Practical Data Science – Tute 9 / Week 10

PDS Teaching Team

RMIT UNIVERSITY

Activity 1 – DBSCAN

- Is it a supervised or unsupervised technique?

Reminder: The goal of clustering is to find structure, or the intrinsic grouping, in a collection of *unlabelled* data. Hence, **clustering** is an *unsupervised* technique.

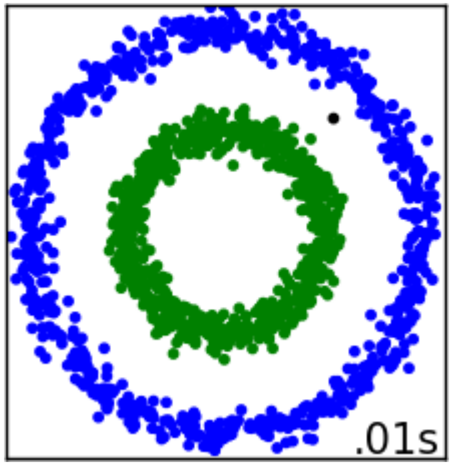
This is in contrast to **classification**, where a human expert has labelled examples into classes, and the goal is to learn from the training data, and to then be able to assign previously unseen instances into a known class.

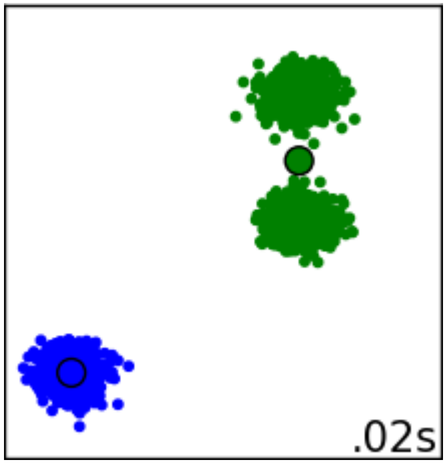
Activity 1 – DBSCAN

- What does a cluster mean in DBSCAN versus K-means?

DBSCAN

❑ Density-based clustering, i.e. a maximal set of density-connected points





K-means

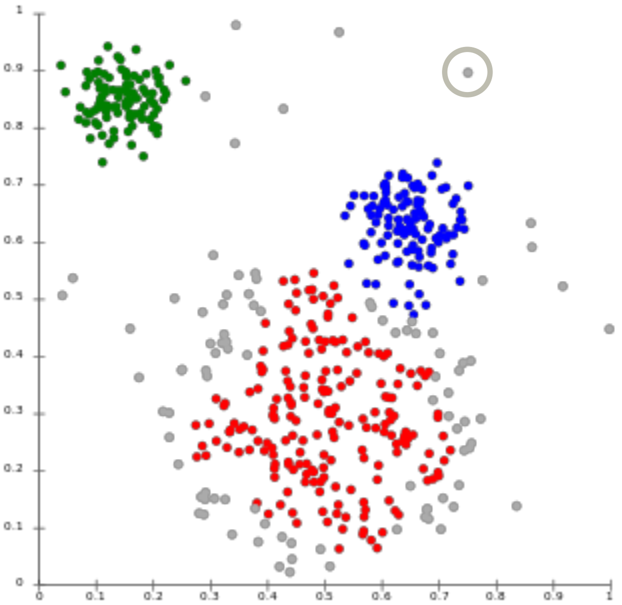
❑ Mean-based clustering, i.e. each observation belongs to the cluster with the nearest mean

http://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html#sphx-glr-auto-examples-cluster-plot-cluster-comparison-py

Activity 1 – DBSCAN

- What does noise mean in DBSCAN?

Outlier points that lie alone in low-density regions (whose nearest neighbours are too far away)



noise

http://upload.wikimedia.org/wikipedia/commons/thumb/2/28/DBSCAN-Gaussian-data.svg/372px-DBSCAN-Gaussian-data.svg.png

Activity 1 – DBSCAN

- Does this algorithm require a specification of the number of clusters in advance?

No, unlike *k-means*. It requires only the density threshold and minimal domain knowledge.

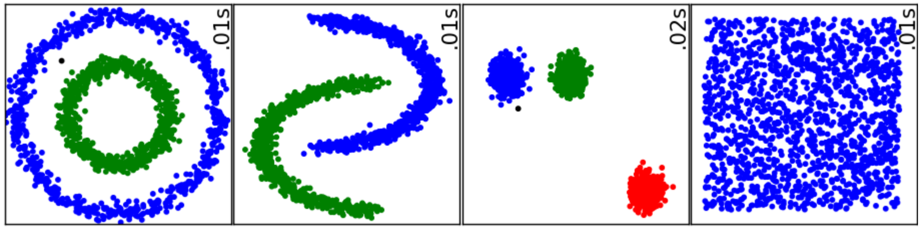
Activity 1 – DBSCAN

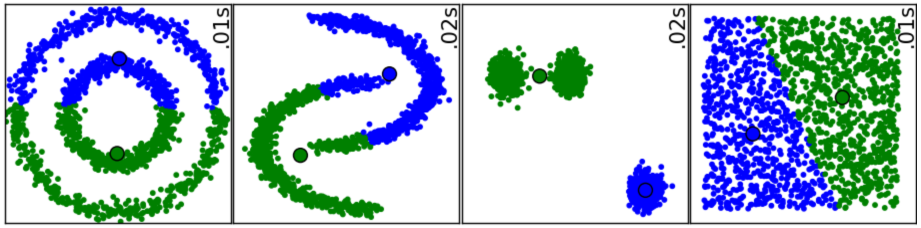
- ➢ Does DBSCAN expect a specific shape of clusters?

DBSCAN

❑ No

❑ Designed to discover clusters of arbitrary shape





K-means

❑ Yes

❑ Assumes shape of the clusters is spherical

http://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html#sphx-glr-auto-examples-cluster-plot-cluster-comparison-py

Activity 1 – DBSCAN

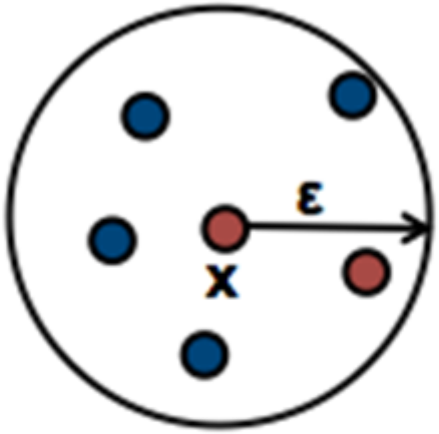
➢ What are the roles of the following two parameters in DBSCAN?

Eps-Neigbourhood

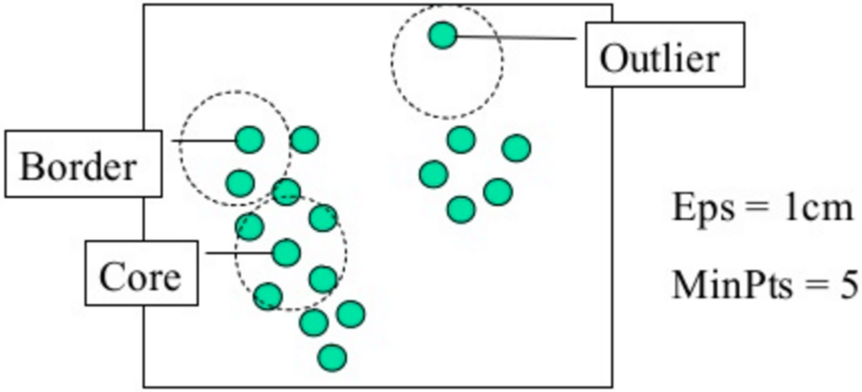
❑ The *Eps*-neighbourhood of a point *x* includes all data points whose distance to *x* is not larger than *Eps*, i.e. maximum radius of the neighbourhood.

MinPts

❑ Minimum number of points required in the *Eps*-neighbourhood of a point to differentiate between *core*, *border* and noise points.



http://www.sthda.com/sthda/RDoc/images/dbscan-principle.png

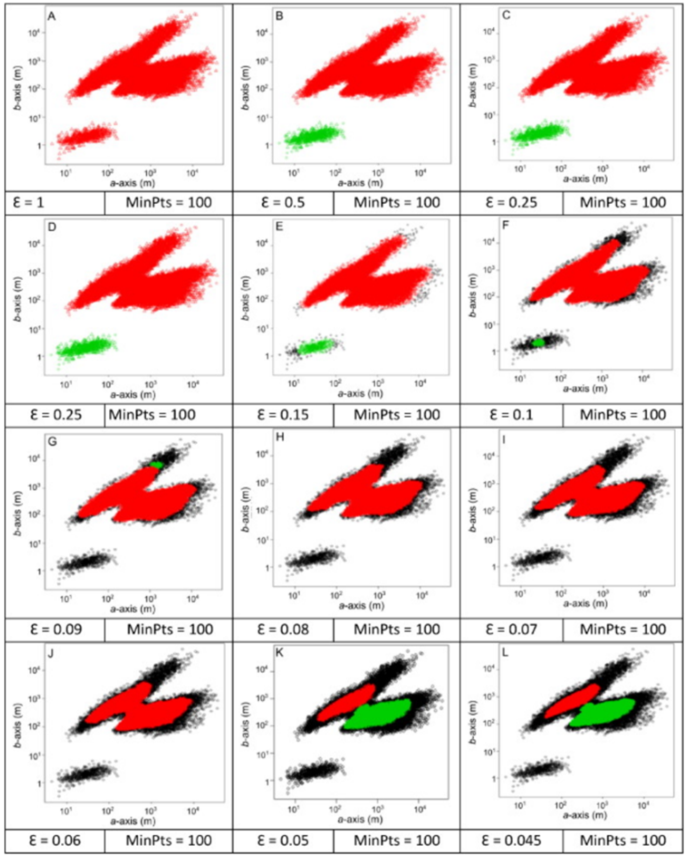


https://bradzzz.gitbooks.io/ga-seattle-dsi/content/dsi/dsi_07_unsupervised_learning/4.2-lesson/assets/images/dbscan.png

Activity 1 – DBSCAN

- ➢ How can a *k*-distance graph help to guide the selection of *Eps*?

$\varepsilon = 1$



$\varepsilon = 0.045$

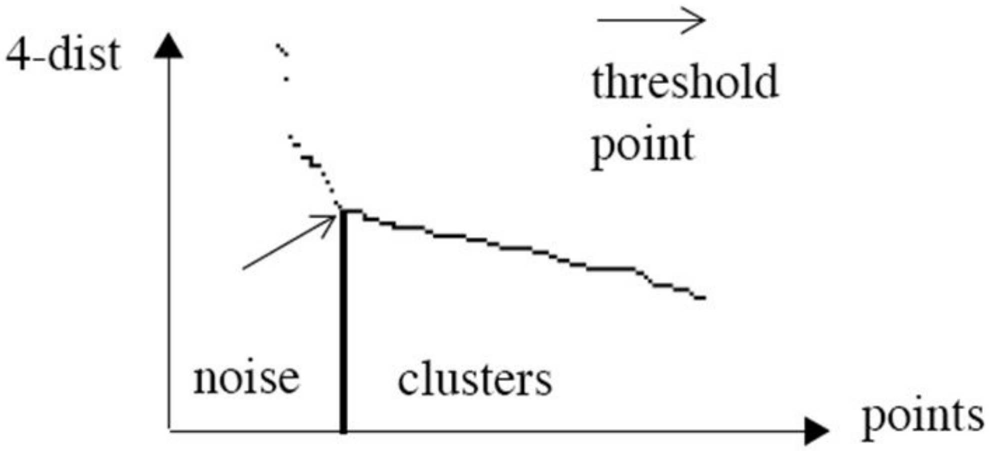
If *Eps* is too high → Clusters will merge and lose distinguishability

If *Eps* is too small → Large part of the data won't be clustered (noise)

https://www.researchgate.net/publication/290010634/figure/fig4/AS:316902506287104@1452566853079/DBSCAN-sensitivity-analysis-to-assess-how-many-clusters-exist-using-different-parameter.png

DBSCAN…

- How can a *k*-distance graph help to guide the selection of *Eps*?



sorted 4-dist graph for sample database 3

Plot the distance to *k* = *minPts* nearest neighbour and choose where this plot shows a strong bend (saturation)

Small values of *Eps* are preferable

http://images.slideplayer.com/39/11093461/slides/slide_25.jpg

Activity 1 – DBSCAN

- What are the advantages and disadvantages of the DBSCAN approach to clustering?

Advantages

- ❑ Doesn't require the number of clusters a priori
- ❑ Can find arbitrarily shaped clusters
- ❑ Robust to outliers
- ❑ Requires just two parameters
- ❑ Mostly insensitive to the ordering of points
- ❑ Designed for use with databases that can accelerate region queries
- ❑ *Eps* and *minPts* can be set by a domain expert

Disadvantages

- ❑ Not entirely deterministic
- ❑ Depends on the distance measure → problematic in high dimensional data
- ❑ Can't cluster datasets well with large differences in densities
- ❑ If no minimal domain knowledge is available, choosing a meaningful *Eps* can be difficult

Activity 2 – DBCAN in sklearn

`sklearn.cluster.DBSCAN(eps, min_samples, metric)`

- What does each of the parameters mean?

❑ **`eps`**

The maximum distance between two samples for them to be considered as in the same neighborhood.

❑ **`min_samples`**

The number of samples (or total weight) in a neighborhood for a point to be considered as a core point. This includes the point itself.

❑ **`metric`**

The metric to use when calculating distance between instances in a feature array. If metric is a string or callable, it must be one of the options allowed by metrics.pairwise.calculate_distance for its metric parameter. If metric is "precomputed", X is assumed to be a distance matrix and must be square. X may be a sparse matrix, in which case only "nonzero" elements may be considered neighbors for DBSCAN.

http://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html

Questions?