Clustering

BY JUAN MANUEL ALONSO

Installation

```
$  pip3 install -r requirements.txt
```

Running

Custom hyperparameters in a textfile i.e. *"./*configs/*config.txt"*.

```
$  python3 experiments.py ./configs/config.txt
```

A *results* folder will contain a timestamp directory with the latest results.

Datasets

- Iris (http://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_iris.html)
- Breast Cancer (http://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic))
- sklearn Toy Dataset: Noisy circles (https://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html)

The first two datasets were explained and shown in the past experiment. The new toy dataset, *noisy circles,* is part of a comparison example by sklearn found on the link from above.

This dataset is composed of two circles in 2D, one containing the other. The number of samples that make them, the standard deviation of Gaussian noise added to the data, the random state to use and the scale factor between the inner and outer circles are all customizable parameters.

Techniques

Hierarchical: Agglomerative Clustering

This type of clustering technique creates a tree-like disposition of the data analyzed according to certain criteria. In this way, a dendrogram can be plotted to see explicitly the relations between the clusters formed, or *similarity*. This similarity is determined by the height of the lowest internal node they share[1].

A similarity computation, or *linkage*, is done between all data points, and the interactively among clusters until a given number of clusters is reached (*k)*. For this experiments, the *average* linkage was chosen, that is, the average of the distances of each observation of the two sets.

Partitional: K-Means

This is the other type of clustering techniques frequently used. It aims to directly obtain a single partition of the datapoint into clusters[2]. As opposed to the previous technique, here there is no hierarchy established. The datapoints or objects are placed directly into a given number (*k*) of clusters, which may be more according depending on the data structure.

The similarity here is done by calculating the distance of *k* representative elements with respect to centroids. The training consists of assigning datapoints to the centroids for each iteration, which are recalculated as the mean of each cluster. This is done until a convergence condition is met, usually defined as absence of change in clusters compared to a previous iteration.

Partitional: Gaussian Mixture

It is interesting to see with last technique, an example of a type of clustering called *Expectation Maximisation (EM),* how Gaussian distributions can fit the data. Firstly, *k* clusters are initialised. Secondly, an expectation step is made, in which each cluster is assigned points for each datapoint, using the conditional probabilities that a certain point belongs to a certain cluster. Lastly, a maximisation step is done, where the model parameters of the distributions are re-estimated. An iteration between the last two step is done until a given cutting condition.

---

[1] FH-Technikum_DataScience-SS_2020-L2.pdf slide 27

[2] FH-Technikum_DataScience-SS_2020-L2.pdf slide 59

Cluster Validity / Evaluation Criteria

External Measures: Entropy, F-Measure

Since the datasets worked with in these experiments is widely known and its clustering can be agreed on, entropy and f-measure are the chosen validity metrics.

Results

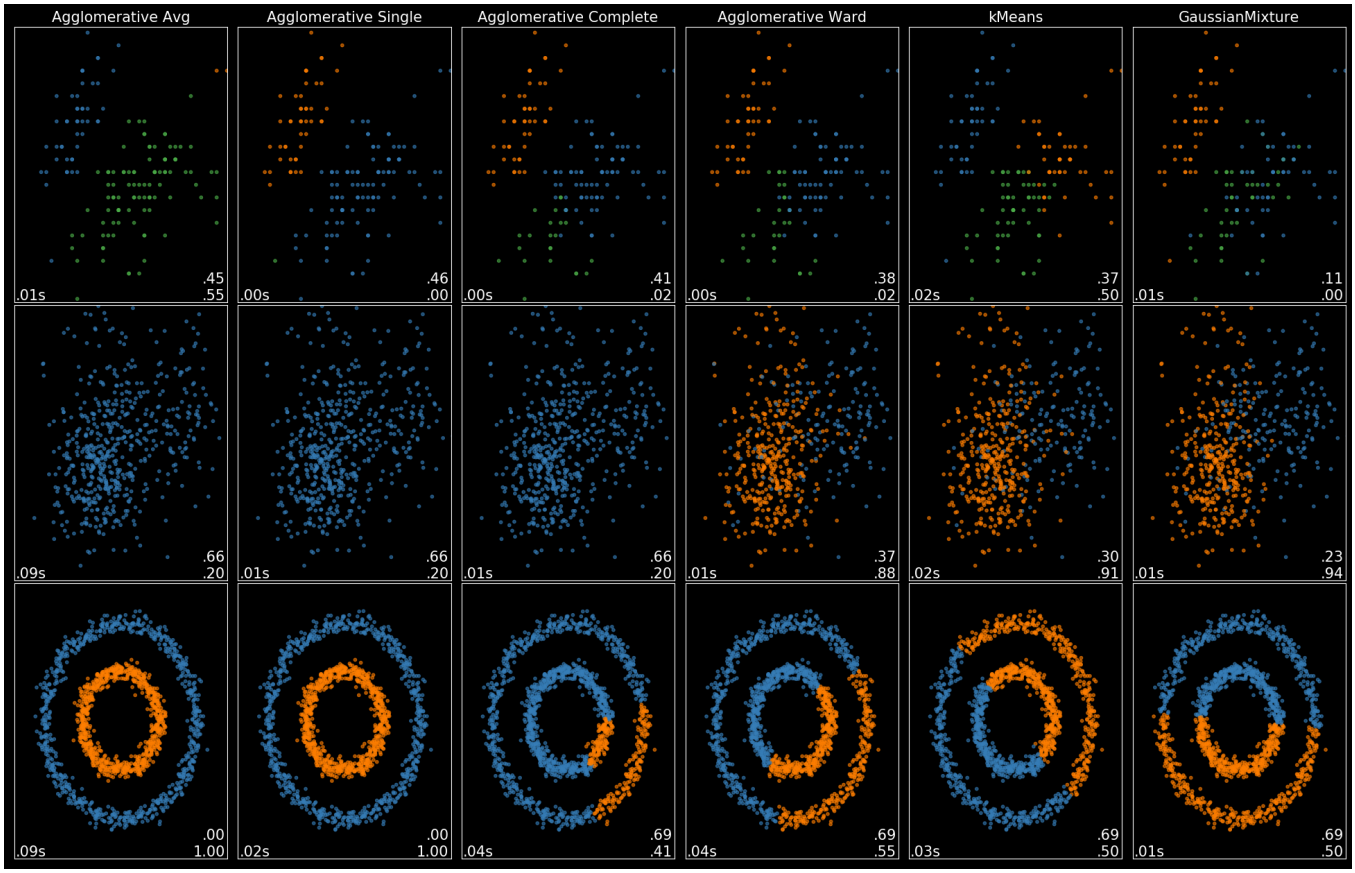

FIG. 1: PLOTS FOR CLUSTERING TECHNIQUES. IRIS, BREAST CANCER, NOISY CIRCLES. THE LOWER LEFT NUMBER IS THE TIME TAKEN TO EXECUTE SUCH ALGORITHM. ON THE OPPOSITE CORNER, THE UPPER NUMBER IS THE OVERALL ENTROPY OF THE CLUSTERS, AND THE LOWER NUMBER IF THE WEIGHTED F1 SCORE.

Based on Figure 1, a number of distinctions can be done from the plots. Firstly, for the Iris dataset (first row), the agglomerative techniques give different results, being the average linkage the one with the best F1 score. This may be due to the structure of the Iris dataset, that seems to favor average inter-cluster distances. On the other hand, the single linkage run results in the worst F1 score, which may make sense if once knows that 2 of the 3 classes in the real projection are rather intertwined, instead of grouped together in the same cluster, o separated by a clear-cut border. The kMeans has similar results in terms of F1 score, but less uncertainty in terms of entropy. Besides, its representation is somewhat closer to the real, known projection for Iris than the. The Gaussian Mixture technique ended up being the closest, with the least uncertainty, but with a poor F1 score, meaning it is a bad for finding true positives.

Secondly, for the Breast Cancer dataset (second row), the agglomerative techniques again give different results, being the ward linkage the only one with plausible plot. The ward linkage minimizes the variance of the clusters being merged, which implies that good-performing centroids are being calculated. The complete (farthest neighbor), single (nearest neighbor) and average linkages incorrectly, placed all data points at one unique cluster. Indeed, of the agglomerative techniques, *ward* was the best performing in both entropy and F1 score, but was outperformed by GaussianMixture and kMeans, in that order.

Lastly, for the noisy circles toy example from sklearn (last row), with the average and single plots yielding correct results in terms of F1 score and entropy. This is be due to the structure of the dataset, that favors average inter-cluster and nearest-neighbor distances. On the other hand, the kMeans and Gaussian Mixture techniques made poor divisions between the circles, which may make sense since they are not suitable to discover clusters with non-convex shapes. Besides, kMeans with different initialization might produce a different clustering.

Regarding time complexity, the agglomerative clustering with average linkage is the most computationally demanding in among all algorithms.

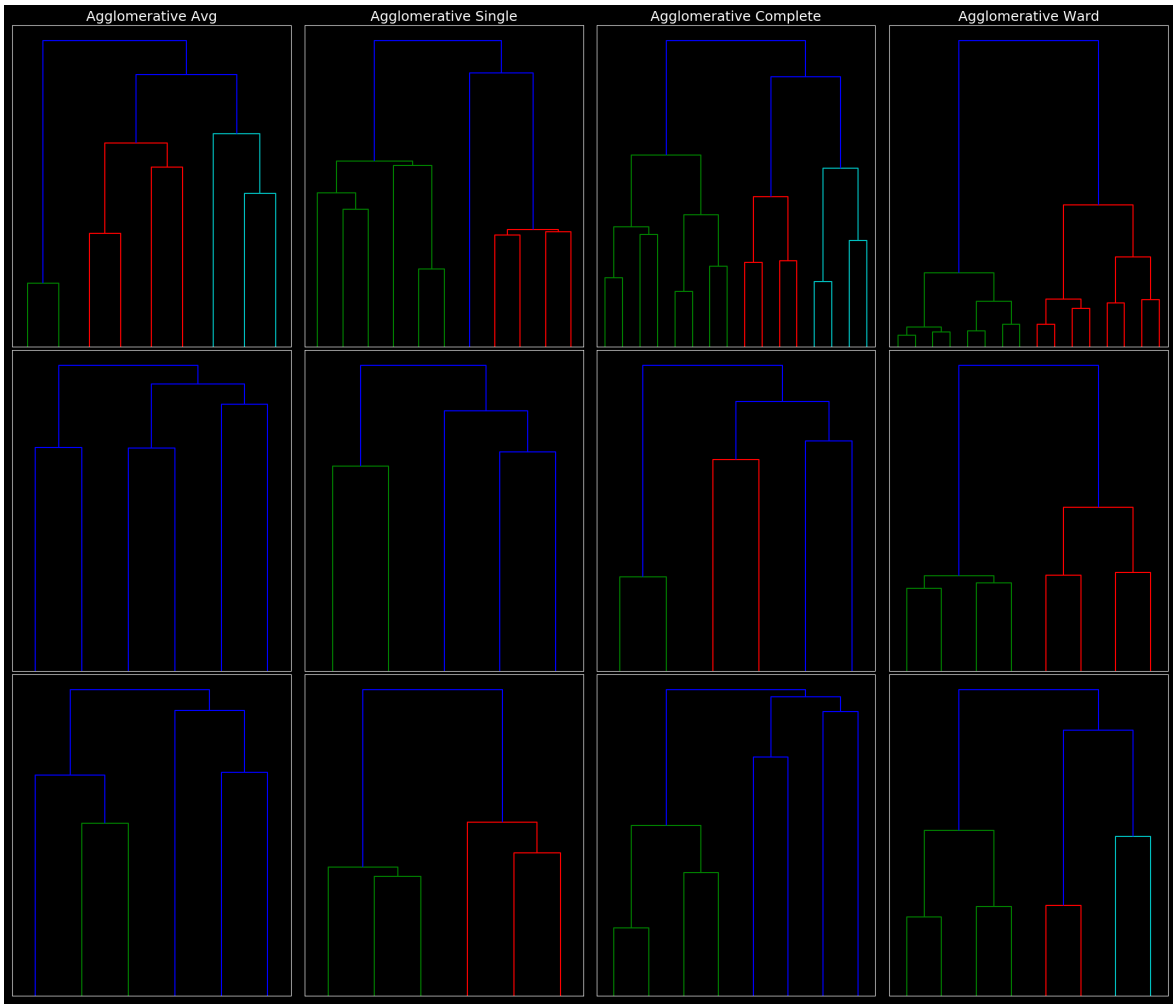

FIG. 2: DENDROGRAMS FOR AGGLOMERATIVE TECHNIQUES. IRIS, BREAST CANCER, NOISY CIRCLES.

Figure 2 shows the hierarchical tree-like decomposition of the datapoints from each dataset (same row order as Figure 1). For this figure the full tree was computed (distance_threshold=0, n_clusters=None). The heigh of a node represents the similarity of the two children clusters[3], which corresponds to the plots of Figure 1, more clearly for the noisy circles dataset. Each lead represents a certain number of points for that node. One can look at each dendrogram and estimate the number of clusters. Though rare, two highly separated subtrees are highly suggestive of two clusters[4].

[3] FH-Technikum_DataScience-SS_2020-L2.pdf slide 26

[4] FH-Technikum_DataScience-SS_2020-L2.pdf slide 93