DMG Assignment 3

Aryan Behal 2019026

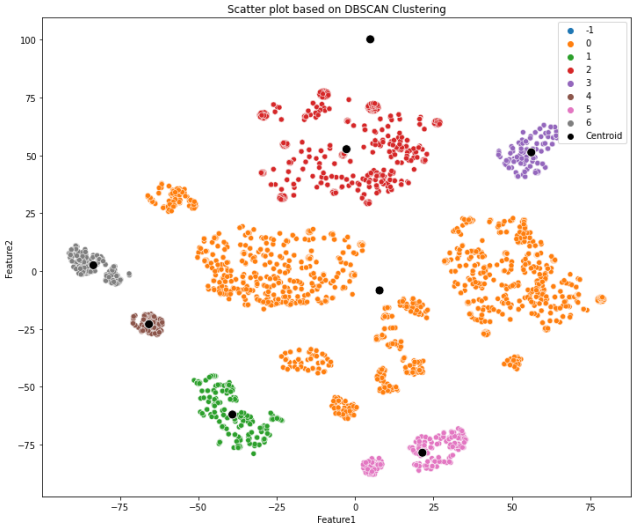
Kesar Shrivastava 2019051

PRE-PROCESSING

- We look at the structure of the data and find that there are many categorical attributes that should be converted into numerical data to perform further analysis.
- For preprocessing we use a label encoder.
- We do label encoding for attributes like elevation, aspect, slope, wilderness, soil type, hillshades, and different distances.
- We checked for null values. There are 0 null values in the dataset.
- We reduce the dimensionality for visualization purposes.

QUESTION 1

- The three clusterings performed are K-means clustering, DBSCAN clustering, and Agglomerative.
- The centroids for each cluster is shown in the plots using a black dot
- For visualization, we have reduced the dimensions to two.



PROCEDURE

- For every clustering technique, we have imported the necessary libraries.
- We then train the model over the dataset to get clusters.
- After we get the clusters, we plot the clusters using the matplot library.

QUESTION 1 (C)

- For every clustering technique, we have printed the number of records under each cluster in increasing order from 0 → 6
- We have also printed the number of records under each ‘target.
- A comparison can be drawn between the different lists

QUESTION 1 (D)

- Adjusted Mutual Information (AMI) is an adjustment of the Mutual Information (MI) score to account for the chance. It accounts for the fact that the MI is generally higher for two clustering techniques with a larger number of clusters, regardless of whether there is actually more information shared.
- This metric is independent of the absolute values of the labels: a permutation of the class or cluster label values won’t change the score value in any way.
- We have shown the AMI scores between Gaussian and all the 3 clustering techniques.
- Scores are as follow:
    - K-mean and Gaussian: 0.838397917001275
    - Dbscan and Gaussian: 0.5298783316173058
    - Agglomerative and Gaussian: 0.8528590658622243

QUESTION 2

PRE-PROCESSING

The pre-processing steps are the same as question 1.

STEPS

- For prediction, we have used K-means clustering.
- We see the elbow plot of the dataset to get the optimal number of clusters.
- The elbow appears at 5 hence we go for five clusters.
- While implementing and training the model we used the test_train_split method to train our model and analyze its accuracy and F1 score.
- The model is trained on the default parameters except for the fact that we create 5 clusters.
- After we train the model, we map the clusters to the true labels.
- The indices of the clusters are different from those of the predicted labels so mapping is required.
- The F1 score that we got on the test set (obtained from splitting the original dataset) is 0.56

REFERENCES

https://www.analyticsvidhya.com/blog/2021/06/k-means-clustering-and-transfer-learning-for-image-classification/
https://machinelearningmastery.com/clustering-algorithms-with-python/
https://www.kaggle.com/satyajitmaitra/detailed-clustering-algorithems
https://www.kaggle.com/shrutimechlearn/step-by-step-kmeans-explained-in-detail
https://towardsdatascience.com/gaussian-mixture-models-d13a5e915c8e
https://www.analyticsvidhya.com/blog/2019/10/gaussian-mixture-models-clustering/