Week 11 Assignment

MD SIMON CHOWDHURY

2024-11-16

Introduction

In this analysis, we will implement the **Global Baseline Estimate** algorithm to predict movie ratings. The data is loaded from an Excel file, processed to calculate global averages, user biases, and item biases, and then used to predict missing ratings.

Steps Overview

1. Load the ratings data from an Excel file.
2. Reshape the data to a suitable format for analysis.
3. Compute the global average rating.
4. Calculate user and movie-specific biases.
5. Predict missing ratings using the formula:

$$P_{u,i} = \mu + b_u + b_i$$

Where:

- μ: Global average rating.
- b_u: User bias.
- b_i: Item (movie) bias.

Import the Ratings Data

We will read the ratings data from the movie_ratings.xlsx file.

```
# Specify url
url <- "https://raw.githubusercontent.com/simonchy/DATA607/refs/heads/main/week%2011/MovieRat
ings.xlsx"

# Create a temporary file path
temp_file <- tempfile(fileext = ".xlsx")

# Download the file to the temporary location
download.file(url, destfile = temp_file, mode = "wb")

# Read the Excel file
ratings <- read_excel(temp_file)

# Display the imported data
print(ratings)
```

```
## # A tibble: 16 × 7
##    Critic  CaptainAmerica Deadpool Frozen JungleBook PitchPerfect2 StarWarsForce
##    <chr>            <dbl>    <dbl>  <dbl>      <dbl>         <dbl>         <dbl>
##  1 Burton              NA       NA     NA          4            NA             4
##  2 Charley              4        5      4          3             2             3
##  3 Dan                 NA        5     NA         NA            NA             5
##  4 Dieudo…              5        4     NA         NA            NA             5
##  5 Matt                 4       NA      2         NA             2             5
##  6 Mauric…              4       NA      3          3             4            NA
##  7 Max                  4        4      4          2             2             4
##  8 Nathan              NA       NA     NA         NA            NA             4
##  9 Param                4        4      1         NA            NA             5
## 10 Parshu               4        3      5          5             2             3
## 11 Prasha…              5        5      5          5            NA             4
## 12 Shipra              NA       NA      4          5            NA             3
## 13 Sreeja…              5        5      5          4             4             5
## 14 Steve                4       NA     NA         NA            NA             4
## 15 Vuthy                4        5      3          3             3            NA
## 16 Xingjia             NA       NA      5          5            NA            NA
```

Reshape the Data

To make calculations easier, we'll transform the data into a long format.

```
# Reshape the data into a long format
ratings_long <- ratings %>%
  pivot_longer(-Critic, names_to = "Movie", values_to = "Rating") %>%
  drop_na()

# Display the reshaped data
head(ratings_long)
```

```
## # A tibble: 6 × 3
##   Critic  Movie          Rating
##   <chr>   <chr>           <dbl>
## 1 Burton  JungleBook          4
## 2 Burton  StarWarsForce       4
## 3 Charley CaptainAmerica      4
## 4 Charley Deadpool            5
## 5 Charley Frozen              4
## 6 Charley JungleBook          3
```

Calculate Global Baseline

Global Average (μ)

First, we calculate the global average rating across all critics and movies.

```
global_mean <- mean(ratings_long$Rating)
cat("Global Average (μ):", global_mean, "\n")
```

```
## Global Average (μ): 3.934426
```

User Bias b_u

Next, we calculate how much each user's ratings deviate from the global average.

```
user_bias <- ratings_long %>%
  group_by(Critic) %>%
  summarize(b_u = mean(Rating) - global_mean)

# Display user biases
print(user_bias)
```

```
## # A tibble: 16 × 2
##    Critic        b_u
##    <chr>       <dbl>
##  1 Burton     0.0656
##  2 Charley   -0.434
##  3 Dan        1.07
##  4 Dieudonne  0.732
##  5 Matt      -0.684
##  6 Mauricio  -0.434
##  7 Max       -0.601
##  8 Nathan     0.0656
##  9 Param     -0.434
## 10 Parshu    -0.268
## 11 Prashanth  0.866
## 12 Shipra     0.0656
## 13 Sreejaya   0.732
## 14 Steve      0.0656
## 15 Vuthy     -0.334
## 16 Xingjia    1.07
```

Item Bias (b_i)

We also calculate how much each movie's ratings deviate from the global average.

```
item_bias <- ratings_long %>%
  group_by(Movie) %>%
  summarize(b_i = mean(Rating) - global_mean)

# Display item biases
print(item_bias)
```

```
## # A tibble: 6 × 2
##   Movie              b_i
##   <chr>            <dbl>
## 1 CaptainAmerica  0.338
## 2 Deadpool        0.510
## 3 Frozen         -0.207
## 4 JungleBook     -0.0344
## 5 PitchPerfect2  -1.22
## 6 StarWarsForce   0.219
```

Predict Missing Ratings

Using the global average, user bias, and item bias, we compute predicted ratings.

```
# Merge biases with the original data
ratings_with_bias <- ratings_long %>%
  left_join(user_bias, by = "Critic") %>%
  left_join(item_bias, by = "Movie") %>%
  mutate(Prediction = global_mean + b_u + b_i)

# Display predictions
print(ratings_with_bias)
```

```
## # A tibble: 61 × 6
##    Critic  Movie          Rating     b_u     b_i Prediction
##    <chr>   <chr>           <dbl>   <dbl>   <dbl>      <dbl>
##  1 Burton  JungleBook          4  0.0656 -0.0344       3.97
##  2 Burton  StarWarsForce       4  0.0656  0.219        4.22
##  3 Charley CaptainAmerica      4 -0.434   0.338        3.84
##  4 Charley Deadpool            5 -0.434   0.510        4.01
##  5 Charley Frozen              4 -0.434  -0.207        3.29
##  6 Charley JungleBook          3 -0.434  -0.0344       3.47
##  7 Charley PitchPerfect2       2 -0.434  -1.22         2.28
##  8 Charley StarWarsForce       3 -0.434   0.219        3.72
##  9 Dan     Deadpool            5  1.07    0.510        5.51
## 10 Dan     StarWarsForce       5  1.07    0.219        5.22
## # ℹ 51 more rows
```

Save the Results

The predictions are saved into a new Excel file named results_ratings.xlsx.

```
# Save predictions to Excel
output_file <- "results_ratings.xlsx"
write.xlsx(ratings_with_bias, output_file, append= FALSE)

cat("Predictions saved to:", output_file, "\n")
```

```
## Predictions saved to: results_ratings.xlsx
```

Conclusion

This analysis implemented the Global Baseline Estimate algorithm to predict missing ratings.