CS 410/510: Explorations in Data Science Project Plan

Matthew Adler, Jerry Smedley, Layton Webber

GitHub Link: https://github.com/webberlh/psu-sum21-eds-final-project

- Project Objective (1-2 sentences)

In this project, we will perform machine learning classification on a dataset of 14 features consisting of medical data for a number of patients collected at hospitals in: Cleveland, Ohio; Budapest, Hungary; Zurich, Switzerland; and the VA Long Beach, California. The classifier will attempt to distinguish between the presence or absence of heart disease in patients given their medical data.

- Project Approach (1-2 paragraphs)

The dataset consists of 4 separate databases corresponding to the 4 hospitals at which the data was collected. We will train one classifier for each dataset for a total of 4, and one classifier on the combined dataset for all 4 hospitals. Each classifier's accuracy will be measured with respect to all 5 datasets, in order to see how well each model is able to generalize.

To classify the data, we will use a Random Forest algorithm. Random forests are a class of machine learning algorithms that combine groups of decision trees in order to learn complex non-linear relationships in data[1]. These algorithms not only perform classification, but also allow for studying the relative importance of each feature in the data set which could reveal some interesting properties[1]. We will use the sklearn Python library implementation of the Random Forest algorithm to perform our classification.

- Team Structure (3-4 sentences)

Our team consists of 3 people: Matthew Adler, Jerry Smedley, and Layton Webber. We will divide up the workload of the project between the 3 of us evenly. Layton was elected as the team leader in a group meeting.

- Project Milestones (~5 milestones)
    - Study the dataset and develop a working understanding of what each feature stands for
    - Develop a pipeline for ingesting and cleaning the data into a form that is compatible with our model. In particular, produce a clean data set for each hospital and one combined dataset
    - Train all 5 models on their corresponding dataset using the Random Forest algorithm
    - Develop visualizations to show the accuracy of each model, both on its own training dataset and on the other datasets.
    - Develop final report out summarizing how well the model is able to perform classification on each of the datasets

Sources:

[1] - Breiman, L. Random Forests. *Machine Learning* **45,** 5–32 (2001). https://doi.org/10.1023/A:1010933404324