Project 3
User and Movie

Brief

In this experiment, you can choose ONE of the following two tasks

Background: MovieLens (http://www.movielens.org/) is a website where users can rate movies. The website released some data regarding user information, movie information, and the ratings

Task 1: User profiling. You need to predict the gender and age of users based on their ratings on movies

Task 2: Movie annotation. You need to predict the genres of movies based on their ratings given by users

Data

There are three types of data files

users.dat: UserID::Gender::Age::Occupation::Zip-code

- Note: for task 1, users.dat is split into ten files named users.dat0, 1, …, 9. They are used for 10-fold cross validation

movies.dat: MovieID::Title::Genres

- Note: for task 2, movies.dat is split into ten files named movies.dat0, 1, …, 9. They are used for 10-fold cross validation

ratings.dat: UserID::MovieID::Rating::Timestamp

For more information, please read the movielens_1m_readme

All files can be opened with Text Editors (e.g. Notepad, Word, …)

Task 1

In this task, you need to perform 10-fold cross validation. In each fold, users.datx (x=0,…,9) is used for test, and all other users.datx are used for training. You can use movies.dat and ratings.dat in both training and test and in whatever way

The evaluation is

- To show the error rate of prediction of gender
- To show the weighted error rate of prediction of age (Remark: in this data set, user's age is divided into 7 ranges; thus, if the predicted age and the true age are in the same range, error weight is 0; if they are in adjacent ranges, error weight is 1; if they are in range 3 and range 5, error weight is 2; …; if they are in range 1 and range 7, error weight is 6; the final weighted error rate is the average of error weights of all test samples)
- To show the runtime of training and test

Evaluation results of each fold and the average results must be reported

Task 2

In this task, you need to perform 10-fold cross validation. In each fold, movies.datx (x=0,…,9) is used for test, and all other movies.datx are used for training. You can use users.dat and ratings.dat in both training and test and in whatever way

The evaluation is

- To show the precision, recall, F1-value of prediction of genres (Remark: you shall calculate the average precision and recall, and then calculate the F1-value; please refer to the next slide)
- To show the runtime of training and test

Evaluation results of each fold and the average results must be reported

Task 2: Average Precision and Recall and F1-value

In this data set, each movie has >=1 genres, thus the prediction of genres is a multi-label classification, and the calculation of average precision and recall and F1-value is as follows

For example, if there are 2 test samples, and the ground-truth genres of the test samples are {C, L} and {T, H, W}, respectively

If the predicted genres are {C} and {T, F}, then the total true positive is 2 (C and T), the total TP+FN=5 (C,L,T,H,W), the total TP+FP=3 (C,T,F); thus the average precision is 2/3 and the average recall is 2/5 and the average F1-value is 1/2

If the predicted genres are {C, L, F} and {T, W, C}, then the average precision is 4/6 and the average recall is 4/5 and the average F1-value is 8/11

…

Report

.pdf is best, .doc and .ppt are acceptable

English is encouraged

List all references

Source code is a plus