Clustering Data and Detecting Outliers

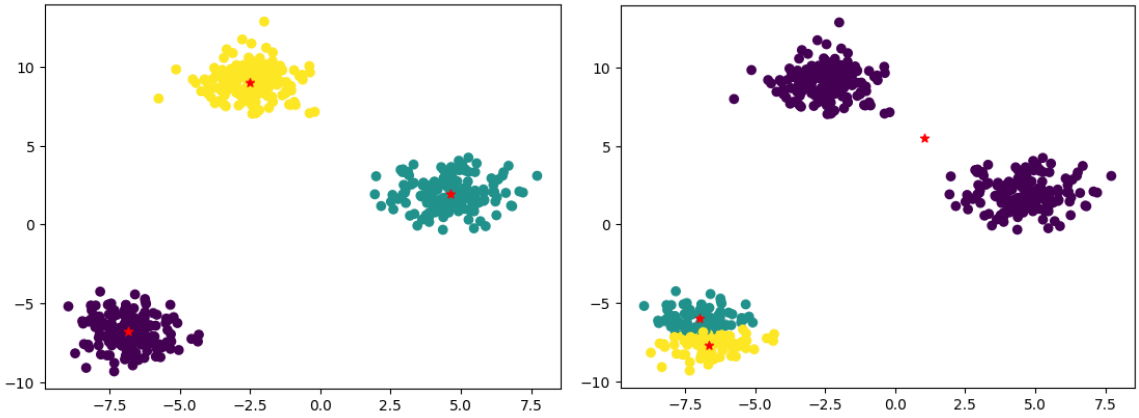
Implementing clustering algorithms from scratch and identifying outliers.

Learning Objectives

- K-Means Clustering
- Gaussian Means Mixture
- Local Outlier Factor
- DBSCAN

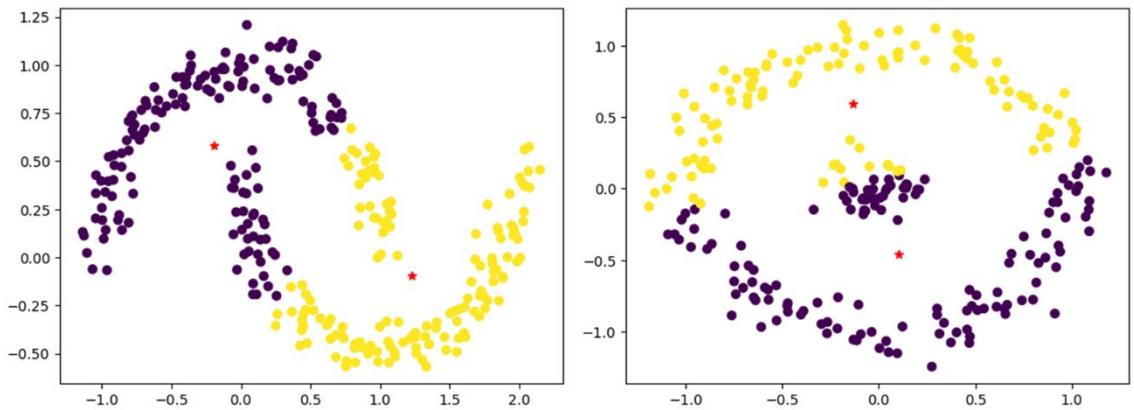
K-Means Clustering

K-Means clustering partitions datapoints by their closest centroid. The locations of the centroids are chosen to minimise the within-cluster variances (squared Euclidean distances). An example is shown below (left), with the centroids marked by red stars.



However, the efficacy of K-Means clustering is dependent on the starting locations of the centroids. If multiple centroids begin too close to the same cluster, that cluster may be incorrectly partitioned, as shown above (right).

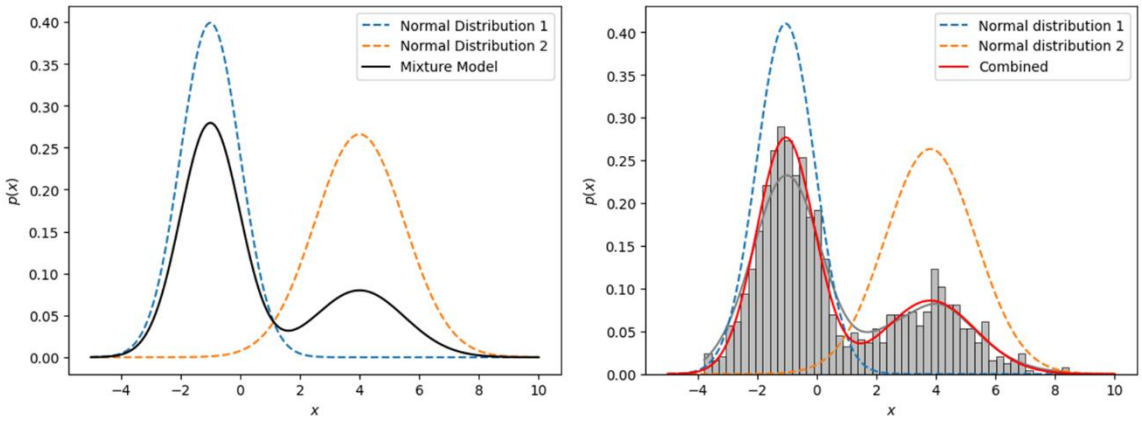
K-Means is effective for spherical clusters. However, it cannot identify more complex structures:



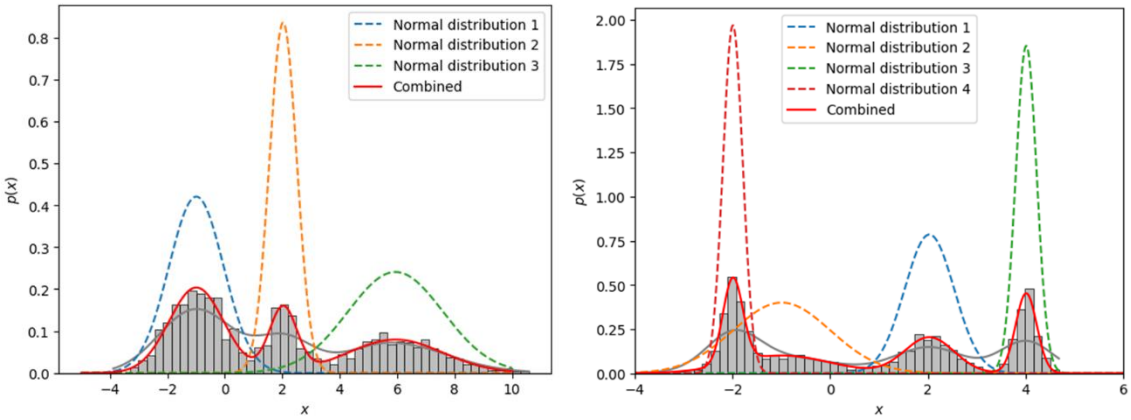
Gaussian Mixture Model

A Gaussian Mixture Model is a probabilistic clustering model that decomposes a dataset into the sum of a series Gaussian distributions. This method assigns each data point a probability of belonging to each cluster.

The diagram below (left) shows the sum of two weighted 1D Gaussian distributions. A dataset was generated by randomly sampling the two distributions. The dataset was then decomposed into its underlying normal distributions by the Gaussian Mixture Model, as shown below (right).



The Gaussian Mixture Model is able to decompose multiple distributions with high degrees of overlap, as shown below. However, the number of underlying distributions must be known in advance.

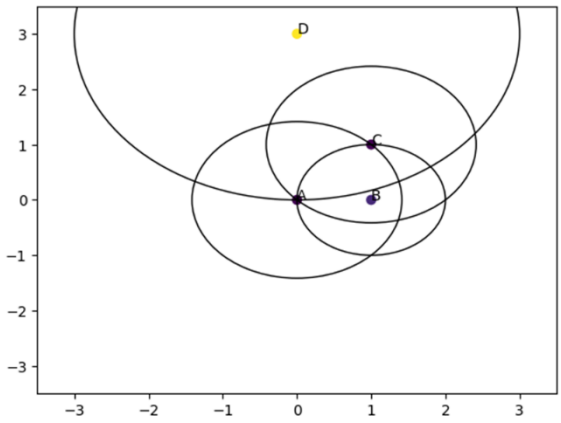


The Gaussian Mixture Model can be extended into multivariate clustering. It is effective for spheroid-shaped clusters but suffers when modelling more complex structures.

Local Outlier Factor

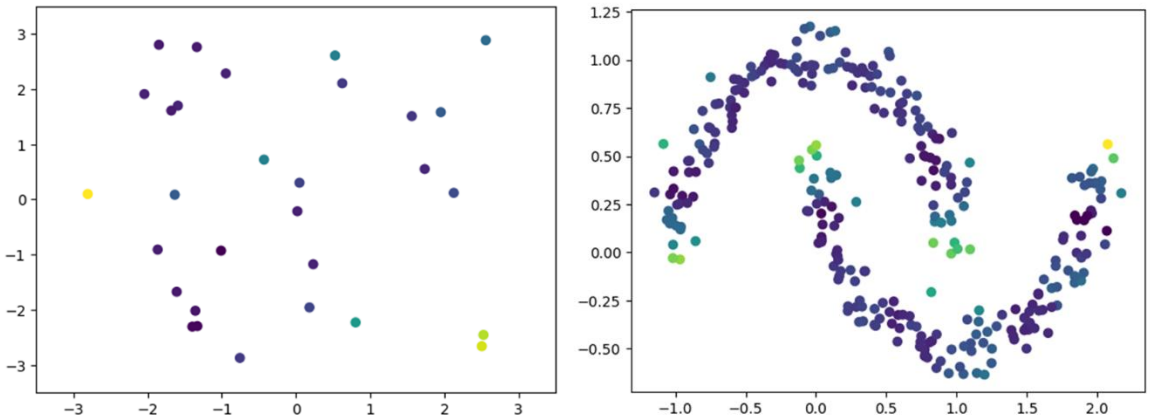
The Local Outlier Factor is a tool for detecting anomalies in a dataset using the local density around datapoints. If a point has a significantly lower local density than its nearest neighbours, it is classified as an outlier.

The local density of a point is calculated using its k-distance, the distance to its kth nearest neighbour. The circular k-distances (k = 2) about four points are shown below.



The local density about D is clearly much lower than the local densities about the above points, suggesting it to be an anomaly.

A larger example using 30 datapoints (k = 5) is shown below (left). Points that are more yellow have a larger outlier factor (and so are more likely to be outliers), whilst points that are more purple have smaller outlier factors.

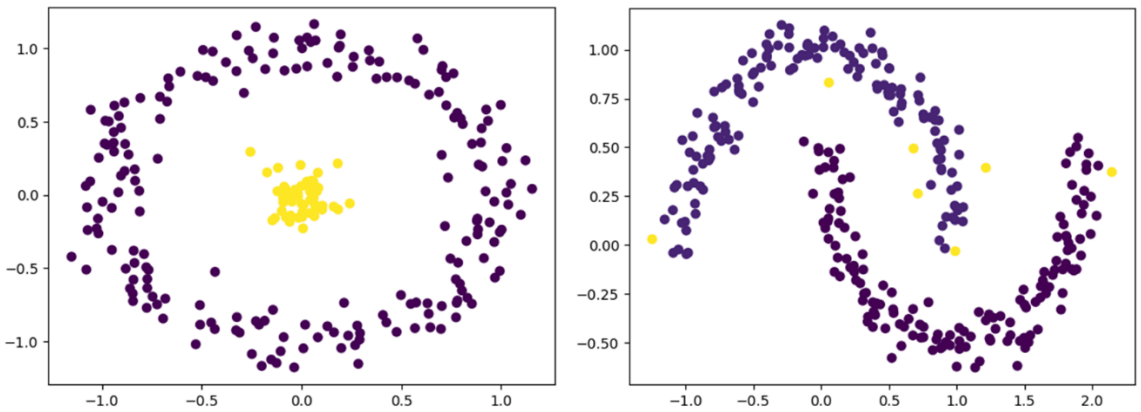


Since local density is used, rather than square Euclidean distance, the Local Outlier Factor can be used on clusters with complex structures, as shown above (right).

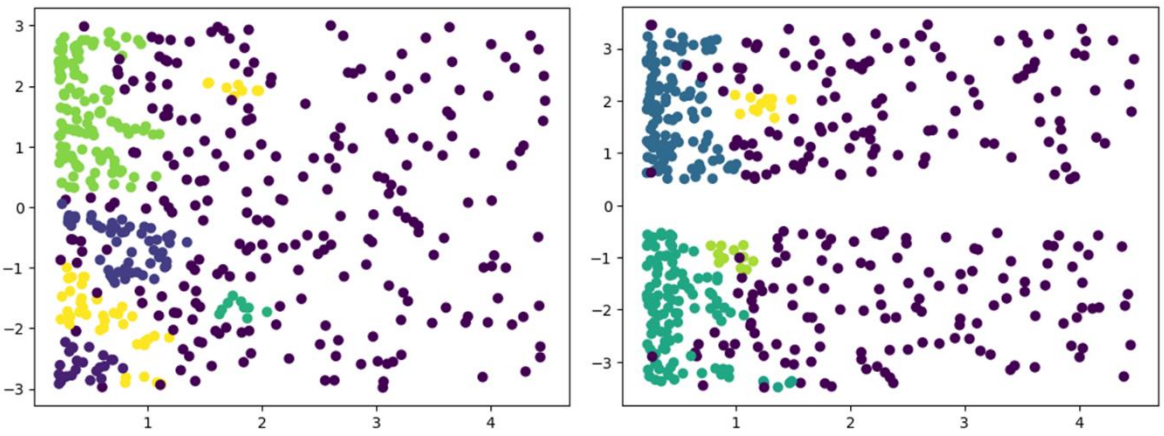
DBSCAN

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) partitions datapoints into groups that are closely packed, and marks points that lie in regions of low density as outliers. Points with many nearby neighbours are marked as *cores*. Points that are nearby *cores* are grouped with the *cores* and this process is repeated until the groups can no longer be extended. Any point that remains ungrouped is an outlier.

Like the Local Outlier Factor, DBSCAN uses local density, rather than square Euclidean distance, allowing it to be used on clusters with complex structures, as shown below.



The major weakness of DBSCAN is its difficulty at identifying clusters in data with varying density:



DBSCAN has been extended into density-based clustering algorithms such as OPTICS, which assigns clusters in increasing order of distance from the *cores*.

Conclusion

Clustering algorithms are power tools for grouping unlabelled data and detecting anomalies.

K-Means clustering is a simple and efficient algorithm for partitioning data into a fixed number of roughly spherical clusters. Due to its distance-based approach, it breaks down when trying to model more complex structures.

The Gaussian Mixture Model improves on K-Means by handling elongated and overlapping clusters. By assuming the data can be modelled as a sum of Gaussian distributions, it assigns probabilities for each point belonging to each cluster. These probabilities can then be used to detect outliers. The Gaussian Mixture Model is often used for spectral deconvolution when processing signals.

The Local Outlier Factor uses local density rather than distance to detect anomalies in a dataset. It can thus handle datasets with a complex structure.

DBSCAN extends the idea of local density to form clusters from extended regions of high density. Its density-based approach allows it to model complex structures with an unknown number of clusters. Any points that are not assigned to a cluster are assumed to be outliers.