Assignment 1 Report

Name: Shihan Wang Kaggle Username: itscoconut

Task 1 Read Prediction

Question: Predict the binary variable about whether the user would read the book (0 or 1). Accuracy is measured in terms of the fraction of correct predictions.

Final Model (Jaccard Similarity):

Split the train_Interacations dataset to the training set (1st-190000th) and validation set (190001st – 200000th). Build a few useful data structures: a collection of the books per user, a collection of users per book, and the collection of the books from each pair of test dataset. Given a (user, book) pair, implement the Jaccard similarity to calculate the mean similarities between the user from test set pair and the user from the training set that read the given book from each pair. Since the test set has been constructed such that exactly 50% of the pairs correspond to read books and the other 50% do not, sort all the users of test set from high Jaccard similarity to low Jaccard similarity and assign the first half positive predication(1) while assigning the other half negative predication(0).

Accuracy: 0.73107

Other approaches attempted:
Use the similarity-based model with cosine similarity to find 2 most similar users, normalize them then do dot multiply to calculate the prediction value. Accuracy=0.55585.
Calculate the maximum/median Jaccard similarities to find the most similar users, perform worse than mean did. Accuracy < 0.60.

Task 2 Rating Prediction

Question: Predict people's star ratings. Accuracy is measured in terms of MSE.

Final Model (Latent factor-based model):
Compute the rating mean of the training set (split as task 1). Implement a simple latent factor-based model to predict using rating mean as alpha, user biases term, and item biases term. Perform gradient descent by running cost function, initial parameter values, derivative functions and the labels(rating) and regularization strength. Try different regularization parameters for user biases and item biases and use the validation set to find the smallest MSE.

Lambda: 0.00001 for user biases, 0.00007 for item biases. MSE: 1.10264.

Other approaches attempted:
Use linear regression with user_average_rating and book_average_rating to create a baseline. MSE: 1.13299.