Assignment1

1 Task 1 Read Prediction

1.1 Methodology

In general, I solve this problem by using a logistic regression model to predict whether a person has read a book or not. Apart from this, I also do a special “trick” to utilize the attribute of the “balance” test set.

1.2 Feature Design

Suppose each piece of datum have the form (u, b), in which u represents the user, b represents the book. Let Ru represents the set of books u has read, Rb represents the set of users who have read b.

Features:

1. |Ru| (Number of books that u has read)
2. |Rb| (Number of users that have read book b)
3. The average rating of book b.
4. For all the books (except b itself) in Ru, selects the one that has the highest Jaccard similarity with b.
5. The average Jaccard similarity of all the books in Ru (except b itself) with b.

1.3 Special Trick

Since the test set is balance (all the users have exactly half of the positive labels and half of negative labels), I first get the probability of how likely the particular label is positive from logistic classifier. Then for each user in test set, I set the label of first half of the books with most confidence to be read by the user to be positive, and set the rest of which to be negative. This increases the performance a lot.

1.4 Future Work

What I do in this task is mainly Item-Based Collaborative Filtering. May be try to compute the similarity between users will also be helpful (User-Based Collaborative Filtering).

2 Task 2 Category Prediction

2.1 Methodology

The approach I use is basically combing tf-idf, unigram and bigrams together to build a Bag-of-words model that counts the words occur in each review text and use logistic regression to compute the result.

2.2 Feature Design

Basically, I use all the unigram and bigram to count all the words in the text, but discards the ones that occur in less than 5 or more than 70000 different texts. Then I compute the tfidf value of each unigram/bigrams and use it as the feature matrix.

2.3 Future Work

The datum contains information more than review text. It may be helpful to put features such as rating and users’ previous reading activity (what kinds of books that user read most) into feature matrix.