Machine learning engineer nanodegree capstone proposal

Gabriel Henriques

August 23, 2018

Proposal

**Domain Background**

Healthcare data these days are being broadly explored in order to build models and analysis for better increase success rates in diagnostics and to develop precision treatments and by trying to find metabolic patterns in patients, therefore, predicting it’s best course of action.

In this project I intend to use classification algorithms and clustering methods to try and predict the presence of heart disease in patients from a given sample.

Also, with my background in biochemistry and biology, I’ll assess the data with a biological feedback to try and gather relevant metabolic information from the analyzed clusters of patients to correlate with medical literature, in order to find something from the biological biomarkers that could better explain illnesses such as metabolic syndrome, oxidative stress or diabetes.

The main goal is to try and find significant biomarkers in order to increase the precision in treatments for groups with same features.

I’m personally motivated for this project because besides being an engineer, I also have a biology degree. Health and technology today are two worlds set apart by skepticism most of the time. I intend to bring the awesome tools of technology to the brilliance of the human health.

With that in mind, we can try to make a better place for people that are in need.

**Problem Statement**

In order to bring two main fields of research together, medicine and technology, I intend to evaluate different ways of classification and clustering of a given dataset (proposed here), to try and draw conclusions with biological depth and relevance, after the data is thoroughly classified and processed.

Since clustering is one of my proposed algorithm in this solution, this gives the problem a different method to partition the domain of numerical valued attributes. Clustering algorithms group similar collections of data together based on a measure of similarity.

In order to decide which clusters should be combined or split, we need to use a metric of clustering dissimilarity between the sets. This is achieved by using an appropriate metric and a linkage criterion. The one proposed in this solution is **Euclidean Distance**:



**Datasets and Inputs**

In this section, the dataset(s) and/or input(s) being considered for the project should be thoroughly described, such as how they relate to the problem and why they should be used. Information such as how the dataset or input is (was) obtained, and the characteristics of the dataset or input, should be included with relevant references and citations as necessary 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.

**Solution Statement**

In this section, clearly describe a solution to the problem. The solution should be applicable to the project domain and appropriate for the dataset(s) or input(s) given. Additionally, describe the solution thoroughly such that it is clear that the solution is quantifiable (the solution can be expressed in mathematical or logical terms) , measurable (the solution can be measured by some metric and clearly observed), and replicable (the solution can be reproduced and occurs more than once).

**Benchmark Model**

In this section, provide the details for a benchmark model or result that relates to the domain, problem statement, and intended solution. Ideally, the benchmark model or result contextualizes existing methods or known information in the domain and problem given, which could then be objectively compared to the solution. Describe how the benchmark model or result is measurable (can be measured by some metric and clearly observed) with thorough detail.

**Evaluation Metrics**

In this section, propose at least one evaluation metric that can be used to quantify the performance of both the benchmark model and the solution model. The evaluation metric(s) you propose should be appropriate given the context of the data, the problem statement, and the intended solution. Describe how the evaluation metric(s) are derived and provide an example of their mathematical representations (if applicable). Complex evaluation metrics should be clearly defined and quantifiable (can be expressed in mathematical or logical terms).

**Project Design**

In this final section, summarize a theoretical workflow for approaching a solution given the problem. Provide thorough discussion for what strategies you may consider employing, what analysis of the data might be required before being used, or which algorithms will be considered for your implementation. The workflow and discussion that you provide should align with the qualities of the previous sections. Additionally, you are encouraged to include small visualizations, pseudocode, or diagrams to aid in describing the project design, but it is not required. The discussion should clearly outline your intended workflow of the capstone project.