UDACITY

‹ Return to "Machine Learning Engineer Nanodegree" in the classroom

Capstone Proposal

REVIEW

CODE REVIEW

HISTORY

Meets Specifications

No real concerns here. You clearly have a solid understanding of your dataset and your approach to this problem.

About the capstone project.

- This is essentially your last project to show off what you have learned throughout this program. Make sure you check out and follow the capstone report template and we look forward to seeing what you can create!

Project Proposal

✓

Student briefly details background information of the domain from which the project is proposed. Historical information relevant to the project should be included. It should be clear how or why a problem in the domain can or should be solved. Related academic research should be appropriately cited. A discussion of the student's personal motivation for investigating a particular problem in the domain is encouraged but not required.

Very solid opening section here, as you have done a great job describing the problem and problem domain.

Glad that you have chosen a real world problem and have referenced other research on such on a problem.

✓

Student clearly describes the problem that is to be solved. The problem is well defined and has at least one relevant potential solution. Additionally, the problem is quantifiable, measurable, and replicable.

> "The CHAMPS group has provided explicit interatomic scalar couplings for 85003 molecules along with information about their 3d molecular structure. Using these to train models, one then has to make predictions for 45772 new molecules. Given that these scalar couplings can in principle take any real value, this is clearly a regression task. "

Problem statement is clearly defined here. And glad that you mention that this would be a regression problem in this section.

✓

The dataset(s) and/or input(s) to be used in the project are thoroughly described. Information such as how the dataset or input is (was) obtained, and the characteristics of the dataset or input, should be included. It should be clear how the dataset(s) or input(s) will be used in the project and whether their use is appropriate given the context of the problem.

Excellent description of your dataset here, as it is clear in what you will be working with throughout this project. Glad that you mention the size of the datasets and analyze how you might utilize the features.

✓

Student clearly describes a solution to the problem. The solution is applicable to the project domain and appropriate for the dataset(s) or input(s) given. Additionally, the solution is quantifiable, measurable, and replicable.

I think you have a good Solution Statement here, as it is clear that you have thought a lot about what your approach is to this problem.

✓

A benchmark model is provided that relates to the domain, problem statement, and intended solution. Ideally, the student's benchmark model provides context for existing methods or known information in the domain and problem given, which can then be objectively compared to the student's solution. The benchmark model is clearly defined and measurable.

Using this kaggle kernel is a fine idea for a benchmark for your problem.

Could also even run a simple linear regression to get a baseline score.

Benchmarking is the process of comparing your result to existing method or running a very simple machine learning model, just to confirm that your problem is actually 'solvable'.

✓

Student proposes at least one evaluation metric that can be used to quantify the performance of both the benchmark model and the solution model presented. The evaluation metric(s) proposed are appropriate given the context of the data, the problem statement, and the intended solution.

Mean absolute error is great to use. In your final report, justification the use of this metric.

Maybe think about how outliers are treated differently?
(https://www.quora.com/What-is-the-difference-between-squared-error-and-absolute-error)

✓

Student summarizes a theoretical workflow for approaching a solution given the problem. Discussion is made as to what strategies may be employed, what analysis of the data might be required, or which algorithms will be considered. The workflow and discussion provided align with the qualities of the project. Small visualizations, pseudocode, or diagrams are encouraged but not required.

You clearly have a good game plan in place here, very solid step by step process. I would also suggest trying to 'create' more features in the beginning as well. Feature engineering is one of the best way to improve the performance of machine learning models. Might want to also check out this post for some more feature creation ideas.

Nice ideas for potential algorithms, might want to also check out using an Xgboost or LightGBM. Here might be a couple of examples of how to implement these powerful algorithms.

- Xgboost example
- LightGBM example

✓

Proposal follows a well-organized structure and would be readily understood by its intended audience. Each section is written in a clear, concise and specific manner. Few grammatical and spelling mistakes are present. All resources used and referenced are properly cited.

DOWNLOAD PROJECT

RETURN TO PATH