Tsiounis Konstantinos, 2630

Polyzos Alexios, 2338

Patterns Recognition

2nd Exercises Set

# Introduction

We used Jupyter Notebook to build this project. If you don’t already have jupyter notebook

or anaconda on your computer, we suggest you to install [Anaconda](https://www.anaconda.com/distribution/) which includes

everything you need to run our notebooks. For this project, we used Scikit-Learn Library for

the models, Pandas Library and NumPy.

About our code, you have it in a zip but you can also see it on clicking [here](https://github.com/ktsiounis/DataClusteringAlgorithms). If you can

download it or run for some reason, you can open the html files in our GitHub repository to see these notebooks with their results.

# Purity Score

After some research we did about purity score, we found that purity can be calculated using confusion matrix. So, to calculate purity, we first created a confusion matrix using sklearn’s metric for clusters and we calculated the sum of the max values in each row divided by the sum of all the values. To do this, we created a function as you can see below.



# F1 Score

The f1 score is calculated using sklearn’s function from metrics package.

# K-Means Clustering

The data loading and pre-processing are the same with the previous exercises set so we won’t repeat this section. After these steps, we created a k-means classifier from SciKitLearn library using 2 clusters, we split our data into training and test data and we fit the model with the training data.





The last step is to compute purity and f1 score.



The above steps repeated for the spambase dataset.

# Agglomerative Hierarchical Clustering

For the Agglomerative Hierarchical Clustering, we created an AgglomerativeClustering object with 2 clusters and linkage parameter equal to average. The other steps were the same with the other models.



# Spectral Clustering

For the Spectral Clustering, we created a SpectralClustering object with 2 clusters. The other steps were the same with the other models.



However, because of the datasets size we couldn’t complete the clustering. As we found after research, Spectral Clustering is difficult to run with big arrays of data. We also tried to run on cloud but we couldn’t reach to an end. So, below at the results, the results for spectral will be empty.

# Results

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 2 Clusters | | 4 Clusters | | 8 Clusters | |
|  | Purity | F1 | Purity | F1 | Purity | F1 |
| Occupancy Detection | 0.946 | 0.047 | 0.936 | 0.023 | 0.971 | 0.038 |
| Spam Dataset | 0.647 | 0.512 | 0.667 | 0.595 | 0.706 | 0.544 |

Table . K-Means Results

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 2 Clusters | | 4 Clusters | | 8 Clusters | |
|  | Purity | F1 | Purity | F1 | Purity | F1 |
| Occupancy Detection | 0.805 | 0.201 | 0.978 | 0.207 | 0.978 | 0.034 |
| Spam Dataset | 0.617 | 0.553 | 0.633 | 0.366 | 0.633 | 0.618 |

Table . Agglomerative Hierarchical Clustering Results

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 2 Clusters | | 4 Clusters | | 8 Clusters | |
|  | Purity | F1 | Purity | F1 | Purity | F1 |
| Occupancy Detection | - | - | - | - | - | - |
| Spam Dataset | - | - | - | - | - | - |

Table . Spectral Clustering Results