BLG 454E Learning From Data (2019)
Term Project Report

Mehmet Yakuphan Bilgiç Sabrina Cara Aleksja Braka

***Abstract*—Autism Spectrum Disorder is a developmental disorder displayed as restrictive, repetitive, verbal and non-verbal behavioral patterns and persistence on following the same routine.In this project, we have hypothesized that these behaviours are an indicator of ADS affecting the similarity in morphology between brain reasons.Thus, we have made a classification project which consists of taking these different features, training a successful model using machine learning methods and algorithms which will result in high accuracy.**

I. INTRODUCTION

In this project, we are training a model in order to determine whether the people taken under examination will show ASD behaviours or not by taking into consideration the euclidean distance between the shapes of their two brain regions, thus making a classification of autistic vs. non autistic. Our kaggle competition information is given as follows:

Group members names: Mehmet Yakuphan Bilgiç, Aleksja Braka, Sabrina Cara. Group name: 150170703_150160916_150160914
Accuracy: 0.65, Rank:36

II. DATA SET USED

We used sklearn.preprocessing RobustScaler method which is a method that scales the data features while also being robust to outliers. According to scikit learn 0.21.2 documentation, RobustScaler method standartizes the data by removing the median from all points, therefore it gives better results when used in data with outliers.[1] Afterwards, we used PCA method in order to perform feature selection and reduce the number of components used.

III. METHODS

In this project we tried different methods including SVM, PCA, Random Forest Tree. Among all these, we found that the model which achieved highest accuracy was SVM (Support Vector Machine) with RBF kernel. Support Vector Machine is a supervised learning method widely used for classification since it is efficient in high dimensional spaces.[2] During our research about SVM we realized there are a lot of parameter tuning needed to be done in order to achieve optimal parameters which will, in turn result in better accuracy. From the code we wrote in order to find optimal parameters, we found that the highest accuracy is achieved when we tune our parameters as such: **C = 100 ; gamma = 'auto' ; kernel 'rbf'**. Also, as mentioned above, we used PCA so as to reduce the number of features used in the testing set. We achieved our highest accuracy when reducing the number of features down to 40, as illustrated in Fig.1.

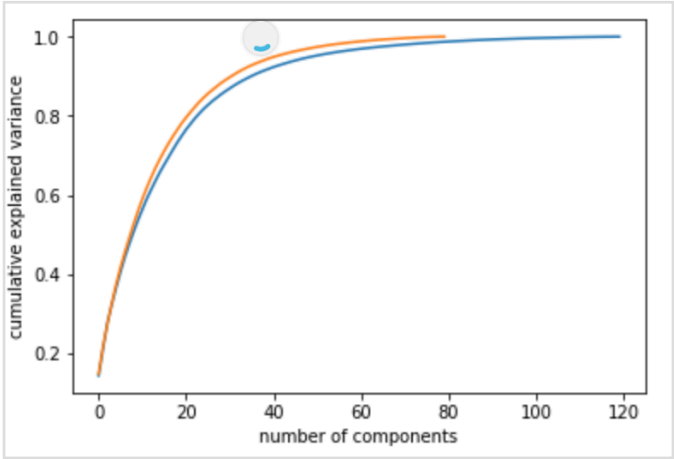

Figure1. Number of components - Cumulative explained variance in PCA

IV. RESULTS & CONCLUSIONS

In conclusion, with our methods used for feature reduction and model training , we achieved the highest accuracy of 65% at parameters explained above, which ranked us 36/57. This is not the best accuracy that we could have achieved. If given more time, or if given other projects as such, another method would have been which would train the data better and avoid overfitting it, as well as tune the parameters better in order to find the most optimal ones, and get relatively higher accuracy. Taking into consideration the fact that ASD is a serious issue, models with a higher accuracy that 65% need to be achieved in order to reduce the number of wrong cases or diagnosis to the maximum point we can.

REFERENCES

1. scikit-learn.org, 'sklearn.preprocessing.RobustScaler', 2007. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html [Accessed: 30- May- 2019].

2. scikit-learn.org, 'Support Vector Machines', 2007. [Online]. Available: https://scikit-learn.org/stable/modules/svm.html [Accessed: 30- May- 2019].