Popularity based product rating system using Bayesian model

by

Ajay Khatri
B.Tech, Kurukshetra University, 2012

A Project Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF ENGINEERING

in the Department of Electrical and Computer Engineering

Popularity based product rating system using Bayesian model

by

Ajay Khatri
B.Tech, Kurukshetra University, 2012

Supervisory Committee

Dr. Fayez Gebali, Supervisor
(Department of Electrical and Computer Engineering)

Dr. Samer Moein, Departmental Member
(Department of Electrical and Computer Engineering)

Supervisory Committee

Dr. Fayez Gebali, Supervisor
(Department of Electrical and Computer Engineering)

Dr. Samer Moein, Departmental Member
(Department of Electrical and Computer Engineering)

ABSTRACT

Rating of products forms an integral part of the online shopping experience nowadays. With an increase in internet penetration, there is a wide array of E-commerce websites that allows a user to purchase and rate products. A product can be termed as popular on the basis of total purchases and number of ratings. This project involves combining the popularity and average rating of a given product to develop a new strategy for 5 star product rating system. The purpose of this new product recommendation system is to allow some amount of weight to number of total votes given to a product. The formulation of this technique is supported by the Bayesian estimation theory of statistics. According to this theory, some of the values are assumed beforehand by the website owner or developer. As the number of user reviews increases with time, the rating changes accordingly. This change in rating helps to recommend best products to a customer and thus improve conversion ratio. In addition to analyzing the rating changes, an open source movie database is used for implementation of this rating system. It was found that the top 10 movie list very closely matches with the live data on Internet Movie Database website.

List of Figures

ACKNOWLEDGEMENTS

I would like to thank:

My parents, for their continuous support, motivation, and love in difficult times. In spite of the financial constraints, they provided me with an opportunity to study in Canada.

My supervisor, Dr. Fayez Gebali, for all the support and motivation which enabled me to succeed as an international student. For all my work terms and course choices, his invaluable guidance was an immense help.

UVIC ECE Dept Admin and graduate office, Amy Rowe, Ashleigh Burns and Scott Baker for assisting me during the course of my degree.

My friend, Pratik Goswami, who helped me a lot throughout my project.

DEDICATION

To my father, **Anil Khatri** and my mother, **Neelam Khatri** for having a dream to see me succeed as a graduate student at a foreign institution. I would like to say special thanks to my brother, **Ravi Khatri,** for his motivational talks during difficult times.

To my supervisor, **Dr. Fayez Gebali,** he is such a kind person who is always willing to help and share his knowledge. I wish him the best of health.

Chapter 1

Introduction

1.1 Overview

E-commerce refers to the transactions occurring over the internet in order to buy or sell things. Online shopping is a subset of e-commerce. In the year 2012, e-commerce sales topped 1 trillion dollars for the first time in history [1]. With increasing online shopping websites, the customer base is expanding at a much faster rate. Most of the websites nowadays allow users to rate a particular product by writing reviews.

Figure 1.1: Representation of an e-commerce transaction [1]

Figure 1.1 shows various options provided to a user under the main umbrella of e-commerce. A customer can buy books, movies, clothes, songs, electronic gadgets and home appliances using a website. e-commerce website use a rating sytem that performs the calculation of overall rating. This rating system is followed by a sorting system which throws data as per the choice selected by user. It can be:

- Sort by low price
- Sort by high price
- Sort by average rating
- Sort by newest first
- Sort by oldest first

Customers have started to rely on the intelligence of other customers when buying a product online. In addition to the sorting and rating system, these websites also recommend top rated products to the customer. The role of a recommender system is to increase the number of sales and provide better sales to conversion ratio. In addition to existing metrices like Rotten Tomatoes, customers primarily prejudge a product with the help of a 5-star rating. 5-star rating technique takes an average weighted mean of a number of ratings. The issue with above 5-star rating system is it does not take into account the number of user reviews. A single user 5-star product ranks higher than a 10 user 4-star product which is unfair from a credibility point of view. In this project, the effect of user reviews on overall rating of a product is studies. Bayesian estimation theory is used to combine the popularity and rating of a product. It takes into account the number of user reviews and the individual rating given by a user.

1.2 Motivation for this work

With tremendous increase in the sale and purchase of products online, there is a huge collection of websites selling a product with exactly similar specifications or features. Amazon recently entered the market of selling grocery over the internet, which has created an entirely new trend. Rating a product using thumbs up-down, star based and positive-negative signs is a way to filter out poor quality products. Due to this reason, there is a need to study the importance of product rating in today's online

era. The work done in this project is bit different from research work done in product rating field and e-commerce market [2][3]. It combines the popularity and ratings of a given product using Bayesian statistics. In simple words, it tries to combine ratings and number of user reviews by training a system for better data output. This data is then used by a recommender system to uplift the number of sales per user.

1.3 Contributions

Normal 5-star product rating algorithms do not take into account the number of user reviews. Once the number of user reviews comes into play, the value of rating either increases or decreases accordingly. The work done in this project contributes to the following:
1.Graphical representation of overall rating variation with user reviews, prior values and live rating.
2.Analysis of mean movie rating variation on an open source dataset.
3.Implementation of Bayesian theory estimated values on given dataset to recommend top 10 movies. It also includes calculation of top 10 movies based on weighted average rating.
This project work will help website owners, business owners, and product managers to implement an efficient technique for recommending the best products to a user visiting their website. As a result, the sales conversion ratios will improve and the problem of single user rated product reaching top of the list will be sorted out. It will also solve the problem of a single user rated product reaching the top when sort by rating option is selected.

1.4 Report organization

This section outlines the organization of this project report and is intended to present the reader with the main focus summary about each chapter.

Chapter 1 provides the reader with the basic scope of this project and work done. The motivation for this project and contributions made are also included.

Chapter 2 describes the background about various type of product rating strategies. After this, it elaborates a common problem faced by recommendation system

based on 5-star rating. This problem served as the source of motivation for this project in addition to a co-op work term.

Chapter 3 explains the theory behind the product recommendation system implemented in this project. It defines the various variables in Bayesian estimation formula and how they are linked to each other. Furthermore, it talks about open source dataset and various Python libraries used in this project.

Chapter 4 contains the simulation results showing the variation of overall rating with various factors such as prior value, prior confidence, the number of user reviews and rating based on those user reviews. It also explains the implementation of Bayesian theory for an open source dataset to find the top 10 products.

Chapter 5 consists of the concluding statements and a description of the solution offered by doing this project.

Chapter 2

Product Rating Background

A rating can be considered as a tool used to support the quality of a given movie, image, product or book. They are accepted online as an indication of consumer opinion of products. In most of the cases, website rating scales allow one rating per user per product.

2.1 Types of Rating Scales

There are 3 widely adopted rating scales used by websites. It includes binary scales such as thumbs up thumbs down, plus-minus and star based rating scales. Star based systems are comparatively more complicated than other style of rating systems.
Thumbs up and down: In this format of voting for a product, a thumbs up means a good quality product or excellent results. A thumb down simply means the user or customer is not satisfied with the product. It is a form of a binary voting system, hence the difficulty lies in a way to measure the satisfaction or poor quality level of the product. A given product cannot be either excellent or extremely poor.

Plus or Minus: In this format of voting for a product, clicking plus means a relevant good quality product. Negative sign simply means the user or customer is not satisfied with the product due to quality or content. It is a form of a binary voting system, hence the difficulty lies in a way to measure the satisfaction or poor quality level of the product. Some websites use the mean of difference between overall positive and negative votes.

Star: A star voting system can be considered as an upgrade to the thumbs up and down style. Figure 2.1 represents a 5 star rating system from a top selling ecommerce website.

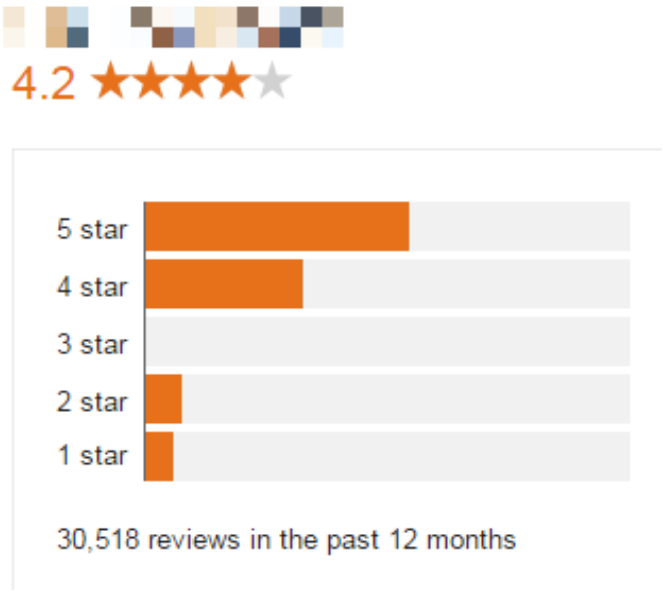

Figure 2.1: Example of 5 star based rating

The current rating of product is 4.2 based on approximately 31 thousand reviews. 1 star means a poor quality product or the product did not meet customers intended needs. 5 star means superior quality and highest level of customer satisfaction. All the ratings in between reflect the variation of how much the customers liked a particular product.

2.2 Overall Rating Calculation

For a thumbs up or down voting rating system, there are multiple methodologies used by websites for sorting the rank of products. The first method is subtracting thumbs up and thumbs down. Second technique can be dividing the total thumbs up by total combined thumbs. However, each of these strategies has their own limitations. Currently, this methodology is used by a famous dictionary website called as UrbanDictionary. For a given 5 star system, the overall rating is calculated by taking the weighted average or weighted mean. This means each star is multiplied by the

number of users who voted for it. The formula can be given as

$$r = \frac{\sum_{i=1}^{n} x_i \times w_i}{\sum_{i=1}^{n} w_i} \tag{1}$$

x represents the rating given by user w. The numerator is a sum of all the star values received, while denominator is a sum of total number of users who gave a rating. This simple equation is used by backend web developers for implementation of a rating calculation system.

2.3 Recommender System

A recommender system [4] is a filtering system whose main purpose is to filter top rated or popular products as per the user search history, purchase history, number of reviews and profile details. These type of systems have been brought into different business classes such as books, movies, restaurants, and music. They involve complex neural networks and artificial intelligence prediction algorithms. The recommender system implemented in this project uses pure data analytics and statistics unlike the one's used by major eCommerce leaders today. A product can be termed as popular on the basis of total number of reviews received and total number of purchases made. A popularity based recommender system can be used for the following purposes:

- Top 20 fiction books
- Top 100 games purchased
- Top 20 music charts for a given month
- Top 10 restaurants in a given city

2.4 Problem

For a given 5 star rating system, what to do with items that have less number of either extremely good or extremely poor ratings? An item with single excellent rating should not be allowed to sit at the top of a sorted list by rating. As a result, the product recommendation sytem based on rating becomes flawed. In Figure 2.2, we can see a list of hard disk drives sorted by user ratings. It is easy to spot that a hard drive

rated by lesser number of users is shown before one which is rated by a significant number of users.

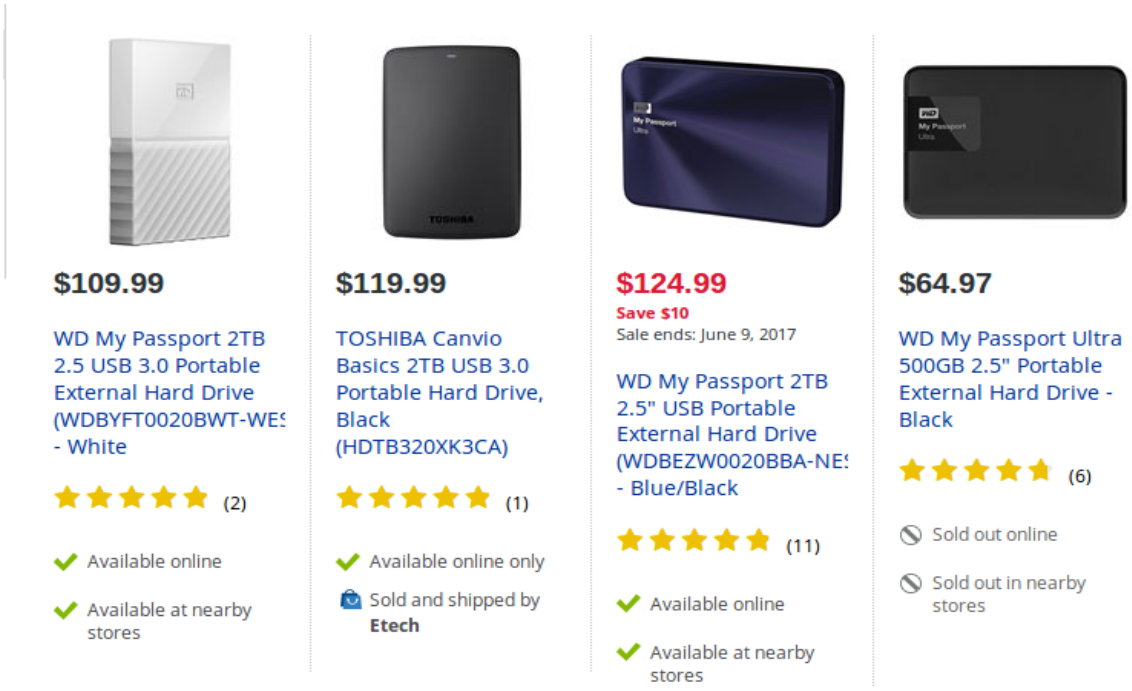

Figure 2.2: Live website with product sorting issue [4]

A product recommendation system which fetches popular products based on this system is not going to yield proper results. The products recommended might be of lower rating than expected and thus hurt the organization in long run. This can also result in lower rate of sales conversation ratio.

Chapter 3

Bayesian Theory and Movie Dataset

3.1 Bayesian theory of statistics

Bayesian theory of statistics [5] is a technique in the field of statistics that deals with the uncertainty and prior beliefs to make predictions about data. The evidence about the true state of the world is expressed in terms of a prior and confidence in prior value. Using Bayesian Statistics, we can analyze the knowledge from the incident that happened before, and predict the incident which might happen in the future. This method adopts the probability, which comes from the comparison of probability among previous and current events. Let's assume X is the probability for an individual to give thumbs up to a movie, initially declared as 0.5. Now we go to a cinema hall parking lot and ask users if they would give the movie thumbs up or thumbs down.We would use these values to update the value of X we assumed. The updated value of X after N observations is called posterior distribution. According to Baye's theorem, the probability of X given O is

$$P(X/O) = \frac{P(O/X) \times P(X)}{P(O)} \tag{2}$$

In Equation 2, $P(O)$ is the probability of observation and $P(X)$ is the prior value before seeing any distribution. $P(O/X)$ is called the likelihood probability. For Bayesian estimation of a 5-star rating system, we are not dealing with scalar values. This is the reason it has to follow a joint distribution [6].

3.2 Bayesian Average Formula

The Bayesian average formula is a simplified outcome of the Bayesian theory of statistics. The average value of a categorical random variable is the weighted average of the values of the random variable weighted by the respected probability values. In other words, the sum of the probability of getting a star, given our observations, multiplied by that star's value for each star value from 1 to 5. Variable P_i represents the probability of getting each star value which is multiplied by the value from 1 to 5. So, for the categorical variable of star ratings, the average value would be

P=$P_1 + 2P_2 + 3P_3 + 4P_4 + 5P_5$

The expected value of average rating based on the posterior is then computed by Equation 3.

$$E((P_1 + 2P_2 + 3P_3 + 4P_4 + 5P_5)/O) = \sum_{i=1}^{5} i \times E(P_i/O) \tag{3}$$

The equation for Bayesian average involves mainly two pre-assumed variables known as c and m. These values of m and c are factored into the calculation and can be applied to a reasonably sized data set. Using Dirichlet distribution, we can calculate the probability of occurence of a star as the ration of Dirichlet parameter for that star to the sum of Dirichlet parameters. This is given in Equation 4. A_i is the Dirichlet parameter used for the estimation of posterior likelihood.

$$E(P_i/O) = \frac{A_i}{\sum_{i=1}^{5} A_i} \tag{4}$$

From Equation 3 and 4, we can equate the two same terms for further simplification. This results in a simpler equation known as the Bayesian average equation shown in Equation 5.

$$X = \frac{c \times m + n \times r}{c + n} \tag{5}$$

Here c is the confidence in prior or minimum votes and m is the prior value or average rating from c votes. r is the weighted average rating and n is the total number of votes responsible for r. By assuming different values of m and c, the rate of variation

of overall product X is affected. As user ratings start to accumulate, the value of X changes dependent on the value of m and c.

3.2.1 Importance of Bayesian average formula for this project

In order to make an efficient recommendation system and also allow better sorting, the total number of reviews need to be taken into consideration. The equation for Bayesian average formula, which is derived from Bayesian theory of statistics, helps us with this scenario. In this project, Bayesian formula is used for the estimation of rating based on pre-existing beliefs denoted by variables m and c. Studies have already been done in the field of spam prediction emails using this equation, but it has never been used for analysis of product ratings involving total number of reviews.

3.3 Movie dataset

The dataset used in this project is available from the Movielens website. This website is maintained by a research group known as Grouplens, located at the University of Minnesota. Grouplens research specializes in recommender systems, digital libraries, open source movie datasets and mobile technologies. Their main ideology is to promote advancements in the practice of social computing by understanding the logic used by real people. In this project, we have used a stable open source table dataset containing 100, 000 ratings from 1000 users on a total of 1700 movies.

3.4 Programming libraries

The main language used in this project is Python3. It is a widely accepted high level programming language for general purpose programming. It is also known as Python 3000 or py3k, one of the most famous backward compatible release of Python. It has a huge collection of standard libraries which is considered one of its strongest points. As per the current Python official documentation website, it has packages to support the following:

- User interface handling
- Web frameworks
- Data analytics

- Scientific computing
- Image processing
- Databases

Pandas is an open source library written in Python, used for data analytics and manipulation of numerical values in a dataset. It was developed in the year 2008 and available under a BSD license. In this project, the role of pandas is to perform data analytics on the given dataset. Matplot lib is a BSD licensed open source library written completely in Python programming language. It was developed in the year 2003 and has been extensively used in embedding plots. It further included sub packages such as Numpy and Scipy. The main advantage of using this open source library is the free cost unlike Matlab.

Chapter 4

Analysis and Implementation

4.1 Variation of Overall Rating

The overall rating here is denoted by the variable X. The factors having an impact over it involves the prior, confidence in prior and number of user reviews. As the number of users rating a particular product vary, the graphs discussed in next section explain variation of the overall rating of a given product.

4.1.1 Weighted Average Rating $> m$

This sub-section is based on the assumption that the ratings given to a product are higher in value than the value of m used by the website owner or the product owner. We assumed $m = 3$ and $r = 5$ for this section. Both of these values will effect the overall rating depending on the variation of value of c. From Figure 4.1, it can be seen that the rate of variation of overall rating for lower values of c is sharp and experiences sudden increase with user reviews. As the number of good user reviews increases on the website, the overall rating becomes extremely close to 5. This is because there is not a lot of confidence assumed in the value of m, so overall rating flows with the live ratings value.

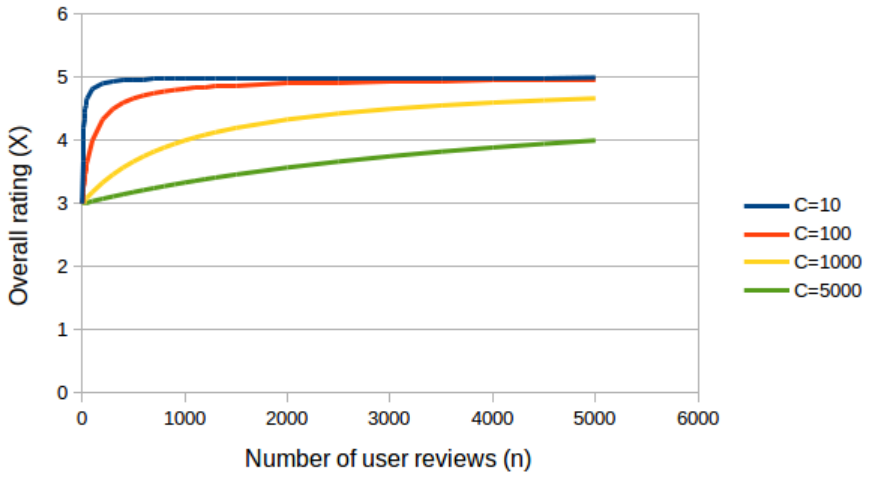

Figure 4.1: Variation of overall rating when m=3 and r=5

As the value of c is changed from 10 to 100. The rate of variation is more gradual now and doesn't experience a very sharp increase with user reviews. As the number of good user reviews increases on website, the overall rating becomes nearly close to 5.

Again, this time the value of c is varied from 100 to 1000. It produces a more gradual curve representing variation of the overall rating. As the number of good user reviews increases on the website, there is no immediate change in the overall rating. Instead, change begins to appear slowly as the user reviews reach values in multiple of thousand. The reason for the slow change is because we are forecasting a high confidence (c) in the prior value (m) of 3. The overall rating takes this into consideration and thus slowly increases on higher values of c. Finally, the value of c is increased five times more. It is observed that the curve starts to become more straight. In spite of having a lot of excellent user reviews, the maximum rating only manages to reaches 4.3.

Summarizing this sub-section, it was observed that the more confidence a website owner shows in the value of assumed rating for a product, the overall rating tries to stick closer to that value.

4.1.2 Weighted Average Rating $< m$

In this sub-section, it is assumed that the ratings given to a product are lower in value than the value of m used by the website owner or the product owner. We assumed

m =3 and r=1 for this section. Both of these values will effect the overall rating depending on the variation of value of c.

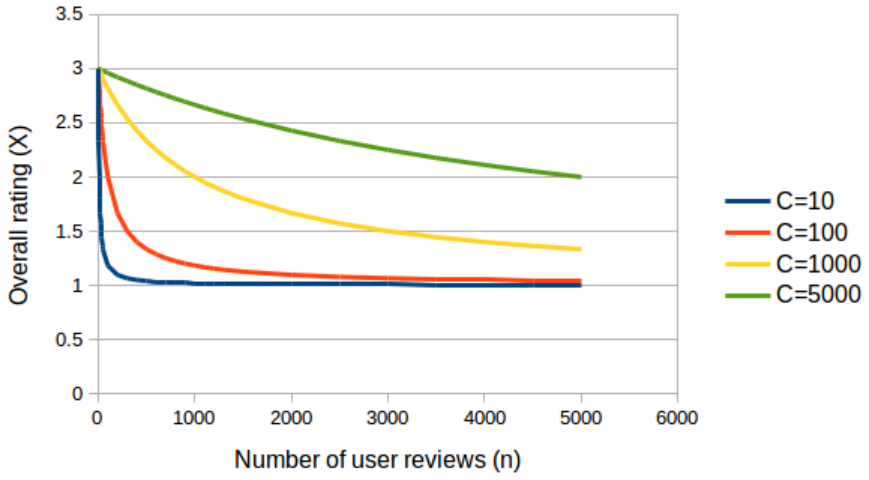

Figure 4.2: Variation of overall rating when m=3 and r=1

As the number of bad user reviews increases on the website, the overall rating becomes extremely close to 1. This is because there is not a lot of confidence assumed in the value of m, so overall rating flows with the live ratings value. As the value of c is changed from 10 to 100. The rate of variation is more gradual now and doesn't experience a very sharp decrease with user reviews. As the number of bad user reviews increase on website, the overall rating becomes nearly close to 1.

Again, this time the value of c is varied from 100 to 1000. It produces a more gradual curve representing variation of the overall rating. As the number of bad user reviews increases on the website, there is no immediate change in the overall rating. Instead, change begins to appear slowly as the user reviews reach values in multiple of thousand. The reason for the slow change is because we are forecasting a high confidence (c) in the prior value (m) of 3. The overall rating takes this into consideration and thus slowly decreases on higher values of c.

Finally, the value of c is increased five times more than previous value of 1000. It is observed that the curve starts to become more straight. In spite of having a lot of bad user reviews, the maximum rating only manages to reaches 2. Summarizing this sub-section, it was observed that the more confidence a website owner shows in the value of assumed rating for a product, the overall rating tries to stick closer to that value.

4.1.3 Weighted Average Rating = m

As long as the weighted average rating given by users stays exactly equal to the value of prior, the number of user reviews has no control over the overall rating of a given product. Change in the value of c creates no impact on the overall rating of the given product. This means the confidence value used by a website owner for the assumed rating gets nullified.

4.1.4 Confidence in Prior=n

Putting c=n in Equation 5, we get a simplified version given as

$$X = \frac{m+r}{2} \tag{6}$$

As long as the confidence in prior is equal to the number of user reviews, overall rating becomes the mean of m and weighted average. In Figure 4.3, we can see three different curves plotted against the number of user reviews which is equal to the value of variable c here. In terms of Bayesian equation mathematics, this means c=n.

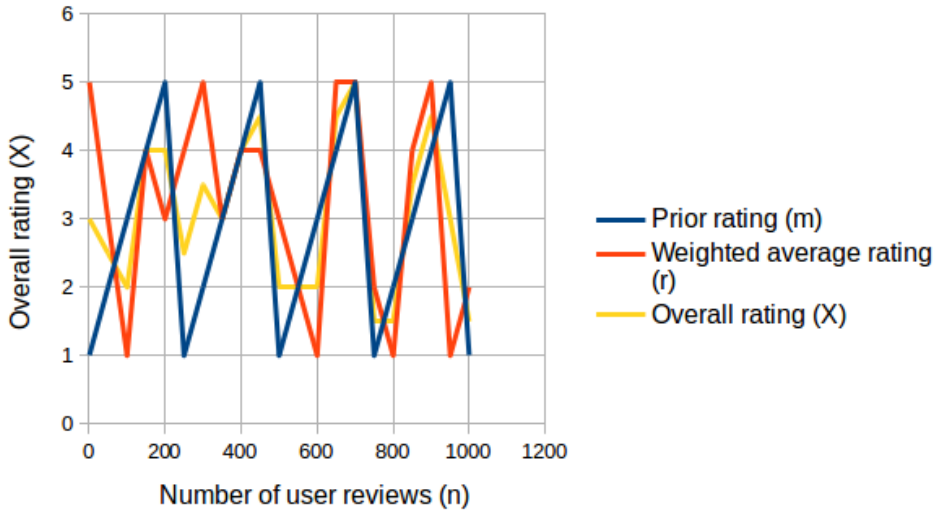

Figure 4.3: Variation of overall rating when c=number of users

4.2 Implementation on Open Source Dataset

For implementation of this Bayesian statistics rating on an open source dataset, this project used Movielens. Movielens is a website that stores movie ratings from an internationally recognized website called Internet Movie Database (http://imdb.com). The ratings are stored on a scale of 1 to 5 and the datasets can be used for academic research and other education purposes. For implementation, Python3 pandas and matplotlib library were used in this project. The file names mentioned below are of great use for us.

- u.data
- u.item

Before moving to any sort of analysis, the dataset IDs need to be joined in a single file. For this purpose, a csv file is created using pandas which contain movie id, rating and title fields. The pd merge function performs a left joint merge on the above two mentioned files.

4.2.1 Top 10 Movies as per Weighted Average

The dataset used in this section contains 100,000 ratings from 943 users on 1682 movies [7]. In order to find the top 10 movies by weighted average rating, a class is created and it gives a list of 10 movies as an output as shown in Figure 4.4.

```
======================== RESTART: /home/ajay/stars.py ========================
                                                    count  mean
title
Aiqing wansui (1994)                                    1   5.0
They Made Me a Criminal (1939)                          1   5.0
Great Day in Harlem, A (1994)                           1   5.0
Saint of Fort Washington, The (1993)                    2   5.0
Entertaining Angels: The Dorothy Day Story (1996)       1   5.0
Someone Else's America (1995)                           1   5.0
Star Kid (1997)                                         3   5.0
Santa with Muscles (1996)                               2   5.0
Prefontaine (1997)                                      3   5.0
Marlene Dietrich: Shadow and Light (1996)               1   5.0
```

Figure 4.4: Top 10 movies as per weighted average

From this list, we can easily spot the problem with weighted average based recommender system. Movies with very high weighted averages but very fewer reviewers are listed, which is incorrect. The top 2 movies Star Kid and Santa with Muscles

both have a rating of 5, but less than 5 reviewers each. The next section involves a bit in depth analysis of the weighted mean and number of reviews for individual movies given in this dataset. We are implementing the logic of Bayesian theory on given dataset to find top 10 recommended movies.

4.2.2 Distribution of Average Rating and User Reviews

For the given dataset, this section would focus on analyzing the distribution of ratings by every user. Python's matplotlib library is used for this purpose and the results are represented with the help of a false colour plot. The ratings present in the dataset are coming from real users and we are analyzing their distribution. The main role of this plot is to explain the variation of given dataset ratings as per the user reviews. The value of rating and number of user reviews are represented by X and Y axis respectively. The colour distribution scale shows the frequency of distribution of movies. A change in colour from red to green signifies an increase in number of movies. Figure 4.5 shows that there are a fair number of movies with approximately 100 reviewers and the rating is around 3.5. On left side of the plot, the rating appears to be 3 but the number of reviewers is less.

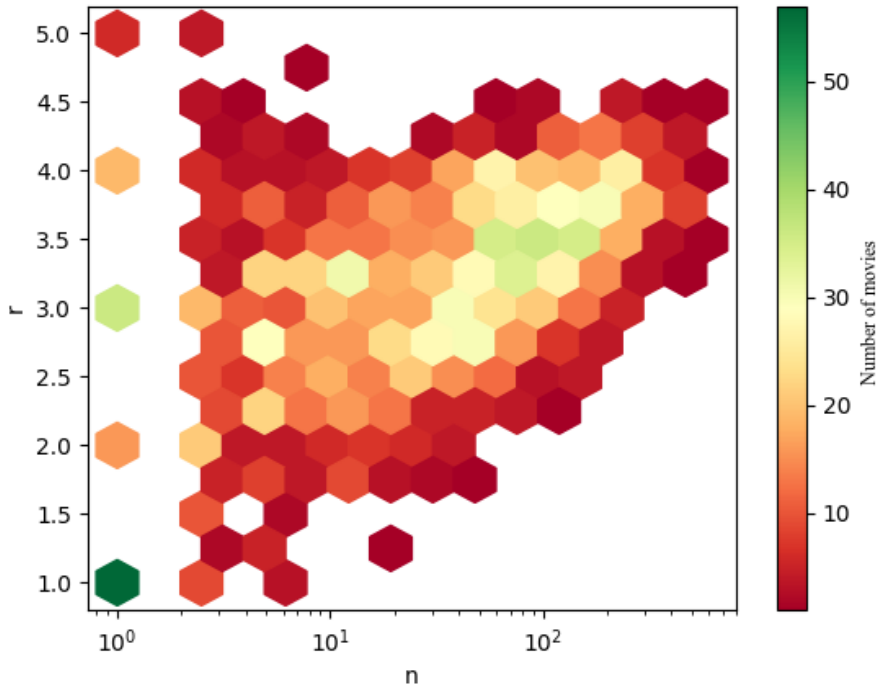

Figure 4.5: Distribution of mean rating with user reviews in the movie dataset

We can summarize the following points from Figure 4.5:

- When there are less than 10 reviewers, there is a lot of highly performing or poorly rated movies.
- As the number of reviews go higher than 10, it is highly probable for a movie to have a rating between 1.5 to 4.5.
- We can say the weighted average is good as long as there are 50 to 110 reviewers. Also the weighted average floats between 3 and 3.5.

Concluding, we used the above graph to find a suitable value of m and c. This is because the ratings are available to us already in the form of a datset, and we want to make the best possible assumption so as to match the data to a greater extent in next section. We will use these two values in order to recommend a list of top 10 movies.

4.2.3 Top 10 Movies m=3.5 and c=50

In section 4.2.2, we did an approximate estimation of the values of m and c, which will be used in this section. m is the average we can expect to receive for a movie and c is the confidence for that prior value. For our dataset, we closely approximated these values to be 3.5 and 50 respectively. These two values were included in the python class created for determination of top 10 movies. It gives us the list shown in Figure 4.6.

```
====================== RESTART: /home/ajay/stars2.py ======================
                                          bayes  count      mean
title
One Flew Over the Cuckoo's Nest (1975)  4.125796    264  4.291667
Raiders of the Lost Ark (1981)          4.145745    420  4.252381
Rear Window (1954)                      4.167954    209  4.387560
Silence of the Lambs, The (1991)        4.171591    390  4.289744
Godfather, The (1972)                   4.171706    413  4.283293
Usual Suspects, The (1995)              4.206625    267  4.385768
Casablanca (1942)                       4.250853    243  4.456790
Shawshank Redemption, The (1994)        4.265766    283  4.445230
Star Wars (1977)                        4.270932    583  4.358491
Schindler's List (1993)                 4.291667    298  4.466443
```

Figure 4.6: Top 10 movies as per Bayesian statistical average

4.2.4 Top 10 Movies m=3 and c=100

Let's consider the case when value of both m and c are changed. Considering the value of m as 3 and projecting a higher confidence value of 100, we get the list shown

in Figure 4.7. We can notice that this time there is higher difference between the bayesian rating and the mean rating, this is because we projected a higher confidence in value of 3.

```
====================== RESTART: /home/ajay/stars2.py ===========
                                    bayes  count      mean
title
Fargo (1996)                     3.965461    508  4.155512
Titanic (1997)                   3.968889    350  4.245714
Usual Suspects, The (1995)       4.008174    267  4.385768
Raiders of the Lost Ark (1981)   4.011538    420  4.252381
Silence of the Lambs, The (1991) 4.026531    390  4.289744
Casablanca (1942)                4.032070    243  4.456790
Godfather, The (1972)            4.033138    413  4.283293
Shawshank Redemption, The (1994) 4.067885    283  4.445230
Schindler's List (1993)          4.097990    298  4.466443
Star Wars (1977)                 4.159590    583  4.358491
```

Figure 4.7: Top 10 movies when higher value of c is used

This logical implementation can form the basis of a recommender system using Bayesian statistics. In simple words, this recommender system would use the ratings and number of reviews to make a top product list related to movies, books and music.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

In this project, impact of the number of user reviews on the overall rating of a given product was studied. This was done by projecting an average rating as prior and using a confidence variable for that prior value. Using this knowledge as the base model, a top 10 list of recommended movies was calculated using Python. It is really fascinating to see the importance of the number of user reviews affecting the overall rating of a given movie. This project can be positively used as the setting stone for implementing a popularity based recommendation system. It will take into account the number of reviews and the ratings provided by those reviews. This way a versatile system can be implemented into live websites.

5.2 Future Work

Cooling factor for considering newer votes more important than really old votes is another major component that can be improved as a future work for this project. This would allow the ratings to be able to follow the latest user trend. Some websites tackle this by showing the number of user reviews but that doesn't impact the rating in any way as it is a totally separate entity. As the process itself is a bit computation intensive on the backend of a website, techniques such as median average could also be considered as an expansion of this project work.

Bibliography

[1] Ryan Ecosta. E-commerce sales topped 1 trillion for first time in 2012. `https://www.emarketer.com/Article/1009649`, June 2014. Online.

[2] Mejo Antony and Nivya Johny. Product rating using sentiment analysis. In *International conference on electrical, electronics and optimization techniques 2016*, pages 2–3. IEEE, November 2016.

[3] Vrushali Karkare and Sunil R Gupta. Product evaluation using mining and rating opinions of product features. In *International conference on electronic systems and computing technologies 2014*, pages 4–5. IEEE, February 2014.

[4] Ulles Inman. *Recommender systems*, chapter 3, pages 112–114. Stanford University Press, Oct 2016.

[5] Jennifer Watkins. *Bayesian theory of statistics*, chapter 2, pages 33–34. Arizona State University Press, January 2015.

[6] Kejian Wang and Guifa Teng. Research on spam e-mails recognition based on bayesian formula and decision tree. In *International conference on neural networks and brain*, pages 2–3. IEEE, October 2005.

[7] F. Maxwell Harper and Joseph A. Konstan. *The MovieLens Datasets*, chapter 2, pages 11–12. University of Minnesota press, December 2015.

[8] Internet movie database. List of top 250 movies by ratings. `https://www.imdb.com/chart/top`, June 2017. Online.