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 its best course of action.

In this project I intend to use logistic regression algorithm 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 logistic regression algorithm as a method of classification, and to use clustering of the proposed dataset, to try and draw conclusions with biological depth and relevance, after the data is thoroughly classified and processed. After the prediction, we can increase probability in defining metabolic pathways that are common amongst patients.

The dataset proposed here provides different possibilities of analysis. First I intend to predict the patients with presence of heart disease from those without (#14 attribute: number of major vessels with >50% narrowing (0,1,2,3, or 4)). After that, I will apply Hierarchical Clustering Algorithm in order to separate groups of patients with similar features, cholesterol/Age/Blood Pressure, and try to find patterns on patients previously classified with heart disease or not.

Datasets and Inputs

The dataset used in this project is the Heart Disease Dataset from the **UCI ML Repository**.
Link to download: https://archive.ics.uci.edu/ml/datasets/heart+Disease

This dataset contains data from 4 different cities but only the Cleveland data is considered. It consists mainly of 303 instances (patients), and 14 attributes. The "goal" attribute is the #14, which refers to the presence of absence of heart disease in patients. It has categorical, integer and real data, and it is multivariate.

Since this dataset consists of patients with diseases or not, and it has important features on patients metabolism, such as lab results, I will be able to apply many statistics tools and ML algorithms to draw conclusions on both perspectives, technology and biology.

- Find the set of features that yields the best accuracy score, using cross-validation.

This dataset was created by:

1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

Dataset Donor:
David W. Aha (aha '@' ics.uci.edu) (714) 856-8779

Relevant Papers:

Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J., Sandhu, S., Guppy, K., Lee, S., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. American Journal of Cardiology, 64,304--310.
[Web Link]

David W. Aha & Dennis Kibler. "Instance-based prediction of heart-disease presence with the Cleveland database."
[Web Link]

Gennari, J.H., Langley, P, & Fisher, D. (1989). Models of incremental concept formation. Artificial Intelligence, 40, 11--61.
[Web Link]

The authors of the databases have requested that any publications resulting from the use of the data include the names of the principal investigator responsible for the data collection at each institution.

Solution Statement

Logistic regression is a machine learning algorithm for classification. In this algorithm, the probabilities describing the possible outcomes of a single trial are modelled using a logistic function.

By classifying our heart disease dataset, we will be generating a model that learns the difference between patients with presence of heart disease for patients with absence, considering its attributes (blood pressure, heart rate, fasting blood sugar and etc).

Logistic regression is designed for this purpose (classification), and is most useful for understanding the influence of several independent variables on a single outcome variable. Using LR gives me the advantage of having a probability estimate for the prediction.

Clustering algorithms groups similar collections of data together based on a measure of similarity. **Hierarchical Agglomerative Clustering** is used in this project as a tool to find common patients based in a selected feature (Cholesterol x Age for example). With this approach, I give the problem a different method to partition the domain of numerical valued attributes.

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 **Manhattan Distance**:

$$d(x, y) = \sum_{k=1}^{n} |x_k - y_k|$$

→ It is the sum of absolute differences between the coordinates. It is also called as Rectilinear Distance, L1-Distance/L1-Norm, Minkowski's L1 Distance, City Block Distance, Taxi Cab Distance.

By combining **clustering** and **classification**, I intend to have a final result with desirable accuracy as a trained a model as well as to shed a light in the biological part of the problem, that is the possible conclusions to be drawn from metabolic pathways in similarity amongst patients with or without heart disease.

Benchmark Model

As benchmark model for my classification problem I intend to use Naïve Bayes classifier in comparison to my results and try to perform better with my proposed method.

As a secondary benchmark model I intend to use Kaggle's competition for this dataset and to make my model to perform better as the top 10 classified (0.48 – 0.64).

And as a third benchmark model, I intend to use this paper https://web.cs.umass.edu/publication/docs/1996/UM-CS-1996-089.pdf as a reference for benchmark for instance based learning at 75.1% accuracy on this dataset.

Evaluation Metrics

I will consider 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.

Standard evaluation of accuracy, precision and recall will be used.

Accuracy:

$$Accuracy = \frac{TP + TN}{TN + FN + TP + FP}$$

Precision:

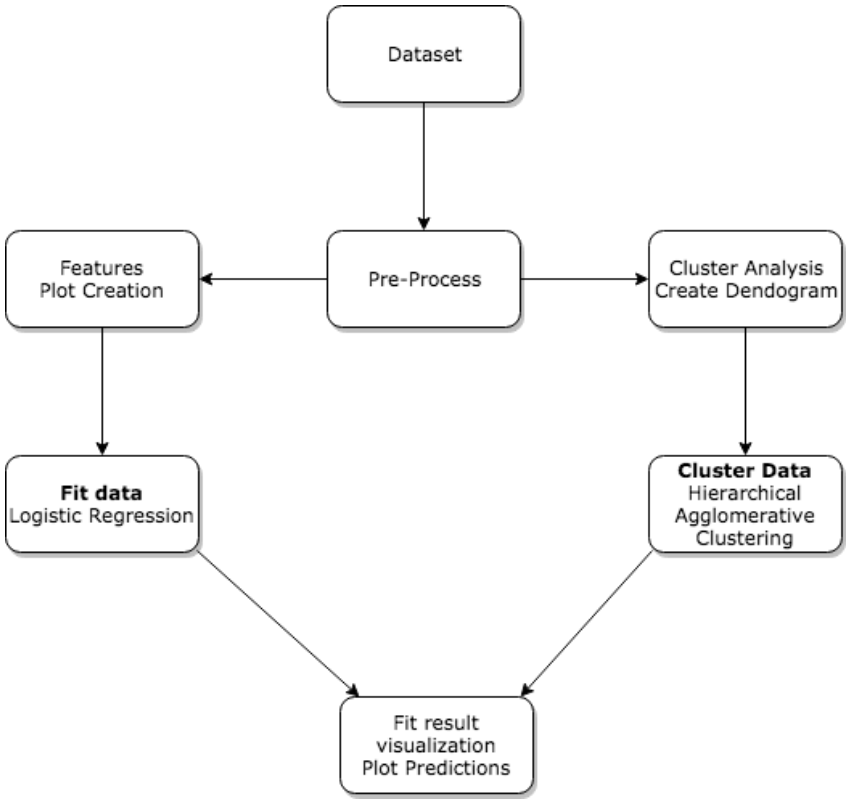
$$Precision = \frac{TP}{TP + FP}$$

Recall:

$$Recall = \frac{TP}{TP + FN}$$

Also to obtain unbiased estimates of the accuracy, precision, and recall properties of the logistic model I will use a three-way cross-validation procedure.

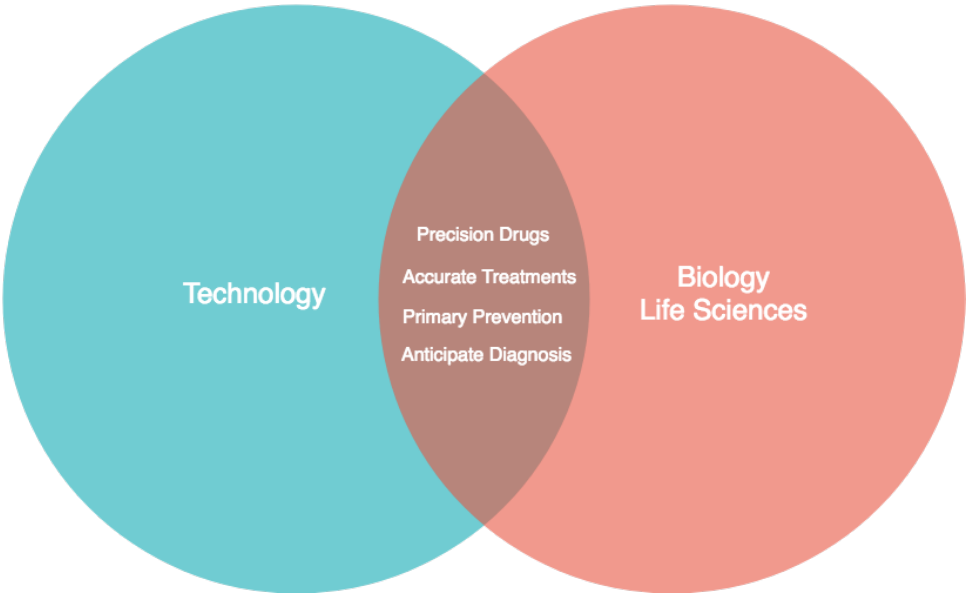
Project Design



To best describe the project design above, I detail the project's workflow:

1. **Dataset** → Download the dataset from the repository
2. **Pre-Process** → Scale and Normalize the data; prepare the features; check for missing values; split the data into test and training sets for cross validation;
3. **Features Plot Creation** → Create plots for the dataset features to best analyze the data
4. **Cluster Analysis Dendogram** → Generate dendogram to inspect the dataset structure and define the best possible number of clusters
5. **Fit data** → Fit the prepared into the logistic regression model
6. **Cluster Data** → Create clusters with the tuned hyperparams and plot the results
7. **Fit result** → Plot predictions for the classification and clusters
8. **Test** → Test the model on the testing dataset
9. **Assess the results.**

After going through the technological process of machine learning to gather data and generate information, I will then try to assess and use the results to understand the patients from a biological standpoint. By bringing the two most important sciences together, from my point of view, we will be able to set some grounds on new research and a new way to analyze data.



Note: I will be using this project as my bachelor degree final paper, for my second graduation (**Nutrition**/**Biochemistry**), with the metabolic pathways approach, from the data I gather after I analyze using the machine learning and statistics point of view. I will also gather more information through several statistical methods besides machine learning, such as standard deviation, mean and variance, in order to obtain more consistent information from the patients data.

References:

http://www.fharrell.com/post/classification/
Classification vs. Prediction

https://www.tandfonline.com/doi/abs/10.1080/01621459.1963.10500845
Hierarchical Grouping to Optimize an Objective Function

https://web.cs.umass.edu/publication/docs/1996/UM-CS-1996-089.pdf
Prototype Selection for Composite Nearest Neighbor Classifiers

https://pdfs.semanticscholar.org/2105/9733ba0f4d4c99666affdb4a7b52242bf386.pdf
Using Rules to Analyze Bio-medical Data: A Comparison between C4.5 and PCL

http://liacs.leidenuniv.nl/~kosterswa/SAC2003final.pdf
Genetic Programming for Data Classification: Partitioning the Search Space

https://en.wikipedia.org/wiki/Hierarchical_clustering
Hierarchical Clustering