BLG 454E Learning From Data

Term Project

Students

Melik Mehmet Bıyık - 150160534
Fatih Kocabaş - 150160539
Furkan Uzun - 150170501

Kaggle Team Name

150160534_150160539_150170501

Contents

1.Introduction

The project, which was selected as an end-of-term project, was applied as a classification competition for the mentally handicapped people who are the subject of the health field from kaggle.com.

Autism, a subspace of mental disability, is a disease that has millions of patients worldwide. There is an important mass to address this disease. Therefore, classifying data related to this disease is of great importance.

Generally, there may be some classification problems due to widespread disease in the world. This difficulty can be examined as the classification of more than 500 features of autistic people and normal people for millions of people.

Our Kaggle team name is **150160534_150160539_150170501** and final score is 60,000% accuracy. We have achieved

2. Datasets

- **NumPy (Numerical Python)** is a math library that enables us to perform scientific calculations faster than normal.

- The **Sklearn** library is a library that enables us to do machine learning.

- **Pandas**, Numpy's column names and the inability to work with non-homogeneous data such as missing issues and produces more solutions on these issues.

- **Matplotlib** graphics drawing package is one of the most important tools of scientific programming with Python. Compared to other packages, Matplotlib is a very powerful package and we can interactively visualize data.

- **Perceptron** is the function that enables us to make predictions within the project.

- PCA(Principal Component Analysis) is a machine learning algorithm to speed up the function and allows us to make the size of the project smaller than 595.

- Within the project, we used **csv library** to read csv files given to students

3. Methods

Firstly, Again we have made reading process for **“train”** dataset. After this fundamental process, we have divided our data (which we used it with an array in numpy methods) into 2 parts. Dividing process made with **“for loop”** When we divide the array with 596 and then use remaining parts, these are our values. Then we named these parts **“Sonuc”** and **“x”**. X ‘s are our features actually (columns in csv file). After this process we use **reshape** function from numpy library. By using this reshape function we transformed our one-dimensional data (with the length of 595x120) to a table format with 120 rows and 595 columns. Actually we divided our X attribute and then write it into X again. We have transformed shape of X as a result. We have **append**ed (which is from numpy again) **“Data”** to **“X”** by using **data[i]** in **for loop** and there is no **“Data”** from now on. **numpy.append’s** first parameter gives us the list that will be add and second parameter of **numpy.append** gives us the element that we added. Then we created a data frame with named **“Dataset”**. We used **data frame** because our machine learning functions working with data frames. Actually data frame is an organized array like a table and our data frame is empty at that moment. We have an array named **“sonuc”** (single-dimensional) and **“dataset”** in type of frame(we have changed type of dataset from array to frame). We have 595 features and we cannot use it in learning algorithms yet. In this part of project we decided to use **feature extraction methods**. From the feature extractions methods we decided to use **PCA (Principal Component Analysis)** method which is suitable when input dimension is high. We have defined **“scaler”** from sklearn (which is library of machine learning). We have imported **StandartScaler()**. These are all preprocessing issues before start to learning processes. We transformed the **“dataset”** into **“scaled_data”**. Then we select **n_component** of PCA as 20. As a result of this issue we have smalled our data with using PCA method. We have only **“x_pca”** now.

Again we have made reading process for **“test”** dataset like we do **“train”** dataset at the beginnig of project. There is no **“sonuc”** column in test dataset. We used **“test_x”** attribute to define. Then we reshaped it with 80 rows and 595 columns. Our main aim to find out 80 results. Then we used data frame again and we used PCA method again like we do it for **“train”** dataset chapter.

Lastly, in learning part of project we have **iteration number** and **eta0**. We use **Perceptron function** in learning part of project. Our main aim is classification, we have a dataset and we need to classify them. Perceptron is an exact point that classifies all datas. After that, we have **fit**ted all learned informations to **ppn**. Then we ask for prediction by using **y_pred = ppn.predict()** . Then we created a **csv** file by using **y_pred**.

4. Results and Conclusions

According to the output from the program written in the project by using the methods of machine learning applied in the world as well as all the work in the field of health.

In the way of the scientific data obtained, we observed the observations and the ability of learning machines to improve during the improvement stages and we explained these paths in details in our report.

Our project, which is the beginning stage, has been a good start in the field of machine learning that everyone dreamed of.

As a result of the feedback received from the group members, we have come to the conclusion that the machine learning is challenging but enjoyable as well as satisfying.

5- References

[1] Alpaydin, Ethem (2010). Introduction to Machine Learning. London: The MIT Press.

[2] Principal Component Analysis - https://towardsdatascience.com/

[3] Char, D. S.; Shah, N. H.; Magnus, D. (2018). "Implementing Machine Learning in Health Care, New England Journal of Medicine.