Module: CSCM45 – Big data and machine learning

Student id: 843042

Coursework 2

Introduction

The world today offers us great technologies like very performant cameras or high quality videos which we all certainly very much enjoy. But what does all this imply and which are main characteristics? From a machine learning point of view we will look at the exponentially growing sizes and we are going to analyse different methods that can process all this data and recognize different patterns.

The two main categories of methods that can be used for pattern recognition are supervised and unsupervised learning. The first one describes situations where real sets of X and Y are known and based on their characteristics further predictions can be made. The unsupervised learning describes a situation where a set of data is given without any other information. Then by using different algorithms patterns should be discovered in order to determine the structure of the data.

Along this document we are going to analyse different supervised and unsupervised learning methods which are to be applied to the following example of 10 image categories, each of them containing 30 pictures. The Aim is to analyse the first 15 pictures of each category and draw conclusions which will be applied to the other half of the sets. Since the two sets of features which we can use for predictions are being given, we can therefore conclude this is a supervised learning situation. However, unsupervised methods can also be used if ignoring the information given by the feature sets. In the next section we are going to look at different supervised learning methods, get a better understanding of them, test and interpret their results.

Method

In order to compute different machine learning methods on the given datasets we first need to adapt its structure. Each cell of the training and test features matrix are other matrixes but the format that we want is a matrix containing only single values and no other matrix in each cell. In order to achieve the proper format for analysis we go through all 10 categories and normalize the data for each first 15 pictures which are our training data. By using a histogram method we reduce each cell/each image to a smaller matrix and therefore we reduce significantly the dimensionality of the data. Since the images have different sizes we also want to reduce their dimensionality to the same number of bins so that each histogram matrix will have the same dimension. Then all the information in the reduced matrix is copied as one row in the X training set. Finally each row of the X training set will consist of each training picture's histogram information in each row. In order to get the Y training set, for each processed image we will save its category number on its corresponding line. Same algorithm is applied for the test data in order to get the proper structure.

In order to test which impact has the number of bins to the accuracy of the methods, all the calculations and predictions have been made for bins varying from 10 to 100. Another attempt of increasing the accuracy of the methods was to eliminate all the null columns from the data sets since they don't offer any relevant information and restrict the number of different values to significant higher level.

Support Vector Machines (SVM)

SVM is also known as a large margin classifier. It constantly looks for the larger minimal distance between different training groups. In order to achieve the aimed results this method tests each category against all others and separates them with a support vector. After computing all the separation lines they are adjusted so that the final result is as accurate as possible for the entire model.

SVM can be used for different type of classifications. One of the most common approach is the linear function which is also known as the "no kernel" method. Just as LDA, it uses a linear approach, but the process is more complicated and therefore it also takes longer time to compute. Since they are so similar we know that if one of them returns good result the other one will as well. The data is computed in a 2 dimensional space. The linear kernel is recommended when there is a small training set and a large number of features.

Further kernel methods are for example Gaussian and polynomial. In this case we look at the data in a multidimensional space and also search for patterns using different computational algorithms. The Gaussian function is supposed to be a better choice especially for a large training set and a small number of features. The polynomial kernel is recommended for positive data [1].

The SVM method first generates a template according to the chosen function (liner/Gaussian/polynomial), then it fits a multiclass SVM classifier which is used for further testing and generation of the predictions.

Linear Discriminant Analysis (LDA)

LDA is a supervised learning method which divides data into classes by maximizing the distance between them and minimize the distances between the elements of the class. The method is very similar with Principal Component Analysis (PCA), which is to be applied for unsupervised learning.

Since LDA focuses on comparing the centroids of different classes, it has a high rate of success when the two sets are clearly separated and definitely not overlapping. If the two centroids are close to each other, no clear conclusion can be drawn.

LDA first examines the prototype classes and then generates discriminative information to new classes, it generalizes the characteristics that were learned from the training set by applying them to the test set. The accuracy can be calculated by dividing the correct matches to the entire number of the set.

Neural Network (NN)

NN is a complex method used for non-linear classification in very large networks. It was mainly developed with the scope of applying it in biology while having a main focus on the study of the brain. As the name

suggest the neurons are a key part of this topic. Among many characteristics that they have they also receive and send away information, they help the information circulate through different layers of the brain. Having said that, the NN contains information about neurons regarding the inputs and outputs, as also the matrix of weights (parameters) reflecting the mapping from a layer to another. Here we also have the concept of a hidden layer which contains values one does not observe in the training set. It is situated between the input and output points of the entire network. In order to implement the propagation of information from the input layer to the hidden layer(s) until the output layer it is used a vectorized implementation of the model while it is also learning its own features as well. The series of weights are randomly assigned to the network at first and they keep being adjusted while going through all the layers of the network in order to learn as much as possible about the network and output a correct result.

Results

In order to get a better understanding of what variables influence our results I have calculated the correlation between the number of bins and the accuracy of the all the used methods. They all use at first training test data in order to determine patterns which are secondly applied to the test data. The accuracy can be calculated by dividing the correct matches to the entire number of the set.

Test set 1 consists of only columns that have at least 100 positive values while test set 2 contains all columns that have at least one positive value. The following conclusions have been made after analyzing a separate set of 50 results for each method. In this section we are going to refer to the various accuracy results from the tests as set of observations and try to find which influence had the variation of bins on these results. By using the correlation coefficient, significance F and R square indices we will conclude if further results can be predicted using this test sets. The predicted results refer to the possibility of determining the accuracy only by defining the number of bins.

In the table below we can see that the maximum accuracy of 65.33 has been achieved by using the LDA method, while the SVM Gaussian model has a maximum accuracy of 61.66. They are both linear models. Therefore, it makes sense that the results are fairly similar. The least accurate methods are for this dataset the SVM Gaussian and NN models, while the SVM polynomial is situated in between.

		SVM linear	SVM Gaussian	SVM Polynomial	LDA	NN
Correlation coefficient (Pearson)	Test set 1	-0.221	N/A	-0.44	**-0.86**	-0.244
	Test set 2	-0.609	N/A	-0.375	-0.76	-0.357
Significance F (ANOVA)	Test set 1	0.139	N/A	**0.001**	1.9E-14	0.1
	Test set 2	7.01E-06	N/A	**0.01**	8.66E-10	**0.014**
R square	Test set 1	0.048	N/A	0.254	0.739	0.059
	Test set 2	0.371	N/A	0.141	0.578	0.127
Max value	Test set 1	60.66	10	44	**64**	20
	Test set 2	61.66	10	48.66	**65.33**	17.33
Min value	Test set 1	50	10	16	44	6.66
	Test set 2	56	10	22.66	52.66	4.66

Figure 1. Test results summary

The Pearson number represents the correlation between 2 variables. In our case it reflects how much influence has the number of bins on the accuracy of the data. According to Pearson's theory the correlation coefficient is between -1 and 1. The closer the value is to 1 or -1 the stronger the bond is as when a result close to 0 describes no connection between the 2 variables. Furthermore, a result close to -1 also reflects that the two variables are changing in opposite direction while 1 shows changes in the same direction. In the table above we can notice the stronger bond is represented for the LDA method which has a correlation coefficient of -0.86. Being close to -1 we can conclude the bin size has a strong influence on the accuracy of the method. For the SVM linear method the coefficient is very close to zero which means there is almost no correlation between the two variables. Having looked at these extreme values we can examine the SVM Gaussian method which is only moderately influenced by the number of bins. Although the influence is not strong it is still meaningful since Pearson's coefficient is smaller than -0.5. Since all the values are negative we can conclude that overall the more the number of bins increases the less accurate a method will be. A greater number of bins also means larger bins. Therefore the larger the bins of data the best accuracy we will get especially in the LDA method.

The R square parameter describes the variation that is explained by the model. Ideally, its value is as close to 1 as possible. This means that the entire model is explained by the independent variables. According to the results in Figure 1, for LDA around 74% and 59% of the two models can be explained the numbers of bins. For all the other methods the percentages are below 40% which shows a very week influence of the number of bins on the accuracy on the model.

In order to check if the predicted results are reliable we need the significance F (ANOVA) coefficient which should be smaller than 0.05. Other greater values show that the independent value(s) don't have a significant impact on the model and one should probably stop using them. In our case, the only valid numbers are returned for the SVM polynomial method for both sets and another time for NN. This means that in these three cases the number of bins can be used to predict the accuracy. However, we already know that they explain between 12% - 25% of the model.

With the scope of increasing the accuracy of the tests, I have also eliminated all null column so that the data is more consistent. Unfortunately, no significant improvement was to be noticed, also not when the amount of relevant values was restricted to at least 100 out of 150. After further analysis I discovered that the variation of the data in the column was low which is why the accuracy only slightly changed.

Conclusions

In this paper we have analysed different supervised methods and also explored a few ways of improving their accuracy. After testing and interpreting the results we can therefore conclude that the content of the datasets has a greater impact on the accuracy than the computing itself. Dimensionality reduction and choosing the proper methods play an import role in preparing the data and analyzing it. A very good understanding of all available methods increases the chances of making a good choice when analyzing data and drawing conclusions upon it.

References

[1] https://www.coursera.org/learn/machine-learning/, Last accessed 28.04.2016