Machine Learning 2020 - Milestone 1 - Unsupervised Learning

Francisco P. Romero - University of Castilla La Mancha 05/10/2020

The purpose of this work is to explore the environmental data collected by various U.S. Federal Government Agencies from two cities (San Juan, Puerto Rico and Iquitos, Peru) to gain a better understanding of the Denge Spread Phenomena.

These data are from a competition of the site DrivenData [1]. Training data will be used [2].

The overall objective is to use unsupervised learning techniques to make a preliminary exploration of the data and to extract conclusions from discarded elements, etc. The specific objectives are as follows:

1. Identification of outliers elements (weeks) in the dataset
2. Use clustering algorithms to identify groups and characterize them.
3. (optional) Feature Selection using clustering algorithms

Template

Student submissions must include a short report following the ACM template.

`https://www.acm.org/publications/proceedings-template`

Deadline: 2nd Nov

Tasks

The following steps could be executed in different order

Dimensionality Reduction

1. Extract the correlation among features and obtain conclusions.
2. Execute PCA and plot the results. Some conclusions are welcomed.

Outlier Identification

1. Find outliers in your data. DBSCAN uses distance and a minimum number of points per cluster to classify a point as an outlier.
2. Analyze why these elements are outliers and decide whether or not to consider them for further analysis.

[1] https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/

[2] https://s3.amazonaws.com/drivendata/data/44/public/dengue_features_train.csv

Clustering by K-means

Apply K-means algorithm to your data.

1. Don't forget to normalize your dataset (excluding the primary keys, of course) or use the PCA data.
2. Specity the chosen number of clusters (k) and a brief explanation about why you have chosen this value of k.
3. Execute k-means, test with different options of initialization (random, k-means++),
4. Try to assign a label to each group. Try to interpret the meaning of each cluster through its centroid.
5. The graphical result of your clustering (only one chart - PCA results- with elements represented by in different colours) must be included in your report.

Additional:

- remove outliers (step 4), select only a subset of features....
- You also can use the centroids as input of the hierarchical clustering algorithm.

Hierarchical Clustering Algorithm

- Compute the similarity matrix. Execute the hierarchical clustering algorithm. Test several cluster-distances-measures and choose the best solution in your opinion.
- Cut the dendrogram and characterize the obtained groups. Try to assign a label to each group. (Don't forget to read the feature descriptions in the competition page)
- Your best dendrogram and a brief explanation of your choices must be included in your report.
- The graphical result of your clustering (only one chart - PCA results- with elements represented by in different colours) must be included in your report.

Data Distribution

AA	San Juan 1990 - 1996	HSJ	Iquitos 2004 - 2010
EEE	San Juan 1997 - 2003	RS	San Juan 1992 - 1998
ER	San Juan 2004 - 2010	PM	San Juan 1992 - 1998
ELM	Iquitos 1990 - 1996	SSJ	Iquitos 1995 - 2000
D	Iquitos 2008 - 2010	others	Years: 2000 - 2004