Project 2

Shipra Ahuja, Nathan Groom, Saheli Kar

June 24, 2016

Introduction

This project uses book ratings data to implement and configure two different recmmender systems: one using a hybrid collaborative filtering technique taking into account both the genre of the books and their user ratings, the other a content based recommendation system.

Collaborative Filtering

This function is essentially a hybrid collaborative recommender function which does the following: first, it calculates the overall average rating of each title (as rated by users) from 1-5 and places these ratings in a data frame. We then create an empty recommendation matrix with users as rows and book titles as columns. Finally we fill in the matrix with each user's recommendations, which are the top rated books, not yet read by that user, sorted with that user's preferred genre appearing first.

```
suppressPackageStartupMessages(library(pROC))

books <- read.csv("/users/nathangroom/desktop/books.csv",header=TRUE)
ratings <- read.csv("/users/nathangroom/desktop/Ratings.csv",header=FALSE)
books_num<-nrow(books)

#Add column names as ISBN of the books
names(ratings)<-c("User", as.character(books$ISBN))

#Get the average rating of each book
books=as.data.frame(cbind(books, avg_score=unname(mapply(mean, ratings[,2:(books_num+1)]))))

readers=ratings[,1]
categories<-unique(books$Category1)

# Matrix to hold the category preference of users
categories_ratings_matrix<-matrix(0,nrow=length(readers), ncol=length(categories),
                        dimnames=list(readers,categories))

Authors<- unique(books$Author)

# Matrix to hold the Athor preference of users
Authors_ratings_matrix<-matrix(0,nrow=length(readers), ncol=length(Authors),
                        dimnames=list(readers,Authors))



for(rownum in (1:nrow(ratings))) {
```

```
  for (colNum in (2:(books_num+1))){
    readerName<-as.character(ratings[rownum,1])
    if(ratings[rownum,colNum]!=0){
      Category1<- books[books$ISBN==(names(ratings)[colNum]),]$Category1
      Category2<- books[books$ISBN==(names(ratings)[colNum]),]$Category2
      Author<-    books[books$ISBN==(names(ratings)[colNum]),]$Author

      categories_ratings_matrix[readerName, Category1] = categories_ratings_matrix[readerName, Category
      categories_ratings_matrix[readerName, Category2] = categories_ratings_matrix[readerName, Category
      Authors_ratings_matrix[readerName, Author] = Authors_ratings_matrix[readerName, Author]+1

    }
  }
}

categories_ratings_matrix=categories_ratings_matrix/rowSums(categories_ratings_matrix)

#Fetch the preferred books for for an user for a particular category
getPreferredBooks<-function(User, Category, booksToFetch){
  #fetch all the books from the category
  booksInCategory<-books[books$Category1==Category | books$Category2==Category,c(1:3,6)]

  #Sort the books with average score
  booksInCategory<-booksInCategory[ order(booksInCategory[, "avg_score"], decreasing = TRUE), ]

  #Initialize vector for the books to recommend
  booksToRecommend<-c()
  count=as.numeric(booksToFetch)

  for(i in 1:length(booksInCategory)){
    bookISBN = booksInCategory[i, "ISBN"]

    #Add the book into the recommendation list if the user haven't read it
    if(ratings[ratings$User==User,bookISBN]==0){
      recommendBook<- paste0("Title::", booksInCategory[i,"Title"], "; Author::", booksInCategory[i,"Au
                             "; ISBN::", booksInCategory[i,"ISBN"])
      booksToRecommend<-c(booksToRecommend, recommendBook)

      #Decrease the count by 1 as one book is recommended
      count=(count-1)
    }

    #If no more boo to fetch then return the list
    if(count==0){
      return(booksToRecommend)
    }
  }
  return(booksToRecommend)
}

#Initialize the empty recommendation matrix
reco_matrix<- matrix(NA, nrow = length(readers), ncol = 5 )
rownames(reco_matrix)<-as.character(readers)
```

```
#Compute recommendation for all the users

for(i in 1:length(readers)){
  #Get users category preference
  temp_df <- cbind(Category=colnames(categories_ratings_matrix), Score=categories_ratings_matrix[i,])
  temp_df<-temp_df[ order(temp_df[, "Score"], decreasing = TRUE), ]

  #Initialize number of books to be recommended
  booksToFetch=5
  count=booksToFetch

  #Start with the  ost preferred category
  Category_index=1
  selected.category=temp_df[Category_index,1]
  booksToRecommend<-c()
  while(booksToFetch>0){

     booksToRecommend<-c(booksToRecommend, getPreferredBooks(as.character(readers[i]), as.character(sel
     booksToFetch=booksToFetch-length(booksToRecommend)

     #if the preferred category doesnt have enough book to offer to the reader go for the next category
     Category_index=Category_index+1

     # If the number of recommended books fetched then stop
     if(Category_index>length(categories)){
       break
     }
    selected.category=temp_df[Category_index,1]

  }

  booksToRecommend=c(booksToRecommend,rep("",(count-length(booksToRecommend))))
  #Add the recommended book in the matrix
  reco_matrix[i,]=rbind(booksToRecommend)

}

getRecommendation<-function(User=NA){
  if(is.na(User)){
    return(reco_matrix)
  }
  else{
    return(reco_matrix[User,])
  }
}
```

Recommending five books for each user

```
head(getRecommendation())
```

```
##          [,1]
```

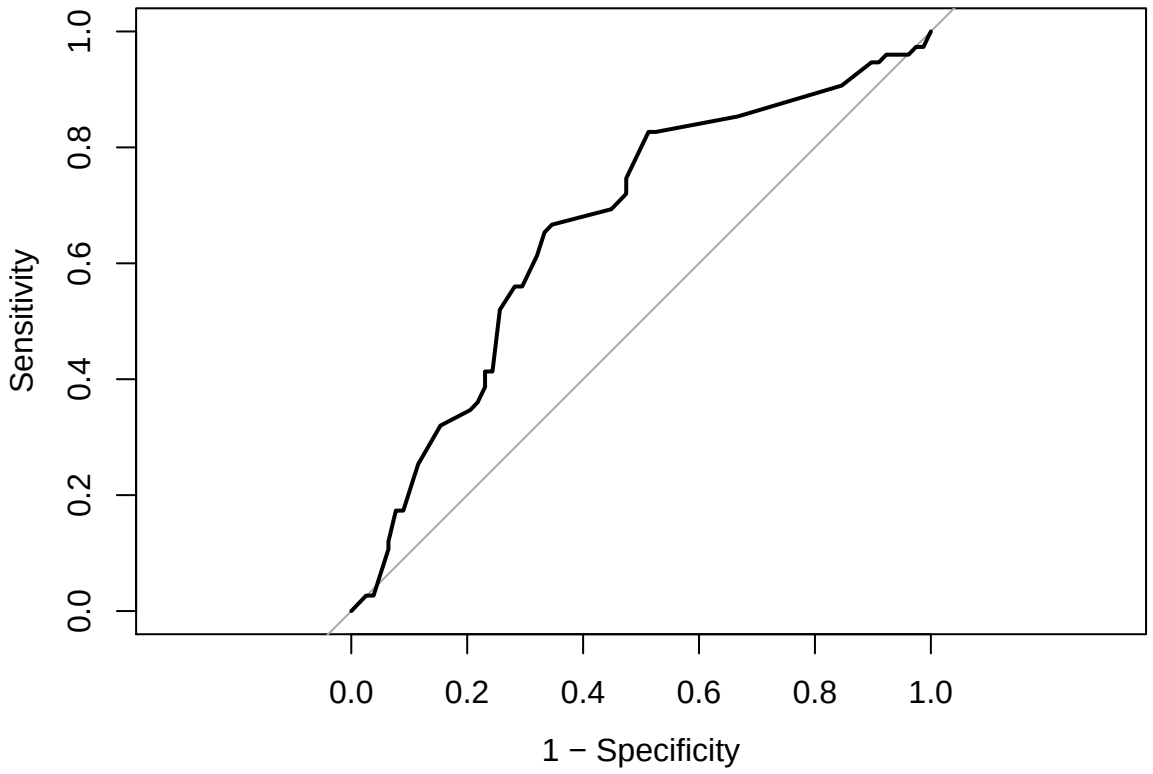
```
## Ben      "Title::Holes; Author::Louis Sachar; ISBN::978-0440414803"
## Moose    "Title::Holes; Author::Louis Sachar; ISBN::978-0440414803"
## Reuven   "Title::Holes; Author::Louis Sachar; ISBN::978-0440414803"
## Cust1    "Title::Holes; Author::Louis Sachar; ISBN::978-0440414803"
## Cust2    "Title::Holes; Author::Louis Sachar; ISBN::978-0440414803"
## Francois "Title::Holes; Author::Louis Sachar; ISBN::978-0440414803"
##          [,2]
## Ben      "Title::Bleach (graphic novel); Author::Tite Kubo; ISBN::978-1421539928"
## Moose    "Title::Hatchet; Author::Gary Paulsen; ISBN::978-1416936473"
## Reuven   "Title::Naruto; Author::Masashi Kishimoto; ISBN::978-1421584935"
## Cust1    "Title::To Kill a Mockingbird; Author::Harper Lee; ISBN::978-0446310789"
## Cust2    "Title::To Kill a Mockingbird; Author::Harper Lee; ISBN::978-0446310789"
## Francois "Title::To Kill a Mockingbird; Author::Harper Lee; ISBN::978-0446310789"
##          [,3]
## Ben      "Title::Bone Series; Author::Jeff Smith; ISBN::978-0439706407"
## Moose    "Title::Naruto; Author::Masashi Kishimoto; ISBN::978-1421584935"
## Reuven   "Title::Bleach (graphic novel); Author::Tite Kubo; ISBN::978-1421539928"
## Cust1    "Title::The Bourne Series; Author::Robert Ludlum; ISBN::978-1780485799"
## Cust2    "Title::The Da Vinci Code; Author::Dan Brown; ISBN::978-0307474278"
## Francois "Title::The Bourne Series; Author::Robert Ludlum; ISBN::978-1780485799"
##          [,4]
## Ben      "Title::Maus: A Survivor's Tale; Author::Art Spiegelman; ISBN::978-0394747231"
## Moose    "Title::Bleach (graphic novel); Author::Tite Kubo; ISBN::978-1421539928"
## Reuven   "Title::Maus: A Survivor's Tale; Author::Art Spiegelman; ISBN::978-0394747231"
## Cust1    "Title::The Hitchhiker's Guide To The Galaxy; Author::Douglas Adams; ISBN::978-0345391803"
## Cust2    "Title::The Hitchhiker's Guide To The Galaxy; Author::Douglas Adams; ISBN::978-0345391803"
## Francois "Title::The Hitchhiker's Guide To The Galaxy; Author::Douglas Adams; ISBN::978-0345391803"
##          [,5]
## Ben      ""
## Moose    ""
## Reuven   ""
## Cust1    "Title::Hatchet; Author::Gary Paulsen; ISBN::978-1416936473"
## Cust2    "Title::The Da Vinci Code; Author::Dan Brown; ISBN::978-0307474278"
## Francois "Title::Hatchet; Author::Gary Paulsen; ISBN::978-1416936473"
```

```
collaborative_reco_matrix<-matrix(NA, nrow=86, ncol=55)

for(i in 1:55){
  collaborative_reco_matrix[,i]<-books[i,]$avg_score
}
```

ROC plot for Collaborative filtering

```
ratings.binary<- unlist(ratings[,2:56])
reco.binary<-as.numeric(collaborative_reco_matrix)
rocCurve<-roc(response=ratings.binary,predictor=reco.binary,threshhold=2)
plot(rocCurve, legacy.axes = TRUE)
```




```
##
## Call:
## roc.default(response = ratings.binary, predictor = reco.binary,     threshhold = 2)
##
## Data: reco.binary in 78 controls (ratings.binary -5) < 75 cases (ratings.binary -3).
## Area under the curve: 0.6648
```

Evaluation metrics for collaborative filtering

```
suppressPackageStartupMessages(library(recommenderlab))
```

```
## Warning: package 'recommenderlab' was built under R version 3.1.3
```

```
## Warning: package 'arules' was built under R version 3.1.3
```

```
## Warning: package 'proxy' was built under R version 3.1.3
```

```
r <- as((collaborative_reco_matrix), "realRatingMatrix")

# Create 90/10 split into training/test datasets
eval <- evaluationScheme(r[1:85,], method="split", train=0.9,
                         k=1, given=9)

# Create a UBCF recommender system using training data
r <- Recommender(getData(eval, "train"), "UBCF")
```

```
# Create predictions for test data using known ratings
pred <- predict(r, getData(eval, "known"), type="ratings")

# Compute the average metrics for all readers - RMSE, MSE, MAE
calcPredictionAccuracy(pred, getData(eval, "unknown"),given=85,goodRating=5, byuser=FALSE)
```

```
##       RMSE        MSE        MAE
## 0.19370294 0.03752083 0.15248871
```

Content Based Filtering

This recommendation system is content based and provides recommendations by normalizing each user's ratings. The algorithm is recommending items for each user that are similar to its past purchases.

```
Category_books_count<- matrix(0, nrow=length(categories), ncol=1)
rownames(Category_books_count)<- categories

for(i in 1:nrow(books)){
  Category1 = books[i,]$Category1
  Category2 = books[i,]$Category2
  Category_books_count[Category1,1] = Category_books_count[Category1,1]+1
  Category_books_count[Category2,1] = Category_books_count[Category2,1]+1
}


books_profile<- matrix(NA, nrow = length(books$ISBN), ncol = length(categories))
rownames(books_profile)<- as.character(books$ISBN)
colnames(books_profile)<- as.character(categories)

idf<- log(nrow(books)/Category_books_count)

for(i in 1:nrow(books_profile)){
  for(j in 1: ncol(books_profile)){
    category1<-as.character(books[i,"Category1"])
    category2<-as.character(books[i,"Category2"])
    idf1=0
    idf2=0
    if(as.character(categories[j])==category1){
      idf1<-idf[category1,]
    }
    if(as.character(categories[j])==category2){
      idf1<-idf[category2,]
    }
    books_profile[i,j]<- books[i,"avg_score"]*(idf1+idf2)

  }
}

avg_rating_by_user = mean(unlist(ratings[,2:56]))

user_profile<-matrix(0, nrow = nrow(ratings), ncol = nrow(books))
```

```
rownames(user_profile)<- ratings[,1]
colnames(user_profile)<- books$ISBN

user_profile<- t(as.matrix(ratings[, 2:56]- avg_rating_by_user))
books_profile = rowSums(books_profile)

rec_profile<- user_profile+books_profile
rec_profile[rec_profile>5]=5
```

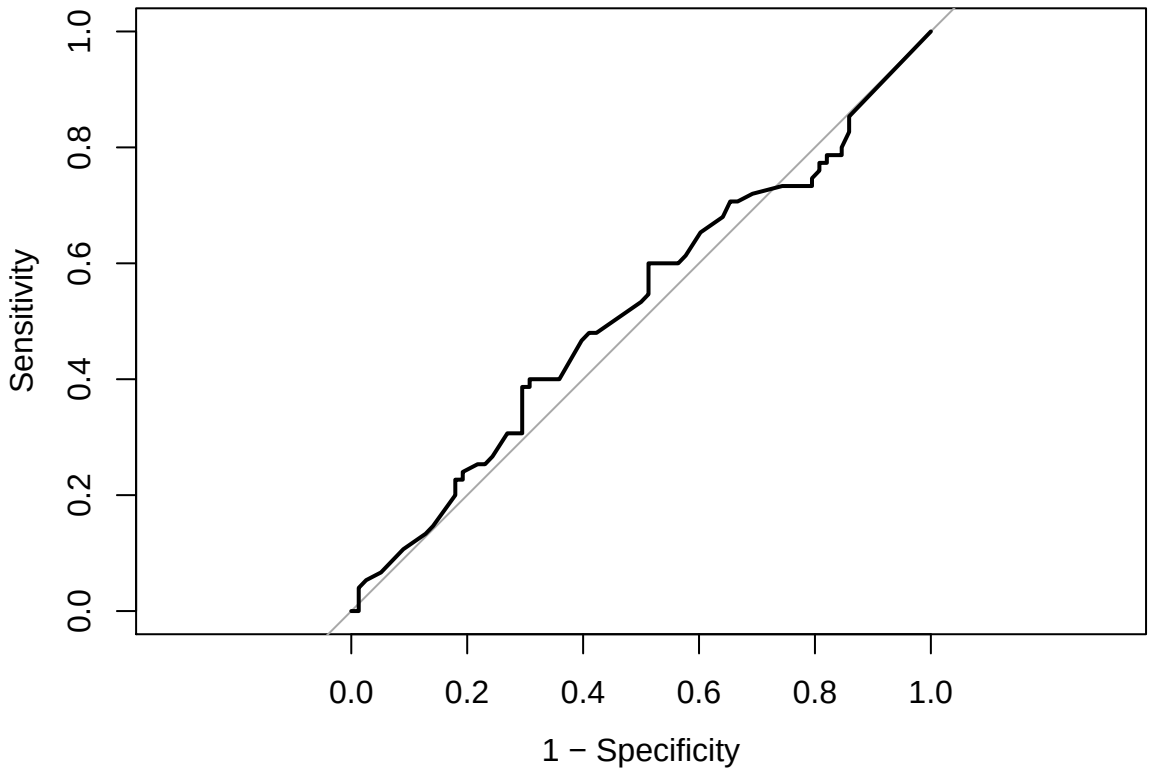
Predcition of ratings by Ben Using Contenet Base Filtering

```
rec_profile[1,]
```

```
##  [1]  5.0000000  5.0000000  5.0000000  5.0000000  5.0000000  5.0000000
##  [7]  5.0000000  4.4974062  4.4974062  4.4974062  4.4974062  5.0000000
## [13]  4.4974062  5.0000000  5.0000000  5.0000000  4.4974062  4.4974062
## [19]  5.0000000  5.0000000  4.4974062  5.0000000  5.0000000  4.4974062
## [25]  4.4974062  5.0000000  4.4974062  4.4974062  4.4974062  4.4974062
## [31]  5.0000000  4.4974062  5.0000000  4.4974062  5.0000000  5.0000000
## [37]  5.0000000  5.0000000  5.0000000  4.4974062  4.4974062  4.4974062
## [43]  5.0000000  4.4974062  4.4974062  4.4974062  4.4974062  4.4974062
## [49]  4.4974062  4.4974062  5.0000000  4.4974062  4.4974062  4.4974062
## [55]  4.4974062  5.0000000  4.4974062  4.4974062  4.4974062  4.4974062
## [61]  5.0000000  4.4974062  4.4974062  5.0000000  5.0000000  4.4974062
## [67]  4.4974062  4.4974062  4.4974062  4.4974062  4.4974062  5.0000000
## [73]  4.4974062  4.4974062  5.0000000  5.0000000  4.4974062 -0.5025938
## [79]  5.0000000  4.4974062  4.4974062  4.4974062  4.4974062  5.0000000
## [85]  4.4974062  5.0000000
```

ROC plot for content based filtering

```
ratings.binary<- unlist(ratings[,2:56])
reco.binary<-as.numeric(rec_profile)
rocCurve<-roc(response=ratings.binary,predictor=reco.binary,threshhold=2)
plot(rocCurve, legacy.axes = TRUE)
```




```
##
## Call:
## roc.default(response = ratings.binary, predictor = reco.binary,     threshhold = 2)
##
## Data: reco.binary in 78 controls (ratings.binary -5) > 75 cases (ratings.binary -3).
## Area under the curve: 0.5222
```

Evaluation metrics for content base filtering

```
r <- as(t(rec_profile), "realRatingMatrix")

# Create 90/10 split into training/test datasets
eval <- evaluationScheme(r[1:85,], method="split", train=0.9,
                         k=1, given=9)

# Create a UBCF recommender system using training data
r <- Recommender(getData(eval, "train"), "UBCF")

# Create predictions for test data using known ratings
pred <- predict(r, getData(eval, "known"), type="ratings")

# Compute the average metrics for all readers - RMSE, MSE, MAE
calcPredictionAccuracy(pred, getData(eval, "unknown"),given=85,goodRating=5, byuser=FALSE)
```

```
##      RMSE       MSE       MAE
## 1.2375988 1.5316509 0.8613343
```

Recommendation using recommenderlab package for UBCF filtering

```
# Get all book names from ratings matrix
ratings_matrix<-as.matrix(ratings[,2:56])

# Get all user names from ratings dataset and place it as individual rows of column 1
rownames(ratings_matrix)<-c(as.character(ratings[,1]))

# Get all books names from books dataset and assign them as column names
colnames(ratings_matrix)<-as.character(books$Title)

# Convert the matrix as realRatingMatrix to compress it
r <- as(ratings, "realRatingMatrix")

# Get unique values of the ratings
vector_ratings <- as.vector(r@data)
unique(vector_ratings)
```

```
## [1]  0 -5 -3  5  3  1
```

```
# Group the ratings
table_ratings <- table(vector_ratings)

# Create recommender system model

reco.model <- Recommender(r[1:nrow(r)],method="UBCF",param=list(method="Cosine",k=30))
```

```
## Available parameter (with default values):
## method    =  cosine
## nn    =  25
## sample    =  FALSE
## normalize     =  center
## minRating     =  NA
## verbose   =  FALSE
```

```
# Recommend books for all users
books.pred <- predict(reco.model,r[1:nrow(r)],n=5)

rec_matrix <- sapply(books.pred@items,function(x){

  colnames(ratings_matrix)[x]
})

head(rec_matrix)
```

```
## [[1]]
## [1] "The Hitchhiker's Guide To The Galaxy"
## [2] "The Five People You Meet in Heaven"
## [3] "Speak"
## [4] "I Know Why the Caged Bird Sings"
##
```

```
## [[2]]
## [1] "The Hitchhiker's Guide To The Galaxy"
## [2] "The Five People You Meet in Heaven"
## [3] "Speak"
## [4] "I Know Why the Caged Bird Sings"
##
## [[3]]
## character(0)
##
## [[4]]
## [1] "The Hitchhiker's Guide To The Galaxy"
## [2] "The Five People You Meet in Heaven"
## [3] "Speak"
## [4] "I Know Why the Caged Bird Sings"
##
## [[5]]
## [1] "The Hitchhiker's Guide To The Galaxy"
## [2] "The Five People You Meet in Heaven"
## [3] "Speak"
## [4] "I Know Why the Caged Bird Sings"
##
## [[6]]
## character(0)
```

EVALUATION OF UBCF RECOMMENDER SYSTEM

Metrics for User Based Collaborative Filtering System

The Root Mean Square Error, Mean Absolute Error and and Mean Square Error have been computed for the UBCF recommendaton system.

Compute RMSE, MAE, MSE

```
# Create real rating matrix

r <- as(ratings_matrix, "realRatingMatrix")

# Create 90/10 split into training/test datasets
eval <- evaluationScheme(r[1:85,], method="split", train=0.9,
                         k=1, given=9)

# Create a UBCF recommender system using training data
r <- Recommender(getData(eval, "train"), "UBCF")

# Create predictions for test data using known ratings
pred <- predict(r, getData(eval, "known"), type="ratings")

# Compute the average metrics for all readers - RMSE, MSE, MAE
calcPredictionAccuracy(pred, getData(eval, "unknown"),given=85,goodRating=5, byuser=FALSE)
```

```
##     RMSE      MSE      MAE
## 1.764983 3.115166 1.348541
```

Compute confusion matrix

```
r <- as(ratings_matrix, "realRatingMatrix")

eval <- evaluationScheme(r[1:85,], method="split", train=0.9,
                         k=1, given=9, goodRating=5)

results <- evaluate(eval,method="UBCF",n=seq(10,80,10))
```

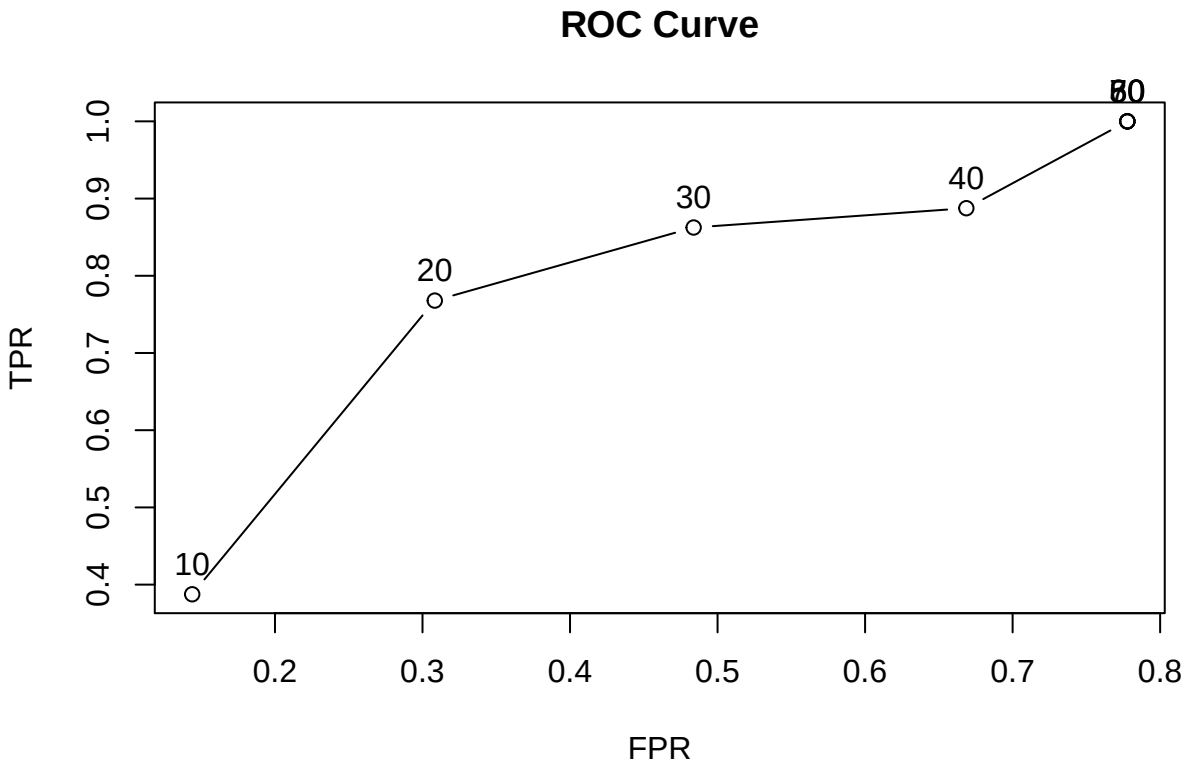
```
## UBCF run fold/sample [model time/prediction time]
##   1  [0.004sec/0.027sec]
```

```
getConfusionMatrix(results)
```

```
## [[1]]
##          TP        FP        FN       TN precision    recall       TPR
## 10 1.666667  6.111111 2.0000000 36.22222 0.2142857 0.3875000 0.3875000
## 20 2.666667 12.888889 1.0000000 29.44444 0.1714286 0.7678571 0.7678571
## 30 3.222222 20.111111 0.4444444 22.22222 0.1380952 0.8625000 0.8625000
## 40 3.444444 27.666667 0.2222222 14.66667 0.1107143 0.8875000 0.8875000
## 50 3.666667 32.111111 0.0000000 10.22222 0.1024845 1.0000000 1.0000000
## 60 3.666667 32.111111 0.0000000 10.22222 0.1024845 1.0000000 1.0000000
## 70 3.666667 32.111111 0.0000000 10.22222 0.1024845 1.0000000 1.0000000
## 80 3.666667 32.111111 0.0000000 10.22222 0.1024845 1.0000000 1.0000000
##          FPR
## 10 0.1438760
## 20 0.3083209
## 30 0.4837819
## 40 0.6686446
## 50 0.7777778
## 60 0.7777778
## 70 0.7777778
## 80 0.7777778
```

Plot ROC and Precision-Recall Curves

```
# Plot ROC Curve
plot(results,annotate=TRUE,main="ROC Curve")
```




```
# Plot Precision-Recall Curve
plot(results, "prec/rec", annotate = TRUE, main = "Precision-recall")
```