"Clustering algorithms"

2nd Homework: A case study

A. Introduction

Main clustering stages (in practice):

1. **"Feeling the data":** In this stage one should perform various simple operations on the data, in order to by aware of their nature. More specifically, he/she should first **check individually each feature,** in terms of whether it is discrete or continuous-valued, the range of its values (minimum and maximum values,…), the way the feature values are distributed within this range(histogram, mean value, standard deviation,…) etc. Then he/she should **consider** the **features in groups** and check for possible correlations of different features. Of course, the possible **existence** of **missing data** should be investigated.
2. **Feature selection/transformation:** Based on the analysis of the previous stage, one should decide, (a) which features will employ to represent the entities involved in the current problem (in our case, the countries) and (b) if he/she will apply certain transformations on the features (e.g., if two features have significantly different range of values, they should be transformed, so that to have comparable ranges of values).
3. **Selection of the clustering algorithm(s):** Having in mind the general picture of the data, one should select the proper clustering algorithm(s) for them. For example, he should take into account the kind of the data (discrete-valued, continuous-valued features, …), the shape of the expected clusters (compact, elongated, ...), the possible existence of outliers etc.
4. **Execution of the clustering algorithm(s):** The selected algorithm(s) should run for various values of its parameters in order to determine possible persistent clusters, which are likely to be physical clusters.
5. **Characterization of the clusters:** Having determined the clusters from the previous stage, one should proceed with their characterization based on the values the features take for each cluster. Thus, in our case study, a cluster may characterized by high mortality and low GDPP.

B. Description of the case study

The **objective** of this study is to cluster the countries using socio-economic and health factors that determine the overall development of the country and to characterize each

resulting cluster (and, consequently, the countries it comprises) based on the relevant values of the above factors. More specifically, the available factors for each country are the following[1]:

1. **Child_mortality:** Death of children under 5 years of age per 1000 live births.
2. **Exports:** Exports of goods and services per capita. Given as %age of the GDP per capita.
3. **Health:** Total health spending per capita. Given as %age of GDP per capita.
4. **Imports:** Imports of goods and services per capita. Given as %age of the GDP per capita.
5. **Income:** Net income per person.
6. **Inflation:** The measurement of the annual growth rate of the Total GDP.
7. **Life_expectancy:** The average number of years a new born child would live if the current mortality patterns are to remain the same.
8. **Total_fertality:** The number of children that would be born to each woman if the current age-fertility rates remain the same.
9. **GDPP:** The GDP per capita (Calculated as the Total GDP divided by the total population).

The relevant data set consists of data from $\mathbf{167}$ ***countries*** and for each country there are available the above $\mathbf{9}$ ***factors***. The data are given in a $167x9$ matrix, called "***Countrydata***", where each row corresponds to a country and each column to a feature. In addition, the array $167x1$ column vector "***country***" contains the names of countries corresponding to the lines of the matrix "*Countrydata*". Both "***Countrydata***" and "***country***" are included in the file named "***data_country***".

Based on the above, you are asked to perform (at least) the following in the current case study.

"Feeling the data": (a) For each **single feature**, determine its kind, its range of values, its histogram, as well as its mean and its standard deviation. Comment very briefly on them. **(b)** *(i)* Compute the linear dependence of each feature with all the others, via the correlation coefficient. *(ii)* Perform 1) the ***standard score normalization*** and 2) the ***minmax feature scaling normalization*** on each feature and for each case compute the linear dependence between the transformed features.

Compare them with those found in (b)-(i) and comment on the results, giving supportive explanations for your statements.

[1] The data are from https://www.kaggle.com/rohan0301/unsupervised-learning-on-country-data.

Feature selection/transformation: (a) Decide which features will be selected for representing each country (one may choose all of them) and comment briefly on your choice. **(b)** Decide for the transformation (if any) you will need to apply on each feature. Comment briefly on your decisions.

Selection of the clustering algorithm: Choose an appropriate clustering algorithm and justify your choice. In the present framework, let the choice be among the cost function optimization algorithms.

Execution of the clustering algorithm: Execute the selected clustering algorithm for various choices of its parameters and (probably) with different initializations, in order to determine the physical clusters formed by the data vectors. It is suggested to consider not only the "best" clustering (according to some criterion) but also the second "best" clustering (if it is also "good enough"). Such a strategy may reveal clusters that are present in both clusterings, which is a very strong indication that these are physical ones.

Characterization of the clusters: Having determined the clusters from the previous step, you should characterize each one of them, based on the values of the features that are encountered among the vectors of each cluster. Quantities/tools like the mean, the standard deviation, the histogram can be used in this direction.

Next, compare the different clusters resulted from the above analysis, based on their characterization.

Clearly, after the application of the clustering procedure, each entity (country in our case) is characterized by the characteristics of the cluster where it belongs.

C. Implementation issues

In the current compressed file there are several .m files that implement the most well-known cost function optimization algorithms, along with some auxiliary .m files. For each of them, there exists an extensive description of what each such algorithm does.
CAUTION: In the above implementations of the algorithms, the **data vectors** are given as **columns** and not as rows.

1. Extract all the material of the current compressed file to a new folder in your computer.
2. Create a new .m file MATLAB and write in it all the instructions you need.
3. In order to load the data you should execute the following.

```
load data_country
```

D. <u>Useful MATLAB/Octave hints</u>

1. To get help for a MATLAB function, type "help" and the name of the function. It is strongly suggested to use it for the functions mentioned below, in order to understand better how they work.
2. The functions ***mean(X)***, ***std(X)***, compute the mean and the standard deviation, respectively, of the rows of a data matrix ***X***.
3. The function ***corrcoef(X)*** computes the correlation coefficient among the **columns** of a data matrix ***X*** (each column is considered as a feature and each row is considered as a data vector).
4. The function ***hist*** returns the histogram of the values of an array.
5. The function ***find(a)*** returns the **positions** of all the **nonzero entries** of ***a***. Writing e.g., ***find(a==1)***, we get the positions of the entries of ***a*** that are equal to **1**.
6. The matrices ***ones(m,n)*** and ***zeros(m,n)*** have all their elements equal to **1** and **0**, respectively, while the matrix ***eye(m,n)*** has all its entries equal to 0, apart from its diagonal ones, which are all equal to 1.
7. Let ***X*** be an *mxn* matrix and ***y*** an mx1 column vector. We wish to **subtract *y*** from **each column** vector of ***X***. This is done via the instruction ***X-y*ones(1,n)***.
8. Let ***X*** be an *mxn* matrix and ***y*** an 1xn row vector. We wish to **subtract *y*** from **each row** vector of ***X***. This is done via the instruction ***X-ones(m,1)*y***.
9. The functions ***plot*** and ***plot3*** create 2-d and 3-d plots, respectively.