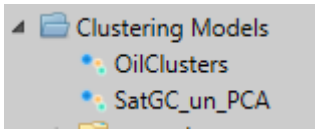
Patterns in the mind – Part 2



Introduction

Recently we have added k-means clustering to p:IGI+, to supplement the data exploration capabilities of the software. In the part 1 we considered what clustering is, how do we know when to use clustering and how can we identify how many clusters we have.

Now, in part 2 of the note, we look at a wider range of clustering methods and using a series of synthetic but useful data sets compare their strengths and weaknesses.

What is clustering?

Clustering is the act of finding similarly behaving groups of samples, based on a set of their characteristics (Everitt, 2011). The definition of similar is typically based on some form of metric, often a Euclidean (straight line) distance. A cluster model is just that, a model. All cluster models are wrong, some are useful. For this reason, there are many clustering methods used, and none is particularly better than any other overall (Estivill-Castro, 2002).

Alternative clustering methods

While HCA and k-means are two of the most widely used clustering methods, others are available, each with their own characteristics. Two additional useful options to consider are both classified as ‘density based clustering’ methods.

The first method is to fit a probability density function, for example based on a mixture of Gaussian distributions, to the data density. The assumption is that each distribution is a ‘cluster’ responsible for generating data. In principle any valid probability distribution could be used, but in general for multidimensional data Gaussian Mixture models are used. These can be efficiently fitted using an Expectation-Maximization (EM) algorithm which is actually very similar to the k-means algorithm. Then each point will be probabilistically (soft) assigned to each ‘cluster’, with the most probable cluster being selected if a hard assignment is desired.

Density based methods have the benefit of providing a soft assignment, can cope with overlapping clusters, and will work well in both the overlapping and the ‘blob’ case. It is also easy to assign a new point to an existing cluster, or update the clustering given new data.

Gaussian mixture model based clustering is a parametric clustering method – the model is defined by a small number of parameters which are estimated as part of the training (or ‘learning’). This is also a ‘generative’ model, in that once we have the model we can easily simulate data from it, which can be useful if we want think about the sorts of models that we might expect prior to seeing data.

An alternative to Gaussian mixture models is the non-parametric DBScan method (Ester et al, 1996). This is based on heuristics and looks for regions of ‘dense clusters’ of data. DBScan is unique in the methods thus far discussed that it allows for a ‘noise’ process and does not assign all points to a cluster. This can be very beneficial in cases where noise is an issue, although a noise model can be included in Gaussian mixture model clustering.

Comparing different methods in practice

To illustrate the differences in clustering methods, and the results they obtain on 2-dimensional data sets we adapted the Python scikit-learn toolkit (https://scikit-learn.org/) demonstration to focus on some key elements. These examples are illustrative and should be used only as guidance.

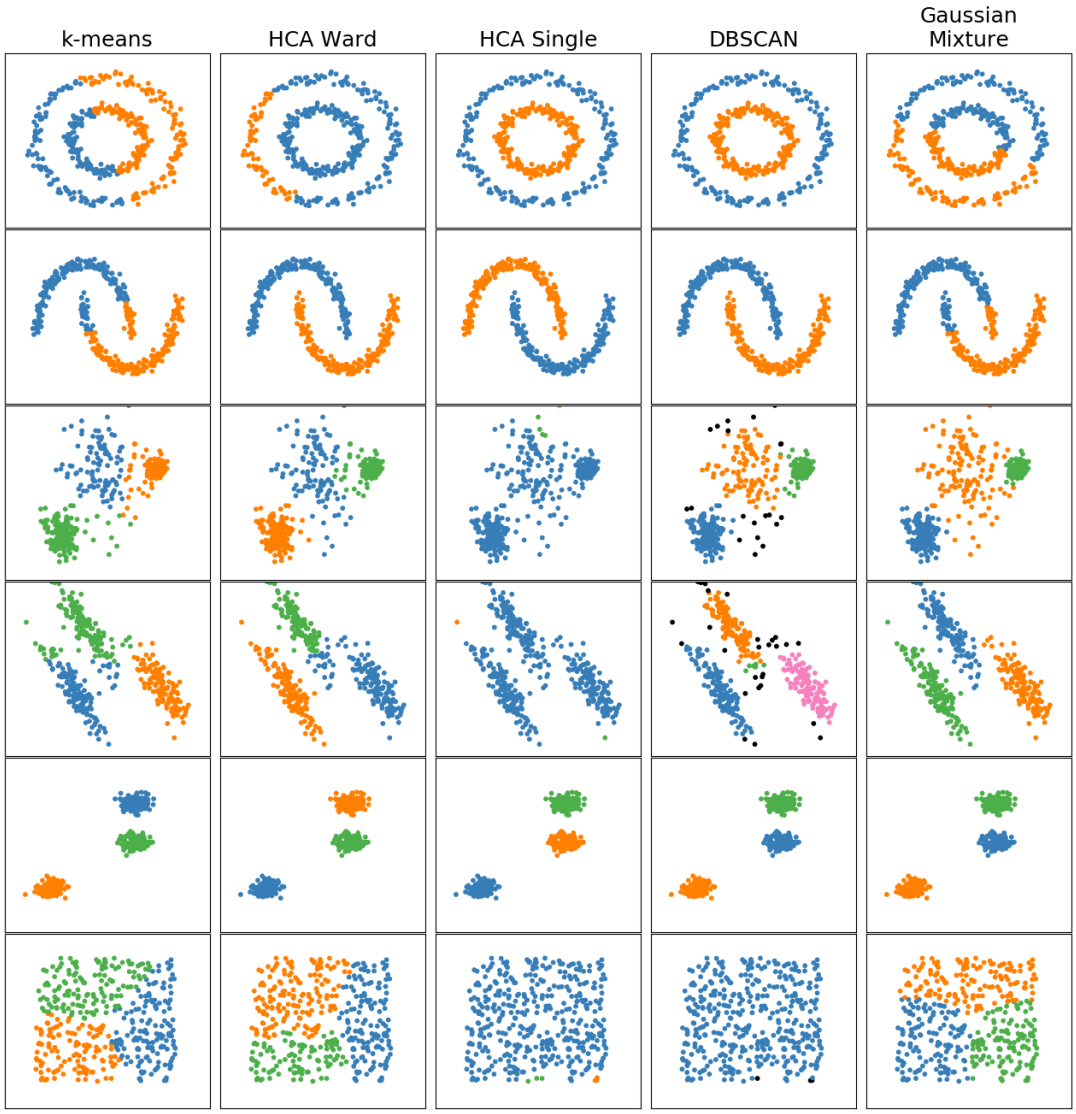
We used 6 different datasets:

- concentric circles – an example of a ‘topological cluster’
- ‘nu’ combination – a slightly different ‘topological cluster’
- overlapping 3 cluster example
- asymmetric 3 cluster example
- distinct 3 cluster example
- uniform random dataset

The intention of using these different datasets is to illustrate the application of a range of tuned algorithms (but not optimised) on these problems to demonstrate empirically their strengths and weaknesses. We used the scikit-learn implementation of the following algorithms:

- k-means – simple implementation
- HCA using the Euclidean distance and the Ward (variance) linkage
- HCA using Euclidean distance and the single (min distance) linkage
- DBScan
- Gaussian mixture model based with hard assignment, and full covariance

The results are shown below for datasets with 400 points.

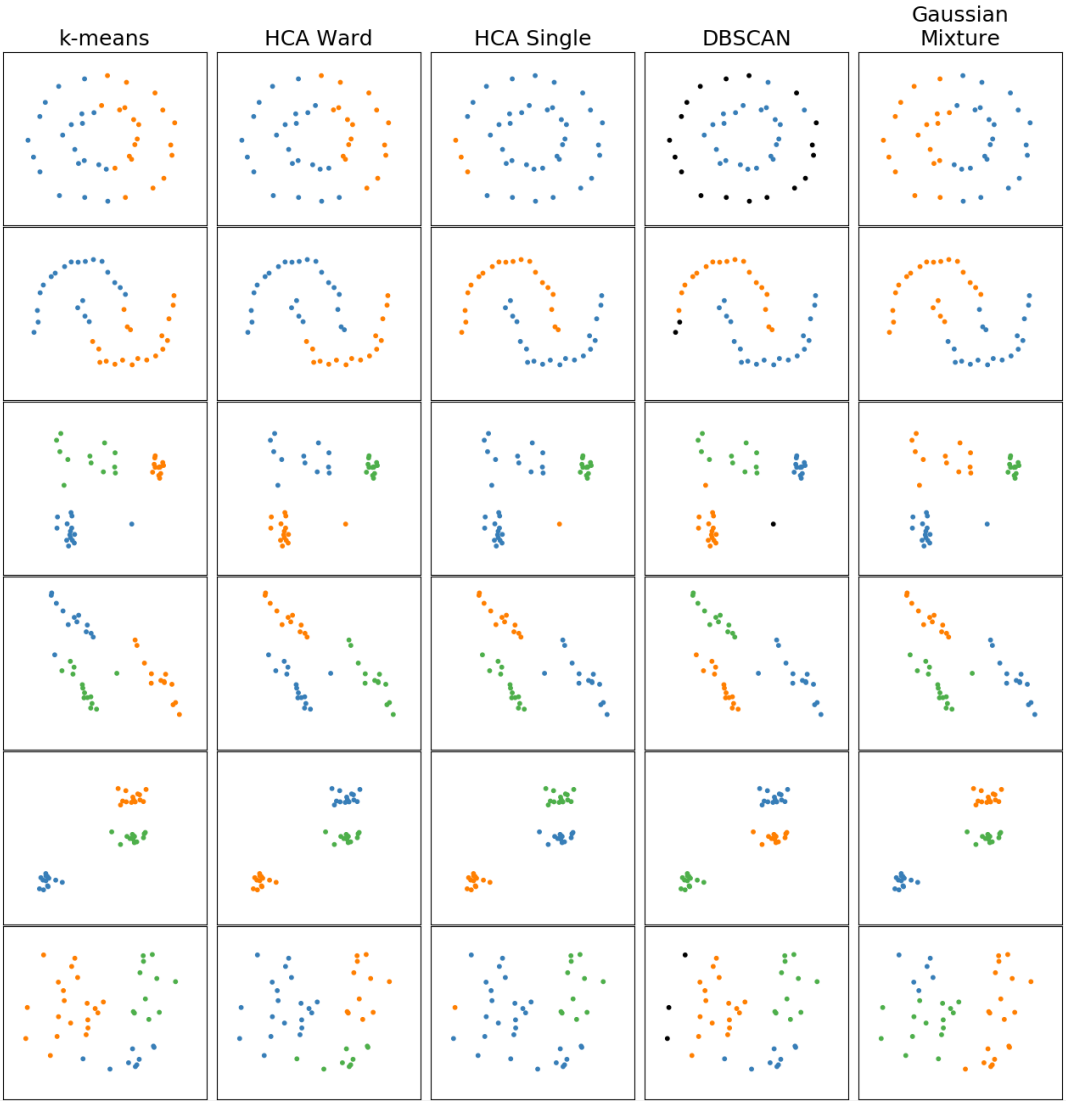


We did not conduct much in the way of algorithm tuning, but did select sensible parameters, and made some adjustments to the DBScan algorithm. What to take from the results?

No single model performs well across all examples, but DBScan, which did require some tuning on the overlapping 3 cluster example, comes closest. Having said that the tuning needed would not have been so simple in higher dimensional space, and the overlapping 3 cluster is example is probably the representative of the most commonly occurring scenario we will encounter with real data.

As expected, the geometrically based k-means and Gaussian mixture model clustering methods perform poorly where the data has a more complex topological structure. However, most algorithms perform similarly on the more typical 3 cluster examples, and all work reliably on the most trivial distinct cluster examples. The behaviour on the uniform data illustrates the perils of applying clustering blindly. Most algorithms will produce a partition of the data set, but the value of this is questionable.

Below we show the plot of the same results, this time with no tuning of the algorithms on a much smaller, and arguably more realistic 40 sample data set.



The results are similar to the 400 sample case but show that some algorithms require more tuning than others – here the DBScan method is prone to suggesting points are simply noise if care is not taken.

It is also worth noting that with the cluster colouring applied to the examples, with 40 data points distributed uniformly (but randomly) it is easy for us to see clusters that are simply artefacts of the random data points.

Summary

Clustering remains an art. There are many methods and options within these, and no single ‘right’ answer. The choice of clustering algorithm to use should consider the likely structure of the clusters in the data set under study. This can be very difficult to conceptualise in high dimensional data. The issues discussed in part 1 remain relevant too – clustering is not the right tool to use if you have mixtures, or labels on your data. Here unmixing or classification would be more relevant. The key to effective use of any clustering method is integration with other views onto the data to support decision making. In essence a clustering is ‘good’ if it helps you understand your data better or improves your decision making.

References

Ester, M.; Kriegel, H.-P.; Sander, J. and Xu, X. (1996). In Simoudis, E.; Han, J. and Fayyad, U. M. (eds.). A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press, 226–231.

Estivill-Castro, V. (2002). Why so many clustering algorithms – A Position Paper. ACM SIGKDD Explorations Newsletter, **4** (1), 65–75.

Everitt, B. (2011). Cluster analysis. Wiley, Chichester, U.K.