Global Baseline Predictors and RMSE

Data 612 Project 1

William Outcault

09 June 2020

Contents

Composing the Dataset **2**

Splitting the Testing/Training Sets **2**

Raw Average and RMSE **3**

Training **4**

Bias 4

Baseline Predictors 5

RMSE for Baseline Predictors 5

Testing Set **6**

Bias 6

Baseline Predictors 7

RMSE for Baseline Predictors 7

Conclusion **7**

Composing the Dataset

The following dataset is made up of 6 rows and 5 columns, the data was fabricated by myself. In context this data will represent movies rated by users A-F (names will be private), and movies will be labelled 1-5 (for sake of convenience I assigned the movies to numbers). This system will recommend movies to its users.

```
m <- matrix(
  c(5,4,4,2,4,4,NA,3,2,2,NA,2,4,5,NA,3,5,5,NA,3,NA,1,4,4,4,4,3,NA,5,4,3,2,1,NA,4,3),
  nrow = 6,
  dimnames = list(c("A","B","C","D","E","F"), c("1","2","3","4","5","6")))

knitr::kable(m)
```

	1	2	3	4	5	6
A	5	NA	4	NA	4	3
B	4	3	5	3	4	2
C	4	2	NA	NA	3	1
D	2	2	3	1	NA	NA
E	4	NA	5	4	5	4
F	4	2	5	4	4	3

Splitting the Testing/Training Sets

Next we will create our training and testing sets by random sampling. We will use 75% of the data for the training and 25% for the testing.

```
smp_size <- floor(0.75 * nrow(m))

set.seed(123)
train_ind <- sample(seq_len(nrow(m)), size = smp_size)

train <- m[train_ind, ]
test <- m[-train_ind, ]
```

	1	2	3	4	5	6
C	4	2	NA	NA	3	1
F	4	2	5	4	4	3
B	4	3	5	3	4	2
D	2	2	3	1	NA	NA

	1	2	3	4	5	6
A	5	NA	4	NA	4	3
E	4	NA	5	4	5	4

Raw Average and RMSE

Now that we have our training/testing sets we will make our first predictions. We will begin by finding the mean for all values, replacing our unknown values with the mean and make our prediction using our mean values.

```
avg <- mean(train[!is.na(train)])
rmse_train <- round(mean((train[!is.na(train)]-avg)^2),3)
rmse_test <- round(mean((test[!is.na(test)]-avg)^2),3)
```

The raw average rating for every user-item combination is 3.05. The RMSE for both our training data and testing data is 1.347 and 1.769 respectively.

Training

We begin by finding user-item bias. This is done by taking the average for each row (user) and column (item).

Bias

```
user_bias <- c()
item_bias <- c()

for (i in seq(1:nrow(train))){
  user_avg <- mean(train[i,!is.na(train[i,])])
  user_bias <- append(user_bias,user_avg-avg)
}

for (i in seq(1:ncol(train))){
  item_avg <- mean(train[!is.na(train[,i]),i])
  item_bias <- append(item_bias,item_avg-avg)
}
```

User Bias
-0.5500000
0.6166667
0.4500000
-1.0500000

Item Bias
0.4500000
-0.8000000
1.2833333
-0.3833333
0.6166667
-1.0500000

Baseline Predictors

Our baseline predictors are found by adding each user-item bias in addition to raw average.

```
baseline_train_matrix <- matrix(
  c(5,4,4,2,4,4,NA,3,2,2,NA,2,4,5,NA,3,5,5,NA,3,NA,1,4,4),
  nrow = 4,
  dimnames = list(c("C","F","B","D"), c("1","2","3","4","5","6")))

for (i in seq(1:nrow(train))){
  for (j in seq(1:ncol(train))){
    baseline_train_matrix[i,j] <- round(avg+user_bias[i]+item_bias[j],2)
    if (baseline_train_matrix[i,j] > 5){
      baseline_train_matrix[i,j] <- 5
    }
    if (baseline_train_matrix[i,j] < 0){
      baseline_train_matrix[i,j] <- 0
    }
  }
}

knitr::kable(baseline_train_matrix)
```

	1	2	3	4	5	6
C	2.95	1.70	3.78	2.12	3.12	1.45
F	4.12	2.87	4.95	3.28	4.28	2.62
B	3.95	2.70	4.78	3.12	4.12	2.45
D	2.45	1.20	3.28	1.62	2.62	0.95

RMSE for Baseline Predictors

We use our baseline predictions to find our route mean square error.

```
which_na <- train[!is.na(train)]
rmse_train <- round(mean((train[which_na]-baseline_train_matrix[which_na])^2),3)
```

The RMSE for the training set using baseline predictors is 0.184.

Testing Set

We repeat this process except we are using our test set.

Bias

```
user_bias <- c()
item_bias <- c()

for (i in seq(1:nrow(test))){
  user_avg <- mean(test[i,!is.na(test[i,])])
  user_bias <- append(user_bias,user_avg-avg)
}

for (i in seq(1:ncol(test))){
  item_avg <- mean(test[!is.na(test[,i]),i])
  item_bias <- append(item_bias,item_avg-avg)
}
```

User Bias
0.95
1.35

Item Bias
1.45
NaN
1.45
0.95
1.45
0.45

Baseline Predictors

```
baseline_test_matrix <- matrix(
  c(5,4,4,2,4,4,NA,3,2,2,NA,2),
  nrow = 2,
  dimnames = list(c("A","E"), c("1","2","3","4","5","6")))

item_bias[is.na(item_bias)] <- 0

for (i in seq(1:nrow(test))){
  for (j in seq(1:ncol(test))){
    baseline_test_matrix[i,j] <- round(avg+user_bias[i]+item_bias[j],2)
    if (baseline_test_matrix[i,j] > 5){
      baseline_test_matrix[i,j] <- 5
    }
    if (baseline_test_matrix[i,j] < 0){
      baseline_test_matrix[i,j] <- 0
    }
  }
}

knitr::kable(baseline_test_matrix)
```

	1	2	3	4	5	6
A	5	4.0	5	4.95	5	4.45
E	5	4.4	5	5.00	5	4.85

RMSE for Baseline Predictors

We use our baseline predictions to find our route mean square error.

```
which_na <- !is.na(test)
rmse_test <- round(mean((test[which_na]-baseline_test_matrix[which_na])^2),3)
```

The RMSE for the training set using baseline predictors is 0.758.

Conclusion

Our RMSE for our training and test set using baseline predictions are 0.184 and 0.758 respectively. Our test RSME is signifcantly larger than the training, intuitively I would believe this is because our test set is relatively smaller than our training set. This would lead to a higher variance therefore a chance for a much higher RMSE. If given larger datasets we would find that our baseline predictions produce similar RMSE scores for both our training and test set.