Machine Learning Engineer Nanodegree

Capstone Proposal

Charu Chhimpa
August 14th, 2017

Domain Background

Advancement in the field of medicine have greatly improved our quality of life which can be clearly seen from the life expectancy rate. From 1816's stethoscope to today's handheld ultrasound machines, doctors have steadily adopted technology to advance healthcare. In past decades the healthcare community has taken major steps by adopting electronic healthcare measures. Machine Learning can contribute significantly in the field of healthcare, it has great usage in cases like early disease detection, finding signs of early breakouts of epidemics, using clustering to figure out regions of epidemics, or finding the best air quality zones in countries with high air pollution.

Heart disease is the leading cause of death in the world over the past 10 years. The World Health Organization reported that heart disease is the first leading cause of death in high and low income countries. There are several methods in the literature individually to diagnosis diabetes or heart disease. There is no automated diagnosis method to diagnose Heart disease for diabetic patient based on diabetes diagnosis attributes to our knowledge. Researches have been using several data mining techniques in the diagnosis of heart disease. Diabetes is a chronic disease that occurs when the pancreas does not produce enough insulin, or when the body cannot effectively use the insulin it produces. Most of these systems have successfully employed Machine learning methods such as Naïve Bayes and Support Vector Machines for the classification purpose. Support vector machines are a modern technique in the field of machine learning and have been successfully used in different fields of application.

Problem Statement

In this project we will train a model using Support Vector Machines to predict that a human being is suffering from a heart disease. We will consider several features to do this classification.

Task : To predict that a person is suffering from a heart disease or not.

Performance : Accuracy - No. of correct predictions.

Datasets and Inputs

I am using the UCI Heart Disease dataset for training the model. This database contains 76 attributes, but the best results are obtained using a subset of 14 of them. In particular, there is data from 4 hospitals but I will use the Cleveland Database because it is properly processed. The dataset have 303 instances. The dataset have a balanced label distribution. The dataset link is : **http://archive.ics.uci.edu/ml/datasets/heart+Disease**

The different features that are used are :

1. age: age in years
2. sex: sex (1 = male; 0 = female)
3. cp: chest pain type
   -- Value 1: typical angina
   -- Value 2: atypical angina
   -- Value 3: non-anginal pain
   -- Value 4: asymptomatic
4. trestbps: resting blood pressure (in mm Hg on admission to the hospital)
5. chol: serum cholestoral in mg/dl
6. fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
7. restecg: resting electrocardiographic results
8. thalach: maximum heart rate achieved
9. exang: exercise induced angina (1 = yes; 0 = no)
10. oldpeak = ST depression induced by exercise relative to rest
11. slope: the slope of the peak exercise ST segment
    -- Value 1: upsloping
    -- Value 2: flat
    -- Value 3: downsloping
12. ca: number of major vessels (0-3) colored by flourosopy
13. thal: 3 = normal; 6 = fixed defect; 7 = reversable defect
14. num: diagnosis of heart disease (angiographic disease status)
    -- Value 0: < 50% diameter narrowing
    -- Value 1: > 50% diameter narrowing

Solution Statement

I will use Supervised Learning approach to learn different features mentioned above and accordingly classify that the person is suffering from heart disease or not. The dataset used is the Cleveland database which is pre processed before to seperate each of the 14 features by a comma. Supervised learning would finally provide us with binary output '0' or '1' based on the patient features.

Final Model that I would use is Support Vector Machines. Support vector machines are a modern technique in the field of machine learning and have been successfully used in different fields of application.

Benchmark Model

Gaussian Naive Bayes

In machine learning, naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features.

When dealing with continuous data, a typical assumption is that the continuous values associated with each class are distributed according to a Gaussian distribution. For example, suppose the training data contains a continuous attribute, x. We first segment the data by the class, and then compute the mean and variance of x in each class. Let μc be the mean of the values in x associated

with class c, and let σ^2_c be the variance of the values in x associated with class c. Suppose we have collected some observation value v. Then, the probability distribution of v given a class x, p(x=v), can be computed by plugging v into the equation for a Normal distribution parameterized by μc and σ^2_c. That is,

$$p(x = v \mid c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} e^{-\frac{(v-\mu_c)^2}{2\sigma_c^2}}$$

I'll be using the Naive Bayes classifier as the benchmark model because it will always predict either of the one class.

I'll be looking to maximize the accuracy of the predictions using this model.

Evaluation Metrics

Prediction Accuracy : The accuracy score will be used as an evaluation Metrics in this case. It would be calculated using the number of data points which are classified correctly.
As it is a binary classification problem, prediction accuracy turns out to be the best evaluation metrics.
If y_hat_i is the predicted value of the i-th sample and y_i is the corresponding true value, then the fraction of correct predictions over n_samples is defined as :

$$\texttt{accuracy}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} 1(\hat{y}_i = y_i)$$

Project Design

Programming Language and Libraries

- **Python2**
- **scikit-learn** : Open source machine learning library for python.
- **numpy** : Python's numberical library.
- **matplotlib** : For plotting graphs and plots.
- **pandas, seaborn** : For data reading and visualization.

Operation

First the dataset would be extracted in a proper format using the scikit-learn functions. We will then convert all the categorical varaibles with more than two values into dummy variables.

And then the Support Vector Machine model would be trained using the processed dataset. SVM functioning is explained above.

Finally we can can calculate the prediction accuracy to see that how well our model is performing.

This would be the over all functioning of the algorithm.

References

- https://link.springer.com/chapter/10.1007/978-3-642-27443-5_25
- http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.401.7117&rep=rep1&type=pdf
- https://dzone.com/articles/a-tutorial-on-using-the-big-data-stack-and-machine
- https://hbr.org/2017/05/how-machine-learning-is-helping-us-predict-heart-disease-and-diabetes

---