Machine Learning Engineer Nanodegree

Capstone Proposal

Dinkar Juyal
March 5th, 2017

Predicting Liver disease from data

Domain Background

In India, delayed diagnosis of diseases is a fundamental problem due to a shortage of medical professionals. A typical scenario, prevalent mostly in rural and somewhat in urban areas is:

1. A patient going to a doctor with certain symptoms.

2. The doctor recommending certain tests like blood test, urine test etc depending on the symptoms.

3.The patient taking the aforementioned tests in an analysis lab.

4. The patient taking the reports back to the reports back to the hospital, where they are examined the disease is identified.

The aim of this project is to somewhat reduce the time delay caused due to the unnecessary back and forth shuttling between the hospital and the pathology lab. Historically, work has been done in identifying the onset of diseases like heart disease, Parkinson's from various features, for example in this paper https://link.springer.com/chapter/10.1007/978-3-319-11933-5_17.In this case, a machine learning algorithm will be trained to predict a liver disease in patients.

Problem Statement

Given a dataset containing various attributes of 583 Indian patients, define a classification algorithm which can identify whether a person is suffering from liver disease or not.

Datasets and Inputs

The dataset for this problem is the ILPD (Indian Liver Patient Dataset) taken from the UCI Machine Learning Repository . Number of instances are 583. It is a multivariate data set, contain 10 variables that are age, gender, total Bilirubin, direct Bilirubin, total proteins, albumin, A/G ratio, SGPT, SGOT and Alkphos. All values are real integers. This data set contains 416 liver patient records and 167 non liver patient records.The data set was collected from north east of Andhra Pradesh, India. Selector is a class label used to divide into groups(liver patient or not). This data set contains 441 male patient records and 142 female patient records. Any patient whose age exceeded 89 is listed as being of age "90".

Relevant sources: Bendi Venkata Ramana, Prof. M. S. Prasad Babu and Prof. N. B. Venkateswarlu, â€œA Critical Comparative Study of Liver Patients from USA and INDIA: An Exploratory Analysisâ€□, International Journal of Computer Science Issues, ISSN :1694-0784, May 2012.

Solution Statement

To solve this problem, I will be using one or more classification algorithms covered in the udacity MLND. I will be trying out Logistic Regression, Support Vector Machine, k-nearest neighbours and one ensemble method.

Different combinations of hyperparameters for individual algorithms , like kernel, degree and C for SVM and weights, n_neighbours and algorithms for k-Nearest Neighbours will be tried across the training sets. Depending on their respective performances on the cross-validation sets, the best algorithm with appropriate hyperparameter tuning will be finalised as the solution.

Benchmark Model

Some models have been created to analyze the chances of liver injury in critically ill patients. The following paper gives one such model: https://www.ncbi.nlm.nih.gov/pubmed/26820880. It assigns a LiFe score from 0 to 10 to patients: with 0 denoting low risk and >8 denoting very high risk. The authors of this paper have stated that: 'a significant positive correlation exists between LiFe score and acute-on-chronic liver failure grade, ($r = 0.478$, $P < 0.001$)'. Since this dataset deals only with critically ill people, it may not be considered to be similar to the problem statement, but can be thought of as an approximate benchmark model.

However the problem lies in finding a dataset where the results are given in such a fashion which is easily comparable with our classification values. In datasets like the one mentioned above, it is intrinsically difficult to compare the scores given with our outputs. Therefore, we will use a simple algorithm like Logistic Regression as our benchmark model and try to improve upon its performance by using other algorithms like SVM, ensemble methods etc.

Evaluation Metrics

Since it is a problem of disease classification, we will generate a confusion matrix so that we can know the False Positives as well as the False negatives. Additionally, we will use the F-scores which take both precision and recall into account, like the F-beta score:
$F_\beta = (1 + \beta^2)(\text{Precision} \times \text{Recall})/(\beta^2 \times \text{Precision} + \text{Recall})$

Project Design

First of all, dataset will be accessed using Pandas and data exploration and visualization will be carried out. Any missing value or outlier will be suitably dealt with. Then, dataset will be split into training, cross-validation and testing set. Testing set will be kept aside for final evaluation, while the training set will be used for training the algorithms. For one classifier, several parameters maybe tested using GridSearchCV technique.

Finally, the best performing algorithm will be tested on the testing dataset and evaluation metrics will be calculated to witness the results.