DATA MINING

ASSIGNMENT 2
04-05-2019

COMPARISON OF TWO CLASSIFIERS

Rashmi Sudarshan (1PE16CS129 rashmisudarshan01.com)
Shashidhar V (1PE16CS144 shashidhar.v10@gmail.com)

PROBLEM STATEMENT

The problem statement here is that an automobile industry is launching its new service chain at a particular place and they have provided an inaugural offer to the customers. The offer is that they will provide free service to all the customers who drop their car on the first day. The problem is that under maximum capacity the service station is equipped to service a maximum of 315 cars per day. But due to this offer there are a huge number of cars to be serviced. Since the company has promised first day service for free they cannot send them back unserviced. So the company approaches a data scientist to solve this problem for them which is to determine which cars need servicing based on a list of parameters which they have recorded during the servicing of the other cars.

Another experimental problem is to analyse the performance of two classifiers. The classifiers are operated on the same dataset and evaluated on the outcomes produced by the models based on some performance measures.

We have chosen to compare between Logistic Regression and k-Nearest Neighbours classification algorithm.

At the end of this analysis we would be able to conclude which classification algorithm performs better and under what circumstances.

STOCK TAKING OF DATA

The dataset consists of 316 samples. There are 6 attributes in the dataset. They are Oil Quality, Engine Performance, Normal Mileage**,** Tyre Wear, HVCA Wear and Service. The first five are numerical valued features and the last one is a categorical feature which says yes or no.

The F-statistic measure was calculated to determine the outliers between 2*sigma and -2*sigma and it was determined that there was no outliers in this particular dataset.

There was no inconsistency observed in the data. But since the data was all numerical except for one attribute normalization of the data had to be done.

BRIEF LITERATURE REVIEW

Another method that can be used to solve the problem of binary classification like the problem statement above is stochastic gradient descent. In this method, only a few samples are randomly selected from the data set for each iteration of calculating the gradient. Hence it is more computationally efficient for modelling very large data sets.

We can also use support vector machines to solve this problem. To do so, all that we need to do is to identify the right hyper plane which most accurately classifies a majority of the data points. SVM would also make a good choice because it is robust to outliers. Tuning the parameters of an SVM helps to increase the performance of the model as well. Granular SVM is a variation of SVM that can be used to solve this problem by building a sequence of information granules and then building an SVM in each granule.

FEATURE ENGINEERING

Feature engineering is the method by which the necessary features(attributes) of the dataset are preserved and the rest are eliminated or merged with another feature to form a new combination of features.

Since this dataset involves five numerical attributes and only one categorical attribute, the numerical attribute values were normalized in order to perform comparison between them. The correlation matrix was constructed out of the normalised data and it was found that there was no high correlation among the data and hence we didn’t need to eliminate any attributes.

Certain attributes like oil quality and engine performance were thought to be interrelated and was therefore combined to make a new attribute but then the accuracy decreased and hence had to be dropped.

CLASSIFICATION

knn Classifier

knn stands for k nearest neighbours classifier. It determines which class a particular data sample belongs to on the basis of votes of the k nearest neighbours. The data sample belongs to the class with the highest number of votes.

In our example there are only two classes which are Yes or No. The test sample is a car and the classifier has to predict whether the car needs to be serviced or not. In our case we have the luxury of already knowing which class the test sample also belongs to so it allows us to compute the accuracy. Otherwise we would have had to stop at the classification itself.

The test dataset contains 135 test samples and the train dataset contains 316 training samples. Service contains two classes yes or no to be determined by the classifier.

Logistic Regression

Logistic regression is a regression analysis method used when the dependent variable is binary as is the case with our data set(the car is classified as either “to be serviced” or “not to be serviced”). It is a predictive analysis technique but instead of predicting the class exactly, it generates the probability that a data point belongs to a particular class.
Hence, for our particular model, if the probability generated is greater than or equal to 50% the point is classified as “Yes” and less than 50% is classified as “No”. In general however this threshold can be set to any value depending on the data set being worked with.

EXPERIMENTAL RESULTS

knn Classifier

The knn classifier performs with a stunning accuracy of 100% with respect to this dataset only at lower numbers of k. But when k was set to larger values the accuracy dropped drastically. This leads to the conclusion that:

1) The lower value of k leads to a very crisp decision boundary and does not allow the inclusion of noise and outliers.
2) The higher value of k leads to inclusion of noise and outliers leading to decrease in the accuracy and increase in chances of misclassification.

There are many ways adopted by data scientists to determine the value of k.

One such method is known as k fold cross validation method where we run the algorithm for k folds for different values of k and then determine from the accuracy the appropriate value of k.

Another method is elbow method where one uses statistics of the data to compute value of k.

Here in this data we use k=3 to determine the class using knn classifier.

Logistic Regression

Logistic regression classifier performs with an accuracy of 91.11% with respect to this data set with a threshold of 0.5. However, changing the threshold to 0.4 increased the

accuracy to 91.8% and further decreasing it to 0.2 brought the accuracy back to 91.11%.
Hence to pick an optimal threshold depends on the scenario. If the objective is to have unbiased predictions, then we must set the threshold in such a way that maximizes sensitivity and specificity. However if the objective is to maximize the accuracy of the model then the cross validation method must be used to determine the best threshold value.
So for the current data set, we choose the threshold value to be 0.4 which maximizes accuracy.

DISCUSSION

knn Classifier

The confusion matrix was computed manually and also using the inbuilt R library. There were no misclassifications and hence an accuracy and kappa score of 1 were obtained. This happens very rarely. I think this is the case because the dataset of the course was fine tuned for this outcome. The specificity and sensitivity values were also 1.

Logistic Regression

Using logistic regression, we obtain the following confusion matrix: true positives=33 , false positives=9, true negatives=90, false negatives=3. We can therefore observe that more number of records were misclassified in this method than in knn. The sensitivity and specificity also dropped significantly to 0.9167 and 0.9091 respectively. The kappa score obtained was 0.7842.

The knn classifier is a non-linear classifier and it works well with various classification problems as well whereas logistic regression is a linear classifier and the problem being solved has to be a linear classification problem. A linear classification problem is the one where a hyperplane is being used to distinguish between two classes.
The next important property is that of hyper parameter tuning. In logistic regression one has to take care of the optimization

of learning rate and regularization parameters to achieve high accuracy apart from optimizing the data. But in knn classifier one has to only determine the optimum value of k and the classification is done.

The next point is that when a point is very close to the decision boundary with respect to logistic regression the chances of misclassification is quite high but with knn since it deals with node votes that problem doesn't occur but it fails miserably when the point is close to outliers. Hence parameter scaling and outlier detection is very important in knn.

Another aspect of logistic regression is that we can derive the confidence interval with which we are making the prediction which is not available to us in knn.

Knn is a non parametric lazy learning algorithm which means to say that it doesn't make any assumptions about the underlying data but logistic regression is a parametric model which means we will have to derive certain parameters out of our training dataset to build the model.

CONCLUSION

Therefore in conclusion, knn(k=3) yields a much better result for this data set due to its non linear nature. Since the decision surface of logistic regression is linear(the hyperplane equation) it is not the best model to use for this type of data set.

CONTRIBUTIONS

Shashidhar worked on the knn classifier and outlier detection using F statistic was a new concept I learnt while doing this assignment.

Rashmi worked on the logistic regression classifier and the literature survey on binary classification models.

REFERENCES

The dataset was takes from the online course on Data Science offered by NPTEL.

The implementation and information about the classifiers were obtained from videos of NPTEL.

The blogs published under Towards Data Science on Medium.com were also referred to.