# save(pred_movie_msd_wt_0.2, "../output/Prediction_Results/pred_movie_msd_wt_0.2.RDa
ta")

## MSD + BNN
# ZScore_movie_msd_bnn_20<- ZScore_Mat(movie_msd, movie_msd_bnn_20, movie_train, movi
e_test)

```

```

# save(ZScore_movie_msd_bnn_20, "../output/Prediction_Results/ZScore_movie_msd_bnn_20.RData")
# ZScore_movie_msd_bnn_40<- ZScore_Mat(movie_msd, movie_msd_bnn_40, movie_train, movie_test)
# save(ZScore_movie_msd_bnn_40, "../output/Prediction_Results/ZScore_movie_msd_bnn_40.RData")

## MSD + combine
# pred_movie_msd_combine_0.05_40 <- avg_dev_pred(movie_train,movie_test,movie_msd, movie_msd_combine_0.05_40)
# save(pred_movie_msd_combine_0.05_40, "../output/Prediction_Results/pred_movie_msd_combine_0.05_40.RData")

## PC + WT
# pred_movie_pc_wt_0.05 <- avg_dev_pred(movie_train,movie_test,movie_pc, movie_pc_wt_0.05)
# save(pred_movie_pc_wt_0.05, "../output/Prediction_Results/pred_movie_pc_wt_0.05.RData")
# pred_movie_pc_wt_0.2 <- avg_dev_pred(movie_train,movie_test,movie_pc, movie_pc_wt_0.2)
# save(pred_movie_pc_wt_0.2, "../output/Prediction_Results/pred_movie_pc_wt_0.2.RData")

## PC + BNN
# ZScore_movie_pc_bnn_20 <- ZScore_Mat(movie_pc, movie_pc_bnn_20, movie_train, movie_test)
# save(ZScore_movie_pc_bnn_20, "../output/Prediction_Results/ZScore_movie_pc_bnn_20.RData")
# ZScore_movie_pc_bnn_40 <- ZScore_Mat(movie_pc, movie_pc_bnn_40, movie_train, movie_test)
# save(ZScore_movie_pc_bnn_40, "../output/Prediction_Results/ZScore_movie_pc_bnn_40.RData")

# PC + combine
# pred_movie_pc_combine_0.05_40 <- avg_dev_pred(movie_train,movie_test,movie_pc, movie_pc_combine_0.05_40)
# save(pred_movie_pc_combine_0.05_40, "../output/Prediction_Results/pred_movie_pc_combine_0.05_40.RData")

```

Step 6 : Valuation

```

## Implementation on Dataset 1: ranked scoring

## MSD + WT
# load("../output/Prediction_Results/pred_ms_msd_wt_0.05.RData")
# RS_ms_msd_wt_0.05 <- Rank_Score(pred_ms_msd_wt_0.05, MS_test)
# RS_ms_msd_wt_0.05
# load("../output/Prediction_Results/pred_ms_msd_wt_0.2.RData")

```

```

# RS_ms_msd_wt_0.2 <- Rank_Score(pred_ms_msd_wt_0.2, MS_test)
# RS_ms_msd_wt_0.2

## MSD + BNN
# load("../output/Prediction_Results/ZScore_ms_msd_bnn_20.RData")
# RS_ms_msd_bnn_20 <- Rank_Score(ZScore_ms_msd_bnn_20, MS_test)
# RS_ms_msd_bnn_20
# load("../output/Prediction_Results/ZScore_ms_msd_bnn_40.RData")
# RS_ms_msd_bnn_40 <- Rank_Score(ZScore_ms_msd_bnn_40, MS_test)
# RS_ms_msd_bnn_40

## MSD + combine
# load("../output/Prediction_Results/ZScore_ms_msd_combine_0.05_40.RData")
# RS_ms_msd_combine_0.05_40 <- Rank_Score(pred_ms_msd_combine_0.05_40, MS_test)
# RS_ms_msd_combine_0.05_40

## PC + WT
# load("../output/Prediction_Results/pred_ms_pc_wt_0.05.RData")
# RS_ms_pc_wt_0.05 <- Rank_Score(pred_ms_pc_wt_0.05, MS_test)
# RS_ms_pc_wt_0.05
# load("../output/Prediction_Results/pred_ms_pc_wt_0.2.RData")
# RS_ms_pc_wt_0.2 <- Rank_Score(pred_ms_pc_wt_0.2, MS_test)
# RS_ms_pc_wt_0.2

## PC + BNN
# load("../output/Prediction_Results/ZScore_ms_pc_bnn_20.RData")
# RS_ms_pc_bnn_20 <- Rank_Score(ZScore_ms_pc_bnn_20, MS_test)
# RS_ms_pc_bnn_20
# load("../output/Prediction_Results/ZScore_ms_pc_bnn_40.RData")
# RS_ms_pc_bnn_40 <- Rank_Score(ZScore_ms_pc_bnn_40, MS_test)
# RS_ms_pc_bnn_40

## PC + combine
# load("../output/Prediction_Results/pred_ms_pc_combine_0.05_40.RData")
# RS_ms_pc_combine_0.05_40 <- Rank_Score(pred_ms_pc_combine_0.05_40, MS_test)
# RS_ms_pc_combine_0.05_40

## Simrank + WT
# load("../output/Prediction_Results/pred_ms_sr_wt_0.05.RData")
# SR_ms_sr_wt_0.05 <- Rank_Score(pred_ms_sr_wt_0.05, MS_test)
# SR_ms_sr_wt_0.05
# load("../output/Prediction_Results/pred_ms_sr_wt_0.2.RData")
# SR_ms_sr_wt_0.2 <- Rank_Score(pred_ms_sr_wt_0.2, MS_test)
# SR_ms_sr_wt_0.2

## Simrank + BNN
# load("../output/Prediction_Results/ZScore_ms_sr_bnn_20.RData")
# RS_ms_sr_bnn_20 <- Rank_Score(ZScore_ms_sr_bnn_20, MS_test)

```



```

# RS_ms_sr_bnn_20
# load("../output/Prediction_Results/ZScore_ms_sr_bnn_40.RData")
# RS_ms_sr_bnn_40 <- Rank_Score(ZScore_ms_sr_bnn_40, MS_test)
# RS_ms_sr_bnn_40

## Simrank + combine
# load("../output/Prediction_Results/pred_ms_sr_combine_0.05_40.RData")
# RS_ms_sr_combine_0.05_40 <- Rank_Score(pred_ms_sr_combine_0.05_40, MS_test)
# RS_ms_sr_combine_0.05_40

### Implementation on Dataset 2: MAE

## MSD + WT
# load("../output/Prediction_Results/pred_movie_msd_0.05.RData")
# MAE_movie_msd_wt_0.05 <- MAE(pred_movie_msd_wt_0.05, movie_test)
# MAE_movie_msd_wt_0.05
# load("../output/Prediction_Results/pred_movie_msd_0.2.RData")
# MAE_movie_msd_wt_0.2 <- MAE(pred_movie_msd_wt_0.2, movie_test)
# MAE_movie_msd_wt_0.2

## MSD + BNN
# load("../output/Prediction_Results/ZScore_movie_msd_bnn_20.RData")
# MAE_movie_msd_bnn_20 <- MAE(ZScore_movie_msd_bnn_20, movie_test)
# MAE_movie_msd_bnn_20
# load("../output/Prediction_Results/ZScore_movie_msd_bnn_40.RData")
# MAE_movie_msd_bnn_40 <- MAE(ZScore_movie_msd_bnn_40, movie_test)
# MAE_movie_msd_bnn_40

## MSD + combine
# load("../output/Prediction_Results/pred_movie_msd_combine_0.05_40.RData")
# MAE_movie_msd_combine_0.05_40 <- MAE(pred_movie_msd_combine_0.05_40, movie_test)
# MAE_movie_msd_combine_0.05_40

## PC + WT
# load("../output/Prediction_Results/pred_movie_pc_wt_0.05.RData")
# MAE_movie_pc_wt_0.05 <- MAE(pred_movie_pc_wt_0.05, movie_test)
# MAE_movie_pc_wt_0.05
# load("../output/Prediction_Results/pred_movie_pc_wt_0.2.RData")
# MAE_movie_pc_wt_0.2 <- MAE(pred_movie_pc_wt_0.2, movie_test)
# MAE_movie_pc_wt_0.2

## PC + BNN
# load("../output/Prediction_Results/ZScore_movie_pc_bnn_20.RData")
# MAE_movie_pc_bnn_20 <- MAE(ZScore_movie_pc_bnn_20, movie_test)

```

```

# MAE_movie_pc_bnn_20
# load("../output/Prediction_Results/ZScore_movie_pc_bnn_40.RData")
# MAE_movie_pc_bnn_40 <- MAE(ZScore_movie_pc_bnn_40, movie_test)
# MAE_movie_pc_bnn_40

## PC + combine
# load("../output/Prediction_Results/pred_movie_pc_combine_0.05_40.RData")
# MAE_movie_pc_combine_0.05_40 <- MAE(pred_movie_pc_combine_0.05_40, movie_test)
# MAE_movie_pc_combine_0.05_40

```

Model-based Algorithm

Step 3: Cluster Model

```

load("../output/movie_train.RData")
load("../output/movie_test.RData")
train <- movie_train
test <- movie_test

N <- nrow(train)
M <- ncol(train)

user <- rownames(train)
movie <- colnames(train)

### cluster model
em_fun <- function(data, C, thres){
  #Input: train_data, number of classes, threshold to determine convergence
  #Output: parameters for cluster models:
  # mu: probability of belonging to class c, vector
  # gamma: probability of scores for a movie given the class, 3 dimentions

  #####
  # Step 1 - initialization
  #####
  set.seed(2)
  mu <- runif(C)
  mu <- mu/sum(mu)
  gamma <- array(NA,c(M,C,6)) #each matrix represents a class
  #the i,j-th element means the probability of rating jth movie with score i in the c
lass
  for(m in 1:M){
    for(c in 1:C){
      gamma[m,c,] <- runif(6)
      gamma[m,c,] <- gamma[m,c,]/sum(gamma[m,c,])
    }
  }
}

```

```

    }
  }

v <- array(0, c(M,N,7))
for(k in 1:6){
  v[,,k] <- ifelse(t(data)==(k-1), 1, 0)
  v[,,k] <- ifelse(is.na(v[,,k]), 0, v[,,k])
  v[,,7] <- v[,,7] + v[,,k]
}

mu_new <- mu
gamma_new <- gamma

## Iterations based on the stop criterion
thres1 <- 1000
thres2 <- 1000
thres1_new <- 0
thres2_new <- 0
count <- 0

while((thres1>thres|thres2>thres)&(abs(thres1-thres1_new)>thres|abs(thres2-thres2_new)>thres))
{
  count <- count + 1
  print(paste0("iteration = ", count))

  thres1_new <- thres1
  thres2_new <- thres2

  mu <- mu_new
  gamma <- gamma_new

  #####
  # Step 2 - Expectation
  #####
  #expectation pi with rows meaning classes and columns meaning users
  phi <- matrix(0, C, N)

  for(k in 1:6){
    phi <- phi + t(log(gamma[,k]))%*%v[,k]
  }
  phi <- phi-rep(colMeans(phi),each=C)

  for(c in 1:C){
    phi[c,] <- mu[c]*exp(phi)[c,]
  }
  phi <- ifelse(phi == rep(colSums(phi),each=C), 1, phi/rep(colSums(phi), each=C))

  #####
  # Step 3 - Maximization

```

```

#=====
mu_new <- rowSums(phi)/N #update mu vector

for(k in 1:6){
  gamma_new[,k] <- v[,k]%*%t(phi)/v[,7]%*%t(phi) #update gamma
}

gamma_new[gamma_new == 0] <- 10^(-100)
if(sum(is.na(gamma_new)) != 0){
  is_zero <- which(is.na(gamma_new))
  gamma_new[is_zero] <- rep(1/6, length(is_zero))
}

## Check convergence
thres1 <- mean(abs(mu_new - mu)) #mean absolute difference of mu
thres2 <- 0
for(c in 1:C){
  thres2 <- max(thres2, norm(as.matrix(gamma_new[,c,] - gamma[,c,]), "O"))
}
print(paste0("threshold1 = ", thres1, " threshold2 = ", thres2))
}
return(list(mu = mu, gamma = gamma))
}

#predict score estimate function
cm_predict <- function(train_df, test_df, par){
  set.seed(2)
  mu <- par$mu
  gamma <- par$gamma
  C <- length(mu)

  v <- array(0, c(M,N,7))
  for(k in 1:6){
    v[,k] <- ifelse(t(train_df)==(k-1), 1, 0)
    v[,k] <- ifelse(is.na(v[,k]),0,v[,k])
  }
  v[,7] <- ifelse(!is.na(t(test_df)), 1, 0)

  ##using Naive Bayes formula
  prob <- array(0,c(N,M,7))
  prob_mu <- matrix(mu, N, C, byrow = TRUE)
  phi <- matrix(0, C, N)
  for(k in 1:6){
    phi <- phi + t(log(gamma[,k]))%*%v[,k]
  }

  phi <- exp(phi)

  den <- matrix(diag(prob_mu%*%phi), N, M, byrow=FALSE)
  #denominator in equation (2) of cluster model notes

```

```

for(k in 1:6){
  print(paste0("k = ", k))

  num <- (t(phi)*prob_mu)%*%t(gamma[,k]) #numerator in equation (2) of cluster model notes
  prob[,k] <- ifelse(num==den & num == 0, runif(1)/6, num/den)
  prob[,7] <- prob[,7] + k*prob[,k]
}
return(prob[,7]*t(v[,7]))
}

### 5-fold cross validation to find best class number C among c_list(2,3,6,12)

set.seed(2)
K <- 5
n <- ncol(train)
m <- nrow(train)
n.fold <- floor(n/K)
m.fold <- floor(m/K)
s <- sample(rep(1:K, c(rep(n.fold, K-1), n-(K-1)*n.fold)))
s1 <- sample(rep(1:K, c(rep(m.fold, K-1), m-(K-1)*m.fold)))

c_list <- c(2,3,6,12)
validation_error <- matrix(NA, K, length(c_list))

train_data <- data.frame(matrix(NA, N, M))
colnames(train_data) <- movie
rownames(train_data) <- user

test_data <- data.frame(matrix(NA, N, M))
colnames(test_data) <- movie
rownames(test_data) <- user

#cv 5 folds
#calculate cv error
cv_fun <- function(train_data,test_data){
  for(i in 1:K){
    train_data[s1 != i, ] <- train[s1 != i, ]
    train_data[s1 == i, s != i] <- train[s1==i, s != i]
    test_data[s1 == i,s == i] <- train[s1 == i ,s == i]
    #write.csv(train_data,paste0("../output/cluster_model_subtrain.csv"))
    #write.csv(test_data,paste0("../output/cluster_model_validation.csv"))

    estimate_data <- test_data

    for(c in 1:length(c_list)){

```

```

cm_par <- em_fun(data = train_data, C = c_list[c], thres = 0.05)
estimate_data <- cm_predict(train_df = train_data, test_df = test_data, par = cm_
par)
validation_error[i,c] <- sum(abs(estimate_data-test_data),na.rm = T)/sum(!is.na(e
stimate_data-test_data))

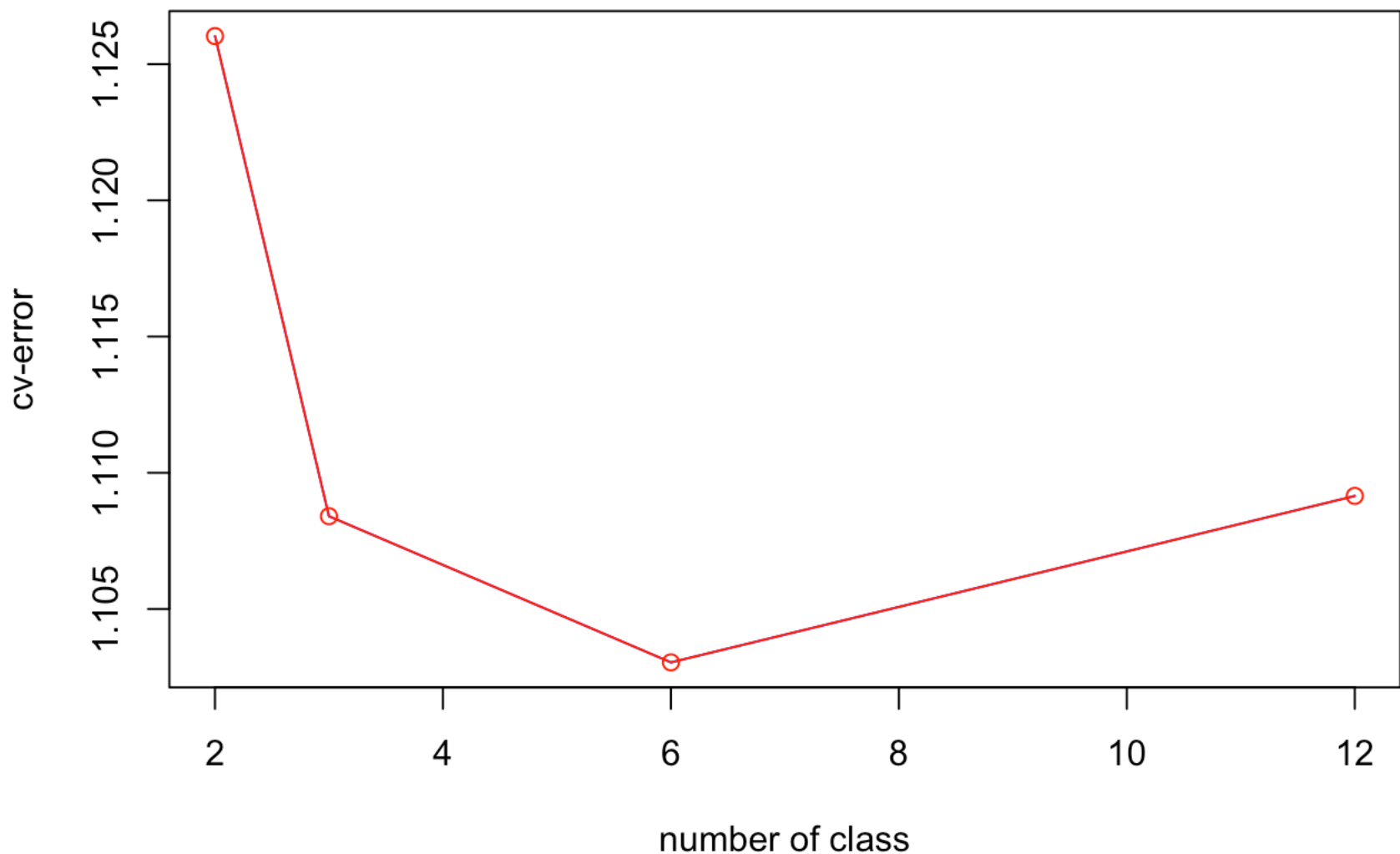
}}
return(validation_error)
}

#validation_error <- cv_fun(train_data,test_data)
#save(validation_error, file=paste0("../output/validation_err.RData"))

# Cluster number comparism
load("../output/validation_err.RData")
cv_error<-colMeans(validation_error)

# setwd("../figs/")
# jpeg(file=paste("cv_err",".jpg") )
plot(c_list,cv_error,xlab="number of class",ylab="cv-error",col="blue",type="l")
points(c_list,cv_error,col="red",type="o")

```



```
#dev.off()
```

```
class = c_list[which.min(cv_error)]  
print(paste("Best class number is", class))
```

```
## [1] "Best class number is 6"
```

```
class <- 6
```

```
#best_par <- em_fun(data = train, C = class, thres = 0.01)  
#save(best_par, file = "../output/best_par.RData")
```

```
load("../output/best_par.RData")
```

```
###estimate scores
```

```
#estimate <- cm_predict(train, test, best_par)
```

```
#write.csv(estimate, paste0("../output/cluster_model_estimate.csv"))
```

```
estimate <- read.csv("../output/cluster_model_estimate.csv")  
estimate <- estimate[,-1]
```

```
# MAE of EM algorithm  
coltest <- colnames(test)  
colnames(estimate) <- movie
```

```
coltest <- which(is.element(movie, coltest))  
estimate <- estimate[,coltest]
```

```
MAE <- function(pred, true){  
  mae <- sum(abs(pred-test),na.rm = T)/sum(!is.na(abs(pred-test)))  
  return(mae)  
}
```

```
error_em<- MAE(estimate,test)  
error_em
```

```
## [1] 2.803069
```

Compare results

```
load("../output/results.RData")
##Data 1
library("kableExtra")
dt1 <- aa[[1]]
kable(dt1, "html") %>%
  kable_styling(c("striped", "bordered")) %>%
  add_header_above(c(" ", "Weight Threshold" = 2, "Best N Estimator" = 2, "Combined"
= 1)) %>%
  add_header_above(c("Dataset 1: Ranked Scoring" = 6))
```

Dataset 1: Ranked Scoring					
	Weight Threshold		Best N Estimator		Combined
	0.05	0.2	20	40	0.05+40
Pearson Correlation	45.72740	44.05147	32.68916	34.62473	33.70090
Mean Square Difference	47.77998	47.77998	33.38771	35.12559	35.04042
SimRank	45.98500	32.14530	32.37750	33.39524	35.05230

```
##Data 2
dt2 <- aa[[2]]
kable(dt2, "html") %>%
  kable_styling(c("striped", "bordered")) %>%
  add_header_above(c(" ", "Weight Threshold" = 2, "Best N Estimator" = 2, "Combined"
= 1)) %>%
  add_header_above(c("Dataset 2: MAE" = 6))
```

Dataset 2: MAE					
	Weight Threshold		Best N Estimator		Combined
	0.05	0.2	20	40	0.05+40
Pearson Correlation	1.187658	1.185793	1.167949	1.167326	1.183761
Mean Square Difference	1.168654	1.168654	1.170718	1.169313	1.166298

```
##Cluster mae
cat("Cluster Model MAE =", error_em)
```

```
## Cluster Model MAE = 2.803069
```