**Homework 3**

**RANDOMFOREST CLASSIFIER FOR FACEALL DATASET**

Tuhena Sen

Mar 25, 2020

**Class:** ALY 6020 – Predictive Analytics

**Professor:** Dr. Marco Montes de Oca

Why it is called random forest? Well, because the forest has trees and in machine learning world means a decision tree. A decision tree is a tool that uses tree model of decision to obtain possible cases. It is a way to display an algorithm that contains conditional control statements. So, in random forest the model creates multiple tress from random sampling hence it is called random sampling and from those sampling decision trees are formed which come up with a decision and the majority of the decision made by various decision trees is the decision of the random forest.

In this assignment I will try 10,50,100 and 500 trees with each having 1, 10 or 20 minimum samples which will therefore be 12 models at the end and will check the accuracy of all the 12 model and finally will compare it with KNN classifier which I have built earlier.

ANALYSIS

Step1 Loading the dataset and loading required libraries and packages.

Here I have loaded the given faceall dataset which has faces of 14 grad students. Each facial outline is mapped onto a one-dimensional series.

Moreover, I have imported randomforestclassifier from sklearn.ensemble.

Table

Description automatically generated

Step2: Preparing the data

Graphical user interface, text, application

Description automatically generated

Creating train x and y and test x and y data. For that I have separated 1st column which has numbering of students 1 to 14 and rest train\_x contains all the dimensions.

Step3: Fitting the model – case1(n\_estimators =10, min\_sample\_leaf=1)

Graphical user interface, text, application

Description automatically generated

Similarly ran all the cases with following accuracy:

|  |  |  |
| --- | --- | --- |
| N\_estimators | Min\_sample\_leaf | Accuracy in percentage |
| 1. 10 | 1 | 60.8 |
| 1. 10 | 10 | 62.2 |
| 1. 10 | 20 | 57.6 |
| 1. 50 | 1 | 80.6 |
| 1. 50 | 10 | 76.0 |
| 1. 50 | 20 | 61.7 |
| 1. 100 | 1 | 76.7 |
| 1. 100 | 10 | 76.9 |
| 1. 100 | 20 | 71.5 |
| 1. 500 | 1 | **82.7** |
| 1. 500 | 10 | 77.1 |
| 1. 500 | 20 | 68.4 |

Out of all the models model with n\_estimators=500 and min\_samples\_leaf=1 has the highest accuracy = 82.7%

Moreover, if compared with KNN model KNN which has following accuracy , I can see that none of knn models has more than 82.7 % accuracy hence I can conclude that random forest is better model than knn for faceall dataset.

|  |  |  |
| --- | --- | --- |
| p | k | accuracy |
| 0.5 | 3 | 68% |
| 0.5 | 5 | 66% |
| 0.5 | 7 | 64% |
| 1 | 3 | 64% |
| 1 | 5 | 62% |
| 1 | 7 | 59% |
| 2 | 3 | 33% |
| 2 | 5 | 30% |
| 2 | 7 | 28% |

References

1. Will(2018). Hyperparameter tuning the randomforest in python. Retrieved from <https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74>
2. Random Forest. <https://www.youtube.com/watch?v=ok2s1vV9XW0&t=600s>