Clustering Algorithms

Let's quickly look at types of clustering algorithms and when you should choose each type.

When choosing a clustering algorithm, you should consider whether the algorithm scales to your dataset. Datasets in machine learning can have millions of examples, but not all clustering algorithms scale efficiently. Many clustering algorithms work by computing the similarity between all pairs of examples. This means their runtime increases as the square of the number of examples n, denoted as $O(n^2)$ in complexity notation. $O(n^2)$ algorithms are not practical when the number of examples are in millions. This course focuses on the **k-means algorithm** (/machine-learning/glossary#k-means), which has a complexity of $O(n)$, meaning that the algorithm scales linearly with n.

Types of Clustering

Several approaches to clustering exist. For an exhaustive list, see A Comprehensive Survey of Clustering Algorithms (https://link.springer.com/article/10.1007/s40745-015-0040-1) Xu, D. & Tian, Y. Ann. Data. Sci. (2015) 2: 165. Each approach is best suited to a particular data distribution. Below is a short discussion of four common approaches, focusing on centroid-based clustering using k-means.

Centroid-based Clustering

Centroid-based clustering organizes the data into non-hierarchical clusters, in contrast to hierarchical clustering defined below. k-means is the most widely-used centroid-based clustering algorithm. Centroid-based algorithms are efficient but sensitive to initial conditions and outliers. This course focuses on k-means because it is an efficient, effective, and simple clustering algorithm.

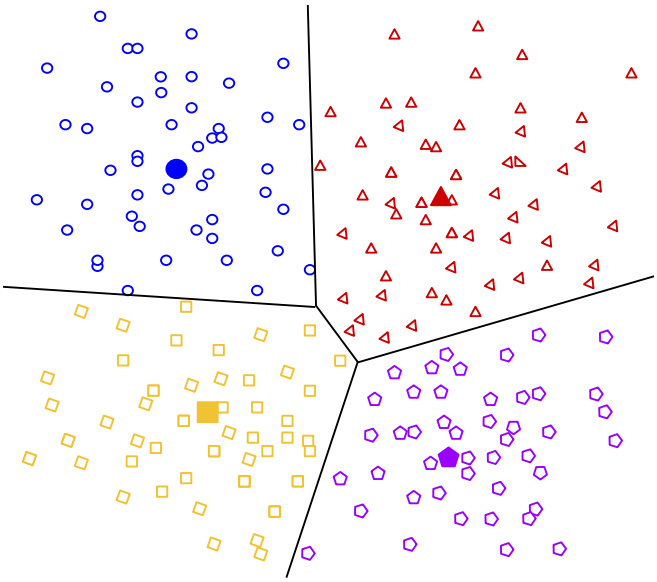

Figure 1: Example of centroid-based clustering.

Density-based Clustering

Density-based clustering connects areas of high example density into clusters. This allows for arbitrary-shaped distributions as long as dense areas can be connected. These algorithms have difficulty with data of varying densities and high dimensions. Further, by design, these algorithms do not assign outliers to clusters.

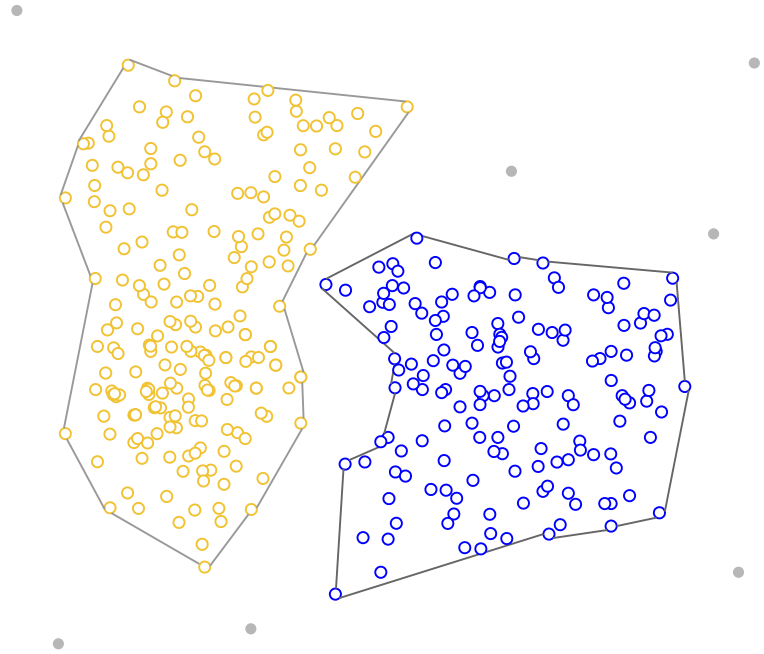

Figure 2: Example of density-based clustering.

Distribution-based Clustering

This clustering approach assumes data is composed of distributions, such as **Gaussian distributions** (https://wikipedia.org/wiki/Normal_distribution). In Figure 3, the distribution-based

algorithm clusters data into three Gaussian distributions. As distance from the distribution's center increases, the probability that a point belongs to the distribution decreases. The bands show that decrease in probability. When you do not know the type of distribution in your data, you should use a different algorithm.

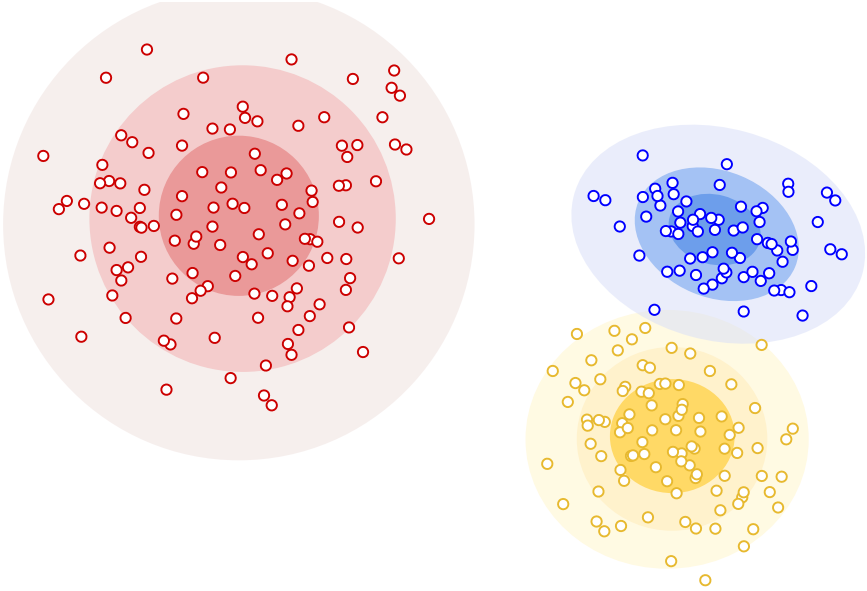

Figure 3: Example of distribution-based clustering.

Hierarchical Clustering

Hierarchical clustering creates a tree of clusters. Hierarchical clustering, not surprisingly, is well suited to hierarchical data, such as taxonomies. See *Comparison of 61 Sequenced Escherichia coli Genomes* (https://www.researchgate.net/figure/Pan-genome-clustering-of-E-coli-black-and-related-species-colored-based-on-the_fig1_45152238) by Oksana Lukjancenko, Trudy Wassenaar & Dave Ussery for an example. In addition, another advantage is that any number of clusters can be chosen by cutting the tree at the right level.

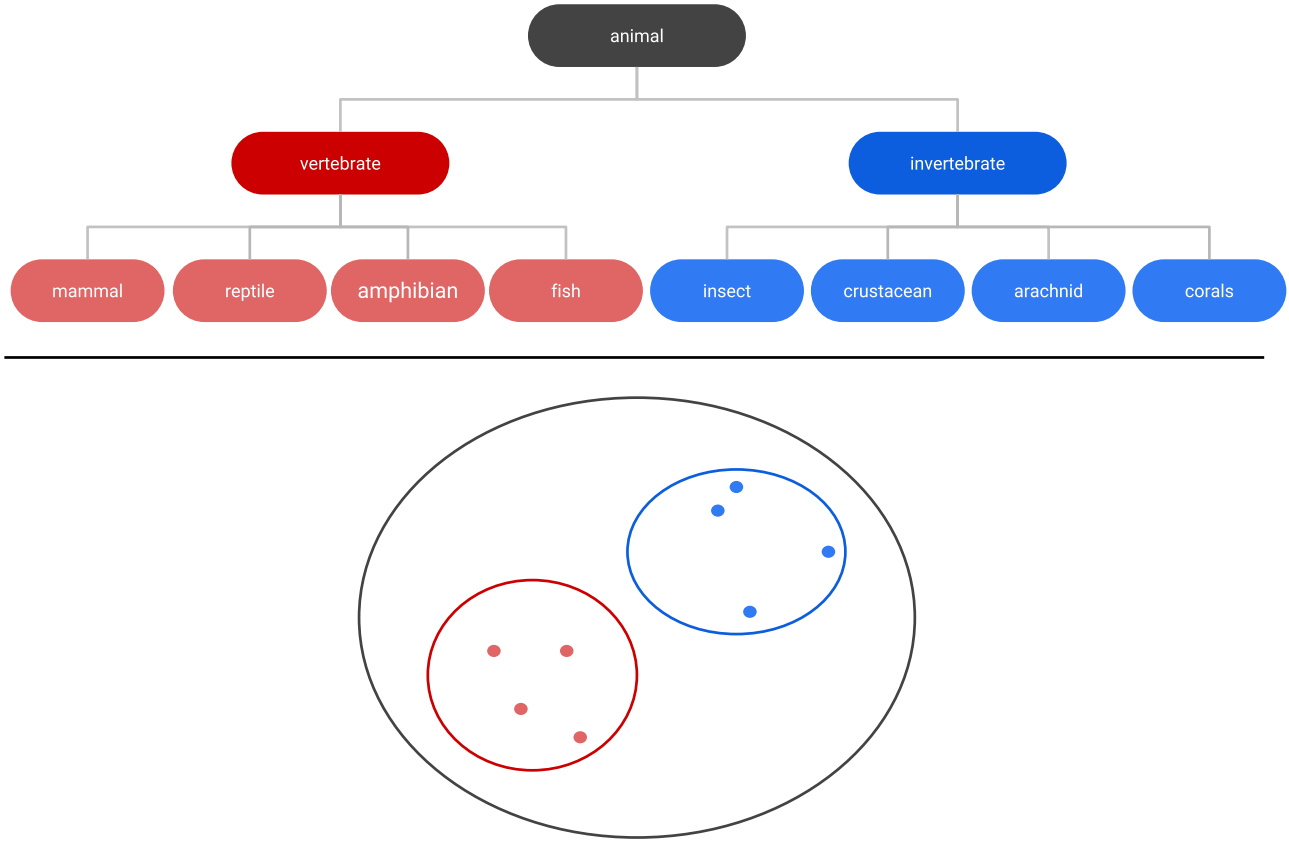

Figure 4: Example of a hierarchical tree clustering animals.

ɪrms:

k-means algorithm (/machine-learning/glossary#k-means)

Gaussian distributions (/machine-learning/glossary#Normal_distribution)

Previous
← What is Clustering? (/machine-learning/clustering/overview)

Next
Overview (/machine-learning/clustering/workflow) →