UDACITY

Machine Learning Capstone Project

A part of the Machine Learning Engineer Nanodegree Program

PROJECT REVIEW

CODE REVIEW

NOTES

+ AUDIT THIS REVIEW

Requires Changes

6 SPECIFICATIONS REQUIRE CHANGES

This is a solid solution to an interesting problem, and your report is well written. You've clearly demonstrated your full understanding of the machine learning process, and analyzed your model and results in a good amount of detail. There are just a few places where adjustments are to be made, but fundamentally you're certainly on the right track. I look forward to your next submission.

Definition

> **Student provides a high-level overview of the project in layman's terms. Background information such as the problem domain, the project origin, and related data sets or input data is given.**

- Solid introduction to the problem you're solving, with a good focus on its origin
- It's also key in this section to highlight why this problem is well suited for a machine learning solution. Why should we want to use ML here?
- An important part of preparing for a project is conducting some form of literature review. This gives you a sense of what work has been done in the field, what techniques were used, and what kind of performance you might expect to see. Is there any interesting academic research that has been

published on this topic? If so, it should be cited here. If not, make a note of that. Again, the point is mainly just to demonstrate that you have gone through this step

The problem which needs to be solved is clearly defined. A strategy for solving the problem, including discussion of the expected solution, has been made.

- The input and output are well defined, which makes for a solid problem statement
- Your approach to the problem is clear, and certainly suitable

Metrics used to measure performance of a model or result are clearly defined. Metrics are justified based on the characteristics of the problem.

- This metric makes sense for the problem, and is well defined here
- Your choice has been clearly justified based on its characteristics and how they align with your goals for the model

Analysis

If a dataset is present, features and calculated statistics relevant to the problem have been reported and discussed, along with a sampling of the data. In lieu of a dataset, a thorough description of the input space or input data has been made. Abnormalities or characteristics about the data or input that need to be addressed have been identified.

- The source of the data, its size, the features and their structure are all clearly described, which makes for a solid overview of the key characteristics
- A data sample is provided as required
- Some descriptive statistics of these features should be summarized as well, or at least the interesting ones. How are these things distributed? Feature distribution is a critical characteristic that influences models greatly, so it should be given proper attention here

A visualization has been provided that summarizes or extracts a relevant characteristic or feature about the dataset or input data with thorough discussion. Visual cues are clearly defined.

- These are important data qualities and certainly things well worth visualizing; good use of this section here

- The visualizations themselves are clean and well presented, with appropriate labels and identifiers, and the right visual encoding for each data type

Algorithms and techniques used in the project are thoroughly discussed and properly justified based on the characteristics of the problem.

- The algorithms you're using are clear, and you did a good job of justifying your choices
- In this section, you should focus on describing how your algorithms *work*, in terms of how they train and predict. What's the procedure behind the scenes? What's the theory behind it? This discussion doesn't have to be mathematically rigorous, unless that suits you better; essentially, however you can best explain these concepts, this is how you should approach it
- As an example, if you were describing the SVM, you might discuss such concepts as "maximizing the margin" and the "kernel trick"
- The purpose of taking this theoretical perspective is that it helps your reader to understand how it's treating the data, and therefore gives more objective reasoning for why it may be optimal in this situation

Student clearly defines a benchmark result or threshold for comparing performances of solutions obtained.

- Comparing to well defined, objective, concrete models / results is always the best approach for obtaining a baseline for our own work, and that's what you have here. Good choice

Methodology

All preprocessing steps have been clearly documented. Abnormalities or characteristics about the data or input that needed to be addressed have been corrected. If no data preprocessing is necessary, it has been clearly justified.

- This is a solid step by step overview of the work required to prepare your data for proper training, and again it's written in a way that's clear and detailed
- It's easy to see what the structure of the data would be before and after each of these transformations

The process for which metrics, algorithms, and techniques were implemented with the given datasets or input data has been thoroughly documented. Complications that occurred during the coding process are discussed.

- The goal of this section is to make our work as reproducible as possible; for any future researchers that read your work and wish to expand on it, they'll have to start by re-implementing what you have done, and they can only do that if your explanation of your work through this report is detailed and accurate. You've certainly met that requirement with your discussion here
- One of the main ways we help with reproduciblity is by clearly documenting the challenges we faced, and how we overcame them. This helps out those that are following our work not to get stuck on the things that we got stuck on. If you can include some discussion of any coding complications you faced in this process, and how you overcame these, that will round out this section nicely. If nothing particularly difficult happened and it was straightforward, discuss why this was the case

The process of improving upon the algorithms and techniques used is clearly documented. Both the initial and final solutions are reported, along with intermediate solutions, if necessary.

- As machine learning engineers, our primary focus should certainly be on building the best feature set we can for the algorithms we use. This is really the most "human" aspect of the ML pipeline. So it's nice to see that your improvements have a particular focus on improving the information in the feature set

Results

The final model's qualities — such as parameters — are evaluated in detail. Some type of analysis is used to validate the robustness of the model's solution.

- Your final results are presented in a way that's easy to analyze and compare
- Another important aspect of this section is discussing robustness. Does your model generalize well to unseen data? Is it sensitive to small changes in the data, or to outliers? Essentially, can we trust this model, and why or why not? You should attempt to answer this question in the most objective way you can, ideally with some well formed tests

The final results are compared to the benchmark result or threshold with some type of statistical analysis. Justification is made as to whether the final model and solution is significant enough to have adequately solved the problem.

Conclusion

A visualization has been provided that emphasizes an important quality about the project with thorough discussion. Visual cues are clearly defined.

- When you have time series data, it's always a good idea to get a sense for its trends by visualizing it. This is a luxury we don't have with high dimensional data, and a part of the EDA process we want to take advantage of. Good use of this section here

Student adequately summarizes the end-to-end problem solution and discusses one or two particular aspects of the project they found interesting or difficult.

- Solid recap of the overall process
- It's great to see how much you learned from this work and how much you benefitted from it. Ultimately, the capstone is a learning experience, so your personal takeaways are the most important lasting effect

Discussion is made as to how one aspect of the implementation could be improved. Potential solutions resulting from these improvements are considered and compared/contrasted to the current solution.

- This is a fairly vague idea; it seems to me you're saying that an improvement would be to, well, improve your score. How specifically might you be able to do that? What machine learning techniques, further data processing, or other evaluation procedures might you use to accomplish this?

Quality

Project report 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 to complete the project are cited and referenced.

Code is formatted neatly with comments that effectively explain complex implementations. Output produces similar results and solutions as to those discussed in the project.

⬇ DOWNLOAD PROJECT