Project Proposal

Applied Machine Learning 2020, Final Project

Student: Fateme Rahimi

Recommendation Engine

Problem Statement:

Online judges provide a platform where many users solve problems every day to improve their programming skills. The users can be beginners or experts in competitive programming. Some users might be good at solving specific category of problems (e.g. Greedy, Graph algorithms, Dynamic Programming etc.) while others may be beginners in the same. There can be patterns to everything, and the goal of the machine learning would be to identify these patterns and model user's behavior from these patterns. The goal of this challenge is to predict **range of attempts** a user will make to solve a given problem given user and problem details. Finding these patterns can help the programming committee, as it will help them to suggest relevant problems to solve and provide hints automatically on which users can get stuck.

Data Files

There are 3 training data files and 1 test set:

1. **train_submissions.csv** - Contains 3 columns ('user_id', 'problem_id', 'attempts_range'). The variable '**attempts_range**' denoted the range no. in which attempts the user made to get the solution accepted lies.
2. **user_data.csv** - This is the file containing data of users. It contains the following features :
    1. user_id - unique ID assigned to each user
    2. submission_count - total number of user submissions
    3. problem_solved - total number of accepted user submissions
    4. contribution - user contribution to the judge
    5. country - location of user
    6. follower_count - amount of users who have this user in followers
    7. last_online_time_seconds - time when user was last seen online
    8. max_rating - maximum rating of user
    9. rating - rating of user
    10. rank - can be one of 'beginner' ,'intermediate' , 'advanced', 'expert'
    11. registration_time_seconds - time when user was registered

3. **problem_data.csv** - This is the file containing data of the problems. It contains the following features :

    1. problem_id - unique ID assigned to each problem
    2. level_id - the difficulty level of the problem between 'A' to 'N'

3. points - amount of points for the problem
4. tags - problem tag(s) like greedy, graphs, DFS etc.

test_submissions.csv - This contains the remaining 66,555 submissions from total 2,21,850 submissions. Contains 1 column (ID). The ‘attempts_range’ column is to be predicted.

- **Source**: Analytics Vidhya