Geo-economic Clustering

Introduction

In this task, we will group most of the countries in the world in 10 different groups by means of the distance among them and the difference in their Gross Domestic Product (GDP) per capita. For that, we need two datasets, one containing the gps coordinates of the countries and other containing their gdp information. The first one is available from Kaggle at https://www.kaggle.com/eidanch/counties-geographic-coordinates which is a copy of the data set published by Google at https://developers.google.com/public-data/docs/canonical/countries_csv. After checking the coordinates in Google Maps, it can be seen that they correspond to the geographical center of the countries. The second one is available from Databank at https://databank.worldbank.org/reports.aspx?source=2&type=metadata&series=NY.GDP.PCAP.CD#, and both the GDP per capita and the GDP correspond to the year 2019, in US$.

Clustering

Since the goal of the task is to classify the countries in 10 different groups, we have decided to use clustering techniques. In particular, we will use the distance among the countries and the center of the clusters and assign the countries to their closest cluster. Mathematically, we need to define the next matrices. Denoting the number of countries by n we have:

- **C**: matrix of dimension n x 3 with the information of the countries: latitude (gps), longitude (gps) and gdp per capita.
- **X**: matrix of dimension 10 x 3 with the information of the clusters: latitude (gps), longitude (gps) and gdp per capita.

Thus, the set of countries is defined as $\{\mathbf{c}_1,\ldots,\mathbf{c}_n\}$ where each country $\mathbf{c}_i$ is a vector with 3 components, and the set of clusters is defined as $\{\mathbf{x}_1,\ldots,\mathbf{x}_n\}$ where each cluster $\mathbf{x}_j$ is a vector with 3 components. $\mathbf{d}_{ij}$ is defined as the distance between country **i** and cluster **j**, and computed as the Euclidean distance between the two vectors. The physical distance (dist) between two points on a sphere can be computed using the Haversine formula, for which gps coordinates will be used. Later, $\mathbf{d}_{ij}$ will be computed as: $\mathbf{d}_{ij} = \text{sqrt} (\text{dist}^2 + \text{gdp}_{\text{difference}}^2)$.

Implementation

The code file includes comments on almost every line to explain all the steps. To classify the countries within the clusters we need to define a cost function, that will be the within-cluster sum of squares. It is the sum of the distance of each country and its corresponding cluster. The objective is to minimize this sum, so we are solving an optimization problem. The steps to implement the algorithm are:

1. Define random centers for the 10 clusters.
2. Assign each country to its closest cluster by computing the distance to all the cluster and selecting the smallest one.
3. Update the center of the clusters by averaging the data of their countries.
4. Repeat step 2 and 3 until convergence.

Results and Conclusions

Below we present a table containing the clusters and their countries and one plot with the average GDP of each cluster.

From a geo-economic point of view, we can see that the algorithm has found a good solution. In general, countries that are relatively close in physical distance and in GDP per capita belong to the same cluster. For example, Cluster 7 has the highest average GDP per capita and we can find countries that are not very far away. Although USA is a bit far in physical distance from the rest of the countries, it is quite similar in terms of GDP per capita. Cluster 2 has the lowest average GDP per capita and we can find countries that are extremely poor in terms of economics and that are quite close in terms of physical distance, since all of them are located in the south of Africa. Cluster 5 also has a low GDP per capita but it is located at the north part of Africa. Another example is Cluster 8, that contains many of the healthiest countries in Europe. Poorer countries in Europe are found in Cluster 6 together with some middle east countries. Cluster 1 contains the countries of South America, whose GDPs per capita are not quite different. Perhaps, the biggest differences can be found in Cluster 0 which contains countries located around the west Pacific. However, some of them are a bit far from each other and there are important differences in terms of GDP per capita for some of them. For instance, the GDP per capita of Australia is 54907 US$ meanwhile the GDP per capita of Tuvalu is just 4049 US$. However, we have to keep in mind that there are many solutions, and this is just the one that we have obtained after randomly initializing the center of the clusters. Thus, we will always have some discrepancies with the results.

Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9
Japan	Brazil	South Africa	Mexico	Mauritius	Nigeria	Saudi Arabia	United States	Germany	China
South Korea	Argentina	Ethiopia	Colombia	Maldives	Algeria	Turkey	Switzerland	United Kingdo	India
Australia	Chile	Kenya	Ecuador	Seychelles	Morocco	Poland	Norway	France	Russia
Indonesia	Peru	Angola	Puerto Rico		Ghana	Egypt	Ireland	Italy	Thailand
Singapore	Uruguay	Tanzania	Dominican Republic		Tunisia	Romania	Denmark	Canada	Philippines
Hong Kong	Bolivia	Congo [DRC]	Guatemala		Cameroon	Czech Republi	Qatar	Spain	Bangladesh
Malaysia	Paraguay	Uganda	Panama		Senegal	Iraq	Luxembourg	Netherlands	Pakistan
New Zealand		Zambia	Costa Rica		Mali	Greece	Iceland	Sweden	Vietnam
Papua New Guinea		Zimbabwe	El Salvador		Gabon	Kazakhstan		Belgium	Sri Lanka
Brunei		Sudan	Honduras		Burkina Faso	Hungary		Austria	Myanmar
Fiji		Botswana	Trinidad and Tobago		Benin	Ukraine		United Arab E	Uzbekistan
Timor-Leste		Mozambique	Jamaica		Guinea	Kuwait		Israel	Nepal
Solomon Islands		Madagascar	Bahamas		Niger	Slovakia		Finland	Cambodia
Vanuatu		Namibia	Nicaragua		Chad	Oman		Portugal	Afghanistan
Samoa		Congo [Reput	Haiti		Equatorial Gu	Bulgaria		Malta	Laos
Kiribati		Rwanda	Barbados		Mauritania	Belarus		Andorra	Mongolia
Nauru		Malawi	Guyana		Togo	Croatia			Kyrgyzstan
Tuvalu		Djibouti	Suriname		Sierra Leone	Lithuania			Tajikistan
		Burundi	Saint Lucia		Liberia	Slovenia			
		Lesotho	Belize		Gambia	Lebanon			
		Central Africa	Antigua and Barbuda		Guinea-Bissau	Libya			
		Comoros	Grenada			Serbia			
			Saint Kitts and Nevis			Azerbaijan			
			Saint Vincent and the Grenadines			Jordan			
			Dominica			Bahrain			
						Latvia			
						Estonia			
						Cyprus			
						Bosnia and Herzegovina			
						Georgia			
						Albania			
						Armenia			
						Macedonia			
						Moldova			
						Kosovo			
						Montenegro