Assignment 1

Abhishek Yadav(2015MT10585)

February 2018

Introduction

In this assignment we were given to study various classification algorithms and model selection algorithms and we had to compare various classification strategies then we were to implement logistic regression from scratch and then used other classification strategies from python libraries and compared the results.

1 Survey of various classification algorithms

1.1 Logistic Regression

It is used for analyzing data sets in which one or more than one variables are used to analyze the data sets and predict an binary outcome.The dependent variable is binary.The main use of the logistic regression is to predict the best fitting model that relates the independent variables to dependent variable
Assumption

- It requires the dependent variable to be binary.
- It requires the observation to be independent of each other.
- It also requires the independent variables not to be highly correlated.
- It usually requires large samples.

Advantages

- The independent variables need not have to be normally distributed.
- It also takes care of non-linear effects.
- The independent variables need not to be highly correlated.
- It usually requires large samples, so better outcomes.

Drawbacks

- It requires very large amount of data to achieve meaningful result.
- It can't predict continuous outcome.
- It may over-fit the model, it overstate it's correctness of prediction.

objective Function

$$logit(p_i) = \beta_0 + \beta_1 x + ... + \beta_n * x^n$$
$$logit(p_i) = \ln \frac{p_i}{1 - p_i}$$

So by solving the above equations we get the correct value of p_i as follows:

$$p_i = \frac{1}{1 + \exp^{-\beta * X_i}}$$

1.2 Decision Tree

It is used to classify the data based on many criteria, it takes one criteria and then classify the data into two parts and then again does the same for the remaining parts.It basically builds a tree based on the fields and classifies the data.Then to predict on the test data, it follows the given trees.
Assumption

- It doesn't much assumption in the pattern of the input data sets.

Advantages

- It specially perform variable screening.
- The data requires much less effort from user to prepare.
- Non-linear relationship in parameters doesn't affect much the performance.
- One of the best and important feature is, it is too easy to understand.

Drawbacks

- It is unstable, very small change in data can cause very large change in the tree.
- Time Complexity of preparing a decision tree is very high, though it is easy to understand but making a decision tree requires much large amount of time.

- Even if the given data set perfectly fits the decision tree still the size of decision tree can be very large. 4.They are sometimes very inaccurate, sometimes with the same data other classifiers perform much better than the decision tree.

objective Function
There is not an objective function as we make the decision tree based on the given data.

1.3 Random Forest

It creates bunch of random trees by randomly selecting the data from training set and then aggregates the votes to decides the last class of test output.
Assumption

- It doesn't much assumption in the pattern of the input data sets.

Advantages

- it reduce the chances of over fitting as compare to decision tree, as it uses set of decision trees.
- Less variance:We reduce our dependency on one classifier which may not perform well because of inter relation between test and train data.
- It is one of the best classifier known as it run efficiently on large data sets.
- It gives an estimate as which are the important variables.

Drawbacks

- It is an predictive modelling tool , if we need an description in our modelling we can't use it we will be using different one.
- Time Complexity of preparing a decision tree is very high, though it is easy to understand but making a decision tree requires much large amount of time(Random forest requires making of decision trees).
- It can sometimes over fit the data sets with noisy classification tasks.
- As we are using multiple decision trees sometimes they don't give lots of insights.

objective Function
As it also works based on decision tree, it also doesn't have objective function.

1.4 Naive Bayes Classifier

It is collection of algorithms which are used to classify the data sets based on bayes theorem, they are group of algorithm and share common characteristics that is the pair of feathers we are going to classify are not dependent.
Assumption

- We assume that the feathers we are going to classify are independent of each other.
- All the features are given importance, none of them are assumed to be irrelevant.

Advantages

- We can very easily implement it.
- It requires small amount of training data to estimate the parameters.
- It is very simple and usually produce good results.

Drawbacks

- One of the major drawbacks of naive bayes classification is that it assumes that the feathers are independent of each other.
- One of the problem arises if the feathers are continuous we need to make them discrete which may cause problem as we may loss lots of information.
- It can't learn interaction between feathers.

objective Function
Given a problem instance to classify, and vector $(x_1, x_2, ..., x_n)$, it assigns to instance probabilities to

$$p(C_k|x_1, x_2, ..x_n)$$

, but using bayes theorem we know,

$$p(C_k|x) = \frac{p(C_k) * p(x|C_k)}{p(x)}$$

, which upon further simplification and assumption will end up giving expression as

$$p(x) = \sum_k p(C_k) * p(x|C_k)$$

1.5 Neural Network

In a neural network classifier, it consist of layers of neurons,in which input vectors are converted to some output vectors which classify them and each layer takes input from previous and we set bias and weights for corresponding fitting of the data.

Assumption

- Depending on the loss function we may have some assumption on the data, it varies on our function.
- We usually don't assume much on the data, we usually do exhaustive search for finding the patterns.

Advantages

- It usually requires less formal statistical training.
- It has ability to find out the complex relationship between the dependent and independent variables.
- It has ability to predict all the relationships between the variables who are predicting and it has many training algorithms.

Drawbacks

- It has drawbacks as it requires large amount of computation.
- It has chances that it can over fit the data sets.
- .it's various algorithms have various drawbacks.

1.6 K-nearest neighbour

It stores all the available cases and classify the new cases based on similarity measure it may be distance function or something else. It is also used in statistic also, it has also been used in pattern recognition.

Assumption

- It assumes that those data points which are near to each other distance wise are similar in properties.
- Weights are sometimes defined to depict the closeness of the data sets.

Advantages

- It is simple to implement.

- It's our choice that we can choose our distances as we want.
- It handles it self the multi-class data.

Drawbacks

- We need to search exhaustively to find nearest neighbour.
- Space complexity can be a problem too.
- We must find a very suitable distance function which can fulfill requirement .

objective Function
Suppose we are given a point Y to classify, we will classify y as the nearest to k neighbour, we will make a circle or any other diagram (depends on the objective function we are using) to in close the k nearest neighbour then depending on the confidence we will classify the sample with the given confidence to a given class.

$$dist(x, y) = (\sum_{n}(x_i - y_i)^2)^{\frac{1}{2}}$$

There are other type of distances also defined over different distance metric but we usually use euclidean distance.

2 Summary of various scoring methods for classification

2.1 Accuracy

It is defined as the number of correct prediction made over all the prediction made by the model.

- It is to be used when target variables are nearly balanced.
- It should not be used when target variables in the data are of one class mainly.

$$acc = \frac{TP \ + \ TN}{n}$$

where acc=accuracy ,TP(True Positive) total positive correct prediction made and TN(True Negative) total negative correct prediction made and n is total model predicted .

2.2 Precision

It is defined as the fraction of output which are actually positive over which were predicted positive.

- It basically means the fraction of relevant instances over all our retrieval instances.

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

where p=precision ,TP(True Positive) total positive correct prediction made and FP(False Positive) false positive prediction made.

2.3 Recall

It is defined as among all the positive fraction which are actually positive what are the fraction we predicted positive.

- It basically means the fraction of relevant instances that we got over all the relevant instances.

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

where r=recall ,TP(True Positive) total positive correct prediction made and FP(False Negative) which we predicted negative and were actually positive.

2.4 F1 Score

It basically takes into account both precision and recall.It is a single measure which takes both precision and recall to give its result.

- It could have been arithmetic mean of both, but that will turn out to be much less influential in some cases.
- So it is basically harmonic mean of precision and recall.

$$p = \frac{2 * p * r}{p + r}$$

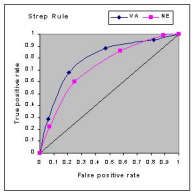
where p=precision ,r=recall.

2.5 Receiver Operating Characteristics(ROC)

It is basically created by taking true positive rate vs false positive rate and plotting them against each other at various threshold setting.

- It is a method of evaluating the performance of diagnostic tests.
- true positive rate and false positive rate ki definition.

2.6 Area Under The Curve

It is equal to the probability that a classifier will rank a randomly chosen positive instance higher than that to a negative one. 

3 Multi-class classification strategies

3.1 One vs One Classification Strategies

If there in an N-way multiclass problem then in OVA one trains N(N-1)/2 binary classifiers for it,samples for different pair of classes are received from training sets and we learn to distinguish among-st them.When we predict, we apply all the N(N-1)/2 classifiers on it and the one with the highest +1 predictions gets predicted.

3.2 One vs All Classification Strategies

In it we train single classifier every class assigning the samples of that class positive and others negative.It need the base classifier to produce an function which gives real valued confidence score for it's decision not simply label of a class.

3.3 comparison

- In OVO the items are distinguished between only two classes, For Example if we have three classes A,B and C then there will be three classifier which will classify them such that in one case only A and B are distinguished and in others only B and C or A and C.
- In OVA the distinction is made between the one class and other i,e. We will divide our classifier as if it belongs to A or other than A.
- In OVO a binary classifier is built between each pair.
- Whereas in OVA classifier is built in one vs remaining.

4 Model Selection Techniques

4.1 AIC(Akaike Information Criterion)

It is one way to choose a model from set of models and it is chosen in such a way that minimizes the KL distance between the model and the truth.It can be

seen as a good fit to the truth value but with few parameters.

$$AIC = 2k - 2ln(L)$$

where k=number of estimated parameters in the model,L is the maximum value of likelihood function of the model. **Advantages**

- It can simultaneously compare multiple nested or non-nested model.
- It allows the estimation of model parameters using all available models.

4.2 BIC

It is criteria of model selection in which the one with lowest BIC is preferred,it is based in part on likelihood function and is closely related to AIC.

$$BIC = ln(n) * k - 2ln(L)$$

where k=number of estimated parameters in the model,L is the maximum value of likelihood function of the model,n=number of data points in x. **Limitations**

- We approximate the formula for BIC and that is only valid when n is much larger than the value of k.
- It can't handle complex collection of models, as feather selection in higher dimension.

4.3 R-Squared

In this case we compute the regression model for each subset of independent variable and finally we choose that criteria which optimizes the selection criteria being used.

$$R^2 = 1 - MSE/TSE$$

Types :

- Backward Selection.
- Forward Selection
- Stepwise Selection.

4.4 Mallow's C_p

It is used to access the fit of a model that has been estimated using least squared model.It is used in the context selection of model in which subsets of parameters are used to predicts the best fitting model. The low value of C_p means more precise. **Limitations**

- The approximation of C_p is valid only if the sample size is large.
- It can't handle the complex selection of model.

4.5 Hyper-parameters

It is the parameters, whose values are initially determined, even before the learning begins.We initially starts with these values and the values of other parameters are determined sub sequentially.

4.6 Tun-able-parameters

They can be defined as those parameters which are optimized in order our algorithm to perform well, so that we can get the best result by optimizing these parameters.

4.7 learn-able-parameters

These parameters are initially randomly selected and then as we feed in the training data they are optimized and corrected so as to give the corrected output.

5 Implementation of Logistic Regression Algorithm

I basically used One Vs All method to learn the data sets and then predict the out comes for various cases.With the value of learning rate to be equal to 0.01, i got 2877 outcomes to predicted correctly out of 3498 total number of points on test data set .We can increase the number of points predicted correctly by using more smaller value of learning rate but that will come at a cost of time, as then it will take more time to converge the algorithm. On training data-set i ran the algorithm to get an accuracy of 0.60 with same learning rate. I have attached the plot over iterations. I choose the score of 0.005 because as we had a large number of data sets,and we have to classify them into 10 classes, considering one of them to be 1 and remaining to be 0, so it had basically required a large number of iteration, if i would have chosen any value which would have been smaller than that, then it would have taken more time and may have over-fitted the data. A large value of learning rate would have caused under-fitting of the curve(data). So to avoid all those things,0.005 seemed a better choice to me. When i used the value of learning score to 0.01, i got my iteration completed in very less time but the accuracy was quite low, it was 1988 out of 3498 test cases.
When i used it on my training data, i got an accuracy of 0.8667. with the learning rate of 0.05 the accuracy was very low, it was just as we randomly classified the data and algorithm terminated very fast, this shows our choice of learning rate was best.

6 Comparison

6.1 KNN Algo

Total 3414 were correctly predicted out of 3498, accuracy=0.9759

6.2 Decision Trees

Total 3207 were correctly predicted out of 3498, accuracy=0.9168

6.3 Naive Bayse Algo

Total 2824 were correctly predicted out of 3498, accuracy=0.8073

6.4 Nearest Centroid Method

Total 2657 were correctly predicted out of 3498, accuracy=0.759

7 Conclusion

I studied various classification technique and then learned about model selection techniques in last while implementing logistic regression from scratch and using other classification from inbuilt libraries, the accuracy given by the algorithms from the libraries where sometimes much better than the logistic regression and some times results where not much better, so i concluded for one type of data one classification may be much better than the other type of data.The classification techniques are made considering various types of data and they are best for only those types and they may or may not be better for other data for classification.

8 Reference

- http://www.brighthubpm.com/project-planning/106005-disadvantages-to-using-decision-trees/
- http://www.simafore.com/blog/bid/62333/4-key-advantages-of-using-decision-trees-for-predictive-analytics
- http://amateurdatascientist.blogspot.in/2012/01/random-forest-algorithm.html
- https://www.quora.com/What-are-some-advantages-of-using-a-random-forest-over-a-decision-tree-given-that-a-decision-tree-is-simpler
- https://medium.com/machine-learning-101/chapter-5-random-forest-classifier-56dc7425c3e1
- https://www.medcalc.org/manual/logisticregression.php

- https://classroom.synonym.com/disadvantages-logistic-regression-8574447.html
- https://victorfang.wordpress.com/2011/05/10/advantages-and-disadvantages-of-logistic-regression/
- https://www.slideshare.net/ashrafmath/naive-bayes-15644818
- https://www.geeksforgeeks.org/naive-bayes-classifiers/
- https://www.sciencedirect.com/science/article/pii/S0895435696000029
- https://medium.com/greyatom/performance-metrics-for-classification-problems-in-machine-learning-part-i-b085d432082b