EM Mixture Model

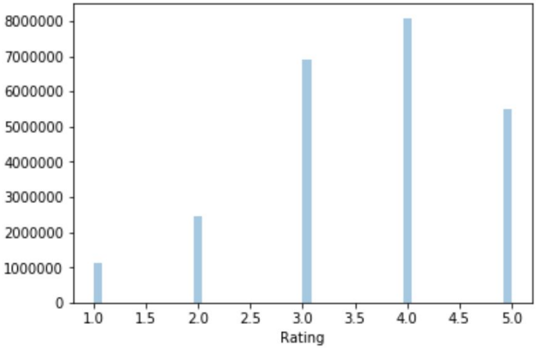
Xin Jin
Xuan Guo

Data Exploration

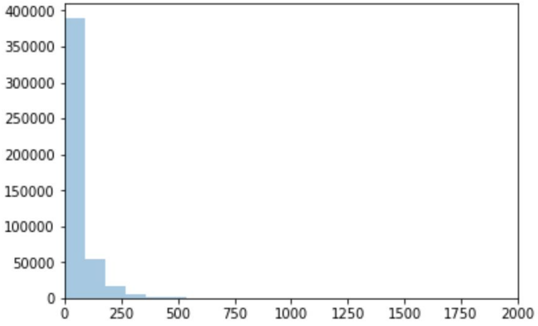
- Original data has **100498277** rows and **3** columns(Time, Customer Id and Ratings);
- We focus on the first data set due to computation limit;
- The dataset includes **4499** movies and **470758** customers.
- In the form of matrix, there exists more than **98%** missing values.

Data Exploration

The Distribution of Ratings

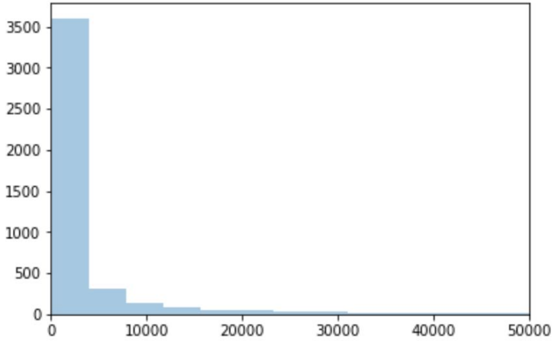


The Distribution of Reviews for Movies



Right-skewed

The Distribution of Reviews for Users



Right-skewed

Data Cleaning

Original dataset

1:		
1488844	3	2005-09-06
822109	5	2005-05-13
885013	4	2005-10-19

Cleaned dataset

Cust_Id	Movie_Id	Rating
1488844	1	3
822109	1	5
885013	1	4

Matrix Form

Cust_Id\ Movie_Id	1	2
1488844	3	.
822109	5	.
885013	4	.

Model Assumptions

Quirky(pi):
In quirky mode, rater i has a private rating distribution with probability mass function q(x|αi) that applies to every movie regardless of its intrinsic merit.
Consensus(1-pi):
In consensus mode, rater i rates movie j according to a distribution with probability mass function c(x|βj) shared with all other raters in consensus mode.

$$q(k|\alpha_i) = \binom{d-1}{k-1} * \alpha_i^{k-1} * (1-\alpha_i)^{d-k}$$

$$L(\theta) = \prod_i \prod_{j \in M_i} [\pi_i q(x_{ij}|\alpha_i) + (1-\pi_i) c(x_{ij}|\beta_j)],$$

$$c(k|\beta_j) = \binom{d-1}{k-1} * \beta_j^{k-1} * (1-\beta_j)^{d-k}$$

EM Algo Implementation for (pi, alpha, beta)

$$\ln\left(\sum_{i=1}^{m}\gamma_i\right) \geq \sum_{i=1}^{m}\frac{\gamma_i^n}{\sum_{j=1}^{m}\gamma_j^n}\ln\left(\frac{\sum_{j=1}^{m}\gamma_j^n}{\gamma_i^n}\gamma_i\right).$$

$$\ln L(\theta) \geq \sum_i\left[\ln\pi_i\sum_{j\in M_i}w_{ij}^n + \ln(1-\pi_i)\sum_{j\in M_i}(1-w_{ij}^n)\right]$$
$$+\sum_i\sum_{j\in M_i}w_{ij}^n\ln q(x_{ij}|\alpha_i) + \sum_i\sum_{j\in M_i}(1-w_{ij}^n)\ln c(x_{ij}|\beta_j) + \sum_i\sum_{j\in M_i}c_{ij}^n.$$

Convergence:

$$\frac{\left|\ln L(\theta^n) - \ln L(\theta^{n-1})\right|}{\left|\ln L(\theta^{n-1})\right| + 1} < \varepsilon$$

Updates:

$$\pi_i^{n+1} = \frac{\sum_{j_{x_{ij}>0}} w_{ij}^n}{m_i}$$

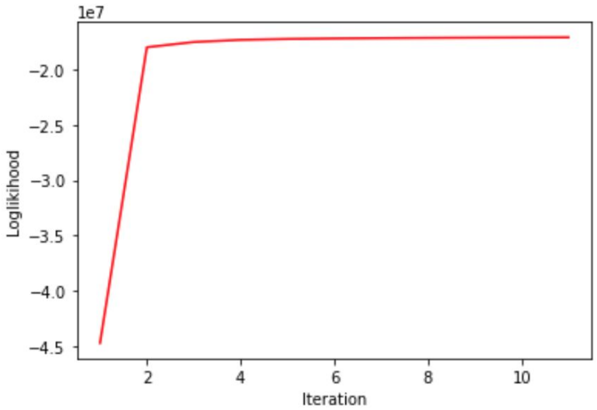
$$\alpha_i^{n+1} = \frac{\sum_{j_{x_{ij}>0}} w_{ij}^n * (x_{ij}-1)}{(d-1) * \sum_{j_{x_{ij}>0}} w_{ij}^n}$$

$$\beta_j^{n+1} = \frac{\sum_i (1-w_{ij}^n) * (x_{ij}-1)}{(d-1) * \sum_i (1-w_{ij}^n)}$$

Reference: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2929029/

Implementation Results

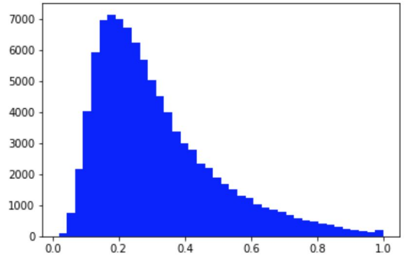
EM Convergence:



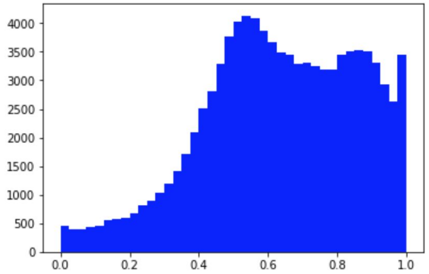

```
c = 0
while c<20:
    w = get_w(pi,alpha,beta,x,w)
    pi = get_pi(pi,w)
    alpha = get_alpha(x,w)
    beta = get_beta(x,w)
    like = loglikeli(pi,alpha,beta,x)
    if ((l[-1]-like)/l[-1])<0.0005:
        break

    l.append(like)
    c += 1
```

Final Parameters

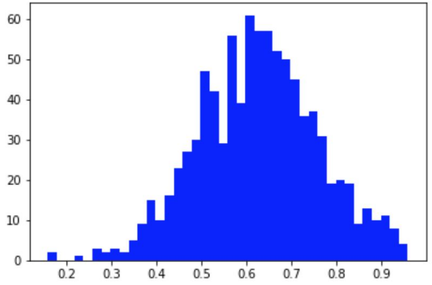
Final Distribution of Pi



Final Distribution of Alpha

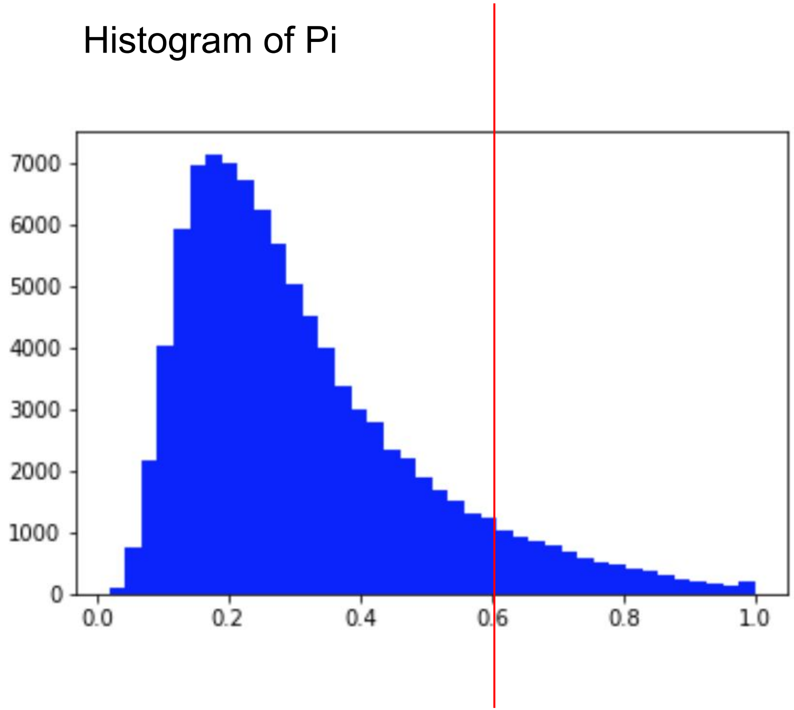


Final Distribution of Beta



Identify Unusual Users

Histogram of Pi

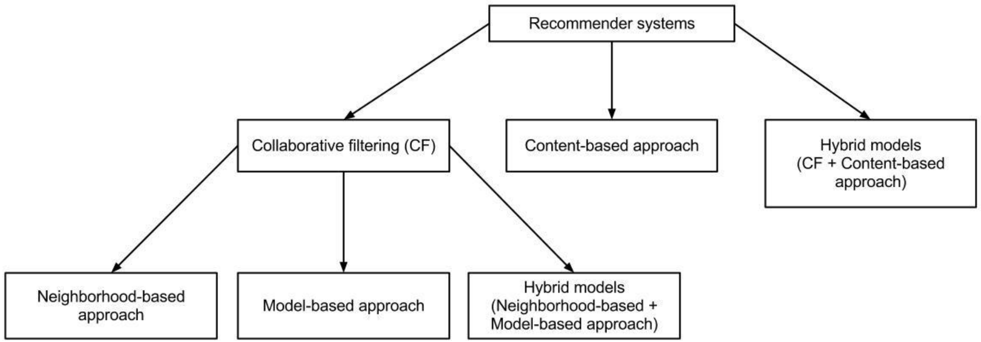


Remove users who have more probability to be in the quirky mode.

The threshold we choose to identify unusual users is pi > 0.6.

Remove 7266 unusual users.

Collaborative Filtering



User-based

Item-based

$$L(\theta) = \prod_i \prod_{j \in M_i} [\pi_i q(x_{ij}|\alpha_i) + (1 - \pi_i) c(x_{ij}|\beta_j)],$$

EM Mixture Model learns both user-based info and iterm-based info.

Collaborative Filtering

Cust_Id\ Movie_Id	1	2	3	4
1488844	3	1	2	3
822109	5	3	2	.
885013	4	.	.	3

Similarity

Pearson-Correlation Similarity

$$\text{simil}(x, y) = \frac{\sum_{i \in I_{xy}} (r_{x,i} - \bar{r_x})(r_{y,i} - \bar{r_y})}{\sqrt{\sum_{i \in I_{xy}} (r_{x,i} - \bar{r_x})^2} \sqrt{\sum_{i \in I_{xy}} (r_{y,i} - \bar{r_y})^2}}$$

Cosine-Based Similarity

$$\text{simil}(x, y) = \cos(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{||\vec{x}|| \times ||\vec{y}||} = \frac{\sum_{i \in I_{xy}} r_{x,i} r_{y,i}}{\sqrt{\sum_{i \in I_x} r_{x,i}^2} \sqrt{\sum_{i \in I_y} r_{y,i}^2}}$$

SVD Algo

The prediction $\hat{r}_{ui}$ is set as:

$$\hat{r}_{ui} = \mu + b_u + b_i + q_i^T p_u$$

Optimization Goal: Min $\sum_{r_{ui} \in R_{train}} \left(r_{ui} - \hat{r}_{ui}\right)^2 + \lambda \left(b_i^2 + b_u^2 + ||q_i||^2 + ||p_u||^2\right)$

Gradient Descent

$$\begin{aligned}
b_u &\leftarrow b_u + \gamma(e_{ui} - \lambda b_u) \\
b_i &\leftarrow b_i + \gamma(e_{ui} - \lambda b_i) \\
p_u &\leftarrow p_u + \gamma(e_{ui} \cdot q_i - \lambda p_u) \\
q_i &\leftarrow q_i + \gamma(e_{ui} \cdot p_u - \lambda q_i)
\end{aligned}$$

$e_{ui} = r_{ui} - \hat{r}_{ui}$

Comparison after Unusual User Identification

MAE	RMSE
0.73327766	0.93505776
0.73297882	0.93468037
0.73231072	0.93444277
0.73095701	0.93125733
0.73001626	0.92982399

MAE	RMSE
0.72625528	0.92592605
0.7286607	0.92794139
0.72662997	0.9263454
0.72764525	0.92663783
0.72870872	0.9272608

	MAE	RMSE
SVD(Before)	0.732	0.933
SVD(After)	0.727	0.926

Future Work

- Normalize Ratings
- Predict using EM mixture Model
- Use the time column
- Improve the data structure to increase computation speed

Thank you!