CDS503: Machine Learning

Assignment 2: Clustering

Problem: In this assignment, you are to work with the countries dataset (countries_geocodes.csv) containing three attributes described below. Given the longitude and latitude of each country, we want to group the countries into continents.

Attribute	Description
longitude	Geographic coordinate that specifies the east–west position of a point on the Earth's surface
latitude	Geographic coordinate that specifies the north–south position of a point on the Earth's surface
country_name	Country name

Your clusters should be formed based on only **2 attributes**: **longitude** and **latitude**.

Part 1: K-Means Clustering (25 marks)

Use the K-Means cluster algorithm to find clusters representing continents.

a) Experiment with some different sizes of k and observe the range of the Sum of Squares Error (SSE) (see Appendix for more details on SSE). What k value would you pick to best cluster the countries into continents? Briefly justify why you select the k value.
b) Report the final parameters you set including the value of k selected to obtain your final clusters. Also, report the coordinates of the centroids, sum of squared differences to centroids and the number of iterations from the best clustering you have found. Explain in one sentence why you think this is your best clustering.
c) Name the continent each cluster represents in the table below. Describe each cluster according to the centroid values of each attribute. For each cluster, be sure to report the attribute centroid in terms of the original attribute values. Also, you can visualize the clusters on a scatter plot to help you describe and identify the continent represented by each cluster. You can also concatenate the cluster labels with longitude, latitude and country names to analyze the countries in each cluster.

Cluster	Centroid	Continent
0		

Part 2: Hierarchical Clustering (25 marks)

Use the hierarchical clustering algorithm to find clusters representing continents.

a) Use the AgglomerativeClustering algorithm. Experiment with different number of clusters (n_clusters) and other parameters (e.g., affinity, linkage, etc.) to get the best clusters to represent continents. Use dendrogram and scatter plot to help you visualize the clusters. Report the best parameters, the number of clusters you have selected. Explain in one sentence why you think this is your best clustering.

b) Name the continent each cluster represent in the table below. To analyze each cluster, observe the data points in each cluster on the scatter plot or look at what country names are in the clusters.

Cluster	Continent
0	

Part 3: Identify the best continent clustering (50 marks)

Based on the best clusters obtained respectively from K-means and AgglomerativeClustering, choose ONE algorithm that would give you the most accurate grouping of countries into the correct continent (final_cluster). Marks will be based on the number of your continent cluster labels matching the actual continent labels.

a) Based on the algorithm you have selected, provide a final mapping of the cluster number to the continent name in the following table.

Cluster	Continent
0	

b) Create a new data frame to store longitude, latitude, country_name and final_cluster. The final_cluster column will contain the cluster labels from the best clusters you have selected.

```
# Create data frame with longitude, latitude,
# country_name and cluster label
# Parameter x is the data frame containing longitude,
# latitude and country_name
# Assuming y_kmeans produces the best cluster labels
# Command should written in one line
final_cluster = pd.concat([x, pd.DataFrame(y_kmeans, columns =
['cluster'])], axis = 1)
```

c) Save final_cluster data frame into “countries_clusters.csv”. Make sure you save your data in csv format.

```
# Save and export final_cluster data frame into csv file
# Command should be written in one line
final_cluster.to_csv(r'countries_clusters.csv', index = False,
header = True)
```

Submission Instructions

Assignment 2 is an **individual** assignment. Submit the following items in the assignment dropbox on elearn latest by **11.59 p.m.** on **Friday**, **18 June 2021**. Please make sure you include your full name and matric number on your report.

a) Assignment report (maximum 5-page Word document file)
b) Data file with the best cluster labels: countries_clusters.csv
c) Jupyter notebook containing all your codes (.ipynb file)

Part 1: Description of K-Means Clustering experiments (25 marks)

Part 2: Description of Hierarchical Clustering experiments (25 marks)

Part 3: Accuracy of the best clustering selected in which each country in your continent cluster will be compared to the actual continent of the country (50 marks)

Appendix

Sum of Squares Error (SSE)

We can use Sum of Squares Error (SSE) to measure the variation within a cluster. SSE is defined as the sum of the squared distance between each member of the cluster and its centroid. The smaller the SSE, the closer the distances of each member to its centroid. Based on the two figures below, SSE for Figure 1 is larger than Figure 2 (Figure 2 produces better clusters than Figure 1).

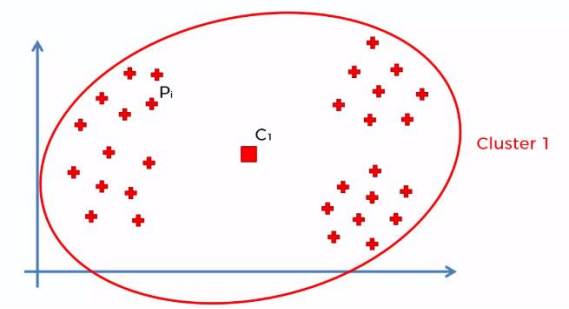

Figure 1

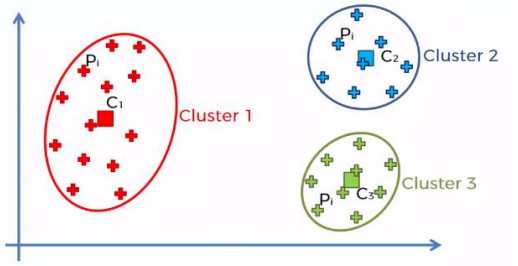

Figure 2

To view SSE in python, fit kmeans on the data and then use the following command:

```
print("Sum of squared distances to centroids", kmeans.inertia_)
```