CSE 258 Assignment 1 A53280596 Yawen Zhao

Task 1

Here we use Logistic Regression to solve this problem. We get the best result when C = 1000

1.1 Feature design

Given the pair (user, book), here we design 6 features of it as following:

(1) Computer the maximum Jaccard, Cosine and Pearson similarity of the book with all the books this user has read before. Denote them as $bookJacSim, bookCosSim, bookPearSim$.

(2) Computer the maximum Jaccard, Cosine and Pearson similarity of the user with all the users who have read this book. Denote them as $userJacSim, userCosSim, userPearSim$.

Here the 3 kinds of similarity can be expressed as below:

$$Jaccard(u,v) = \frac{u \cap v}{u \cup v}$$

$$Cosine\ Similarity: Sim(u,v) = \frac{\sum_{i \in I_u \cup I_v} R_{u,i} R_{v,i}}{\sqrt{\sum_{i \in I_u \cap I_v} {R_{u,i}}^2 \sum_{i \in I_u \cap I_v} {R_{v,i}}^2}}$$

$$Pearson\ Similarity: Sim(u,v) = \frac{\sum_{i \in I_u \cup I_v} (R_{u,i} - \overline{R_u})(R_{v,i} - \overline{R_v})}{\sqrt{\sum_{i \in I_u \cap I_v} (R_{u,i} - \overline{R_u})^2 \sum_{i \in I_u \cap I_v} (R_{v,i} - \overline{R_v})^2}}$$

Note that we computer Jaccard Similarity, we not only consider each element as the book the user read, or the user who read the book, but we also consider the rating the user gave to the book.

Then our feature to be train can be denoted as:

$x = [1, bookJacSim, userJacSim, bookCosSim, userCosSim, bookPearSim, userPearSim]$

Here corresponding y is just whether this user has read this book, 1 as read, 0 as not read.

1.2 Data Seperation

Given 200000 (user, book) pairs of the books the users have read, we separate the data as 2 parts:

(1) 190000 (user, pair, rating) used to generate the feature of each book and each user to computer their similarity.

(2) 20000 (user, pair) used to train the logistic regression classifier.

Regarding the 20000 data used to train the logistic regression classifier, as we only have 10000 (user, book) pairs that indicate the user has read this book as possitive data, we need to generate 10000 (user, book) pairs that indicate the user has not read this book as negative data. Here we simply generate these 10000 pairs based on the 200000 pairs concerning about all the books each user has read.

Task 2

Here we predict the rating as: $f(u,b) = \alpha + \beta_u + \beta_b$. The problem then becomes optimize the following equation:

$$\arg min_{\alpha,\beta} \sum_{u,b} \left(\alpha + \beta_u + \beta_b - R_{u,b}\right)^2 + \lambda \left(\sum_u {\beta_u}^2 + \sum_b {\beta_b}^2\right)$$

We can computer the derivatives for the 3 parameter and get the following equation. We can repeat the following updates until convergence:

$$\alpha = \frac{\sum_{u,i \in \text{train}} (R_{u,i} - (\beta_u + \beta_i))}{N_{\text{train}}}$$
$$\beta_u = \frac{\sum_{i \in I_u} R_{u,i} - (\alpha + \beta_i)}{\lambda + |I_u|}$$
$$\beta_i = \frac{\sum_{u \in U_i} R_{u,i} - (\alpha + \beta_u)}{\lambda + |U_i|}$$

By changing λ, we can get the best MSE when $\lambda = 3$. The we just use $f(u,b) = \alpha + \beta_u + \beta_b$ to predict the rating of user u give to book b.