Peer-graded Assignment: Kaggle Competition Peer Review

Yelp Recruiting

by Dan Sergiu Romescu
May 24, 2017   ♡ like   ⚑ Flag this submission

PROMPT

Part 1: Problem Description. Give the name of the competition you selected and write a few sentences describing the competition problem as you interpreted it. You want your writeup to be self-contained so your peer-reviewer does not need to go to Kaggle to study the competition description. Clarity is more important than detail. What's the overall goal? What does the data look like? How will the results be evaluated?

RUBRIC

Is the problem description clear and comprehensible?

○ 0 pts
The submission makes no real attempt to describe the problem.

○ 2 pts
The submission makes an attempt to describe the problem, I don't really understand what is being described.

○ 5 pts

Help Center

Example: "The task is to predict whether a given passenger survived the sinking of the Titanic based on various attributes including age, location of the passenger's cabin on the ship, family members, the fare they paid, and other information. Solutions are evaluated by comparing the percentage of correct answers on a test dataset."

Track 3 community-powered metrics of review quality: Useful, Funny, Cool. Over time, a good review will accumulate lots of votes in these categories from the community. However, another extremely important quality feature is the freshness of a review. What if we didn't have to wait for the community to vote on the best reviews to know which ones are high quality?

A Yelp review receives votes from the community in three categories: Useful, Funny, and Cool. The task is to predict how many "Useful" votes a new Yelp review will receive.

The problem description is clear enough; I understand what is going on.

PROMPT

Part 2: Analysis Approach.   Write a few sentences describing how you approached the problem. What techniques did you use?

 Example: "I split the data by gender and handled each class separately.

RUBRIC

Is the approach to the problem described clearly? Do you have some idea how you might employ these techniques to solve the problem?

For the females, I trivially classified all of them as "survived." For the males, I trained a random forest as a classifier. I ignored the pclass atribute that indicated the location of the passenger's cabin because I didn't think it was relevant."

Useful reviews may be written by reviewers who have a history of being useful. For each known reviewer, I calculated the average Useful score that their reviews received. I then assigned this score to their future reviews. In the test data, 40% of the reviews were written by reviewers whose usefulness history was not known. I assigned these scores to be "0" (the most frequent Useful score, consisting of 42% of the training data).

○ **0 pts**
No real attempt was made to describe the approach to the problem.

○ **2 pts**
There's a description, but I don't fully understand what it's saying.

○ **5 pts**
The description is clear enough; I understand the approach.

PROMPT

Part 3: Initial Solution. Write a few sentences describing how you implemented your approach. What languages and libraries did you use? What challenges did you run into?

 Example: "I partitioned the data by gender manually using Excel. I used Weka to build the random forest."

RUBRIC

Is the initial solution implementation clearly described? Are the tools, languages, and libraries used reasonable?

○ **0 pts**
No significant attempt was made to answer the question.

○ **2 pts**

As coming from software engineering I used python to develop, special jupyter - y can find the notebook on my github account.

Data was provided in the form of json files. I used the pandas library in python to load the data into DataFrames and to manipulate the data.

The implementation is described, but I don't fully understand how it works.

○ 5 pts
The implementation is well-described; I understand how it works.

PROMPT

Part 4: Initial Solution Analysis. Write a few sentences assessing your approach. Did it work? What do you think the problems were?

 Example: "My approach did not work so well, achieving a score of 0.65. This is less than the sample solution. I suspect I should not have ignored the pclass attribute."

My approach achieved a score of 0.58610. This was better (i.e. lower) than the benchmark scores. The benchmark scores were 0.72327 (Global mean value benchmark) and 0.72745 (All zeros benchmark).

RUBRIC

Is the initial solution analysis comprehensible and reasonable?

○ 0 pts
No significant attempt was made to assess the approach.

○ 2 pts
There's an assessment, but I don't really understand what is being described.

○ 5 pts
The assessment is adequately described; I understand what's going on.

PROMPT

RUBRIC

Part 5: Revised Solution and Analysis. Write a few sentences describing how you improved on your solution, and whether or not it worked.

 Example: "I included the pclass attribute and ignored the ticket number attribute. My score improved to 0.68."

Reviewers that have written lots of reviews (have a high "review count") may write useful reviews. In training data, I found a correlation supporting this hypothesis. I assigned the training data into 10 bins, based on review counts of their associated reviewers, and calculated the median Useful score of each bin. In the test data, I assigned Useful scores to reviews based on the appropriate reviewer word count bin. For reviews in which this information was not known (6% of test data), I assigned the value "0".

A submission based solely on the review count of reviewers received a score of 0.60070

Combining my first and second strategies, received a score of 0.58324

Is the revised solution and analysis presented clearly?

○ 0 pts
No significant attempt was made to describe an improved solution.

○ 2 pts
There's a description, but I don't fully understand it.

○ 5 pts
The improvement is adequately described; I understand it.

OVERALL ASSIGNMENT RUBRIC

Overall evaluation-Given the description provided, do you feel you would be able to reproduce the result?

○ 0 pts
No, the description is too sketchy, incomplete, or unclear for me to be able to reproduce this result.

○ 2 pts
No, but primarily because I don't have a strong enough background in the area; there is a fair amount of jargon used.

○ 4 pts
Yes, I think I could reproduce the result with some effort.

○ 5 pts
Yes, the solution is clearly and thoroughly described -- I could follow them easily.

Evaluating the analyses of others' is a great way to pick up new insights and viewpoints. What is one helpful thing you got out of reviewing this submission? (e.g. maybe ideas on analyzing solutions, incentive to go try this competition, etc.)?

(This question will not affect the submitter's score.) You're not expected to actually try and reproduce the submitter's result, but you may choose to do so. Did you try to reproduce the result?

○ No

○ Yes, but my result was different

○ Yes, and my result was the same.

Submit Review

Comments

Comments left for the learner are visible only to that learner and the person who left the comment.

share your thoughts...