Machine Learning Engineer Nanodegree

Using Supervised Learning to predict whether a Patient has a liver disease or not.

ESVK Sri Harsha

27th June, 2018

Proposal

Domain Background

The number of patients for Liver disease in the state of Andhra Pradesh, India has been increasing continuously because of many factors. Some of such factors are excessive consumption of alcohol, inhale of harmful gases, intake of contaminated food, pickles and drugs. In the case that they were admitted in the hospital, it becomes really important to know whether the patient has the liver disease or not. Hence, this dataset was used to evaluate prediction algorithms in an effort to reduce burden on doctors.

A similar thing has been done here for the classification of liver diseases:

Link: https://www.researchgate.net/publication/319181775_Disease_Classification_Using_Machine_Learning_Algorithms-A_Comparative_Study

This dataset is similar to the one we're using:

Link: https://www.researchgate.net/publication/309210947_Heart_Disease_prediction_using_Machine_learning_and_Data_Mining_Technique

Dataset that we're going to use: https://www.kaggle.com/uciml/indian-liver-patient-records

Problem Statement

Patients with Liver disease have been continuously increasing because of excessive consumption of alcohol, inhale of harmful gases, intake of contaminated food, pickles and drugs. This dataset was used to evaluate prediction algorithms in an effort to reduce burden on doctors.

The doctors need us to classify the patients checking if they have the liver disease or not. Hence, this is a classification problem.

Inputs:

Different levels of the chemicals in liver - These can be used to check if the patient has the disease or not.

Age – This is an important factor as people mostly get diseases as they gradually age.

Gender – This is also an important factor.

Output:

Whether the patient has the Liver disease or not.

Datasets and Inputs

This data set contains 416 liver patient records and 167 non liver patient records collected from North East of Andhra Pradesh, India. The "Dataset" column is a class label used to divide groups into liver patient (liver disease) or not (no disease). 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").

The features in the dataset:

- Age of the patient
- Gender of the patient
- Total Bilirubin

- Direct Bilirubin
- Alkaline Phosphotase
- Alamine Aminotransferase
- Aspartate Aminotransferase
- Total Proteins
- Albumin
- Albumin and Globulin Ratio
- Dataset: field used to split the data into two sets (patient with liver disease, or no disease)

These features tell us about the age, gender, levels of different chemicals, levels of proteins etc.

A total of 583 rows and 11 columns.

Solution Statement

The current situation is that there are a lot of people dying of liver diseases in hospital and it has become a burden for the doctors to diagnose the patients. Our job is to predict whether a patient has the liver disease or not, given the features like Age, Gender, Bilirubin, Proteins etc.

I will first clean the data by filling up any null values in the data set by the respective sets mean. Then, I'll apply one-hot encoding on the gender feature, which will divide the males and females into two different features. Then, I'll apply normalization on the data by using one of the normalization techniques like MinMaxScaler() or StandardScaler(). Then, I'll use the train_test_split to get the X_train, X_test, y_train, y_test. I want to apply one of the machine learning models on them to train the data, predict and test it on the prediction. Then, we can use one of the evaluation metrics to evaluate how the model performs on the data.

Benchmark Model

I consider Logistic Regression model as the benchmark model because the data is linear. I choose fbeta score to be the benchmark evaluation metric. I'll further test with other models like ensemble methods and decision trees to see if they are performing better. Also, I'll try

using other evaluation metrics like Accuracy score, Confusion matrix etc. to get the best of my model.

Evaluation Metrics

Precision tells us what proportion of messages we classified as spam, actually were spam. It is a ratio of true positives (words classified as spam, and which are actually spam) to all positives (all words classified as spam, irrespective of whether that was the correct classification), in other words it is the ratio of

[True Positives/(True Positives + False Positives)]

Recall (sensitivity) tells us what proportion of messages that actually were spam were classified by us as spam. It is a ratio of true positives (words classified as spam, and which are actually spam) to all the words that were actually spam, in other words it is the ratio of

[True Positives / (True Positives + False Negatives)]

For cases like this dataset, precision and recall come in very handy. These two metrics can be combined to get the F1 score, which is weighted average (harmonic mean) of the precision and recall scores. This score can range from 0 to 1, with 1 being the best possible F1 score (we take the harmonic mean as we are dealing with ratios).

Project Design

I will first load the data from the csv into a variable. Then, I'll copy the 'Dataset'(Which has the information on whether the patient has the disease or not) into 'labels'. After that, we clean the data by filling up any null values in the data set by the respective sets mean. Then, I'll apply one-hot encoding on the gender feature, which will divide the males and females into two different features. Then, I'll apply normalization on the data by using one of the normalization techniques like MinMaxScaler() or StandardScaler(). Then, I'll use the train_test_split with the input of transformed_features and labels to get the X_train, X_test, y_train, y_test.

I'll first apply the benchmark model by fitting the data with the logistic regression and predicting the values. Now, I'll check with the benchmark metrics to check how my model is working.

Later, I'll apply other models from ensemble methods like Gradient Boosting, AdaBoost, XGBoost etc. and Decision Trees like Decision Tree Classifier and Random Tree Classifier, applied with the evaluation metrics to see how the data is working with different models.

Of these, I'll choose the best model based on the metrics and I plan to further improvise and optimize it by applying the GridSearchCV.

I might add some visualizations to even better the evaluation process.

References

- https://werlabs.co.uk/liver-function/bilirubin/
- https://medium.com/greyatom/performance-metrics-for-classification-problems-in-machine-learning-part-i-b085d432082b
- http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html