Clustering

CSC 461: Machine Learning

Fall 2022

Prof. Marco Alvarez
University of Rhode Island

Unsupervised learning

- Algorithms/methods for **uncovering** latent structure in the data
  - ✓ unlabeled data
- Observations / data instances / data points:

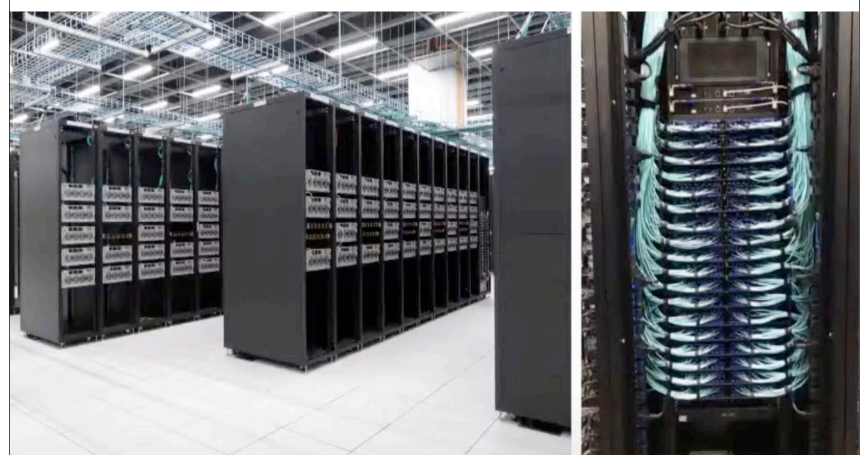
$$\mathscr{D} = \{\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}\}$$

usually $\mathbf{x_i} \in \mathbb{R}^d$

Labels may be available with the data, but should be ignored if unsupervised learning is applied.

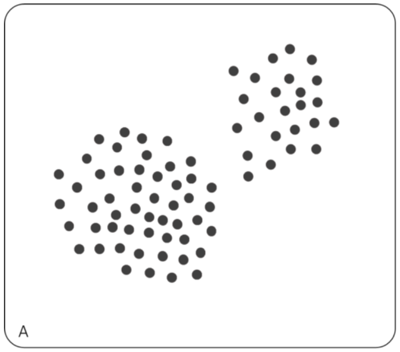
Cluster Analysis

Not this type of clusters …



Credit: https://blogs.nvidia.com/blog/2021/06/22/tesla-av-training-supercomputer-nvidia-a100-gpus/

Clustering unlabeled data



Given **n** feature vectors

group them into K clusters based on pairwise similarities

Credit: https://towardsdatascience.com/a-step-by-step-guide-for-clustering-images-4b45f9906128

Subjective nature

(a) Original points. (b) Two clusters.

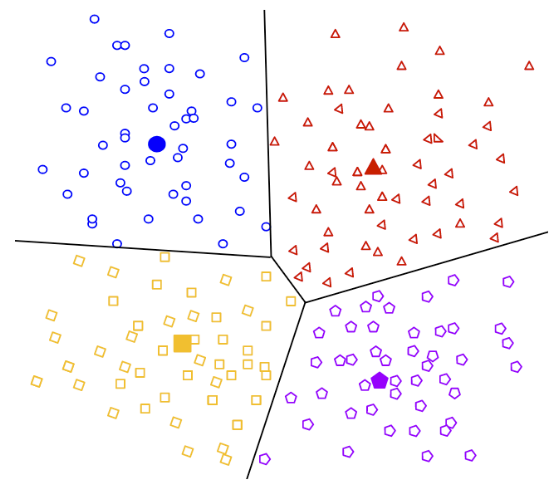
(c) Four clusters. (d) Six clusters.

Credit: https://www-users.cse.umn.edu/~kumar001/dmbook/index.php

Types of clustering

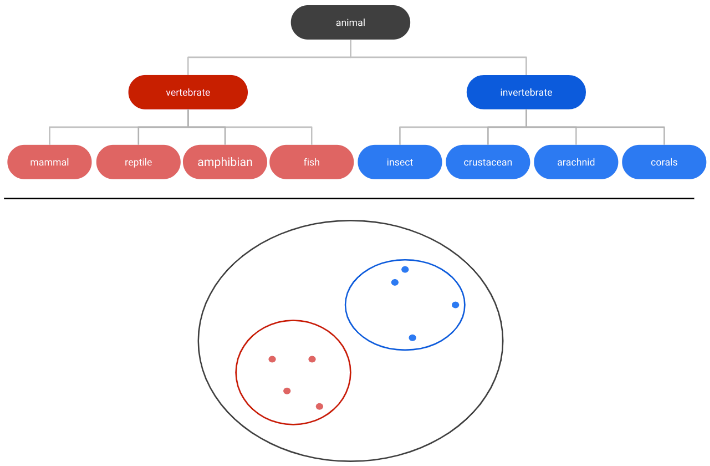
- **Partitioning** clustering (centroid-based)
  - ✓ each observation belongs to one partition (centroid)
  - ✓ each partition has a centroid
  - ✓ observations are assigned iteratively to the cluster with the closest centroid
- **Hierarchical** clustering
  - ✓ defines a hierarchy (tree) of clusters
  - ✓ sensitive to the definition of distance between observations

Partition clustering



Credit: https://developers.google.com/machine-learning/clustering/clustering-algorithms

Hierarchical clustering

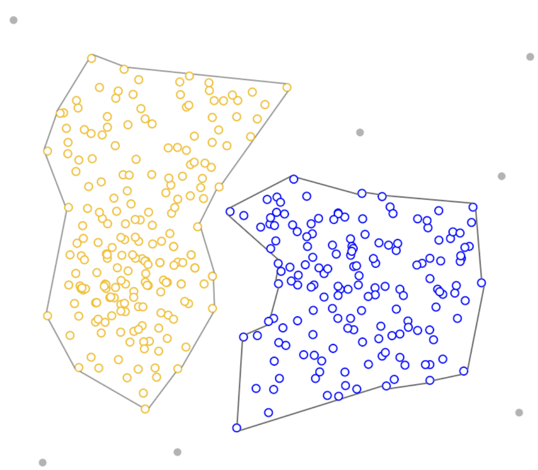


Credit: https://developers.google.com/machine-learning/clustering/clustering-algorithms

Types of clustering

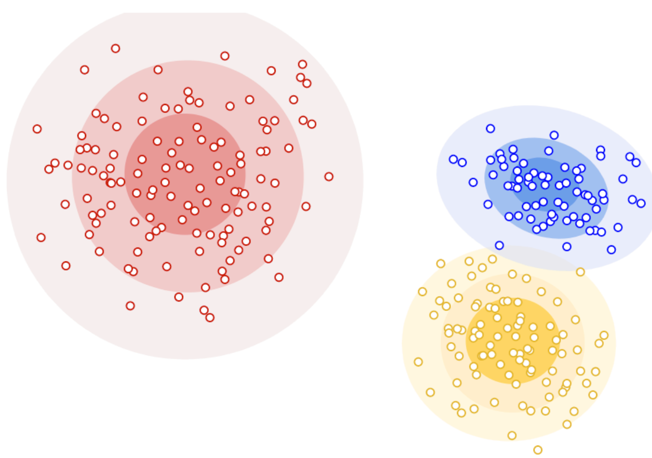
- Density-based
  - ✓ clusters are formed in high-density areas
  - ✓ can form arbitrary-shaped clusters
  - ✓ difficult to deal with varying densities and high dimensions
- Distribution-based
  - ✓ assumes clusters are generated by underlying distributions (e.g. gaussian distribution)
  - ✓ requires defining the proper distribution
- and more
  - ✓ fuzzy clustering, spectral clustering, etc.

Density-based



Credit: https://developers.google.com/machine-learning/clustering/clustering-algorithms

Distribution-based



Credit: https://developers.google.com/machine-learning/clustering/clustering-algorithms

K-Means

K-means clustering

- Given **n** observations (feature vectors) …

$$\mathscr{D} = \{\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}\} \qquad \mathbf{x_i} \in \mathbb{R}^d$$

- … partition the data into **K** clusters such that the within-cluster-distance is minimized for all clusters $\mathbf{C_i}$:

$$\arg\min_{C_1,\ldots,C_K} \sum_{k=1}^{K} WCD(\mathbf{C_k})$$

Formally

- Each cluster is denoted by $C_i \subseteq \{1,\ldots,n\}$
  - ✓ if $\mathbf{x_i}$ is assigned to cluster j then $i \in C_j$
  - ✓ subject to:

$$\bigcup_i C_i = \{1,\ldots,n\} \quad \text{and} \quad C_i \cap C_j = \varnothing, \text{ for } i \neq j$$

- Defining the clustering goal:

$$\arg\min_{C_1,\ldots,C_K} \sum_{k=1}^{K} \frac{1}{2|C_k|} \sum_{i,j \in C_k} \|\mathbf{x_i} - \mathbf{x_j}\|_2^2$$

Credit: https://www.cs.cornell.edu/courses/cs4780/2022fa/lectures/UnsupervisedLearning.html

Using centroids

- A **centroid** $\mu_{\mathbf{k}}$ is the mean of all observations (datapoints) in cluster C_k

$$\mu_{\mathbf{k}} = \frac{1}{|C_k|} \sum_{i \in C_k} \mathbf{x_i}$$

- The clustering goal can also be defined using centroids:

$$\arg\min_{C_1,\ldots,C_K} \sum_{k=1}^{K} \sum_{i \in C_k} \|\mathbf{x_i} - \mu_{\mathbf{k}}\|_2^2$$

Credit: https://www.cs.cornell.edu/courses/cs4780/2022fa/lectures/UnsupervisedLearning.html

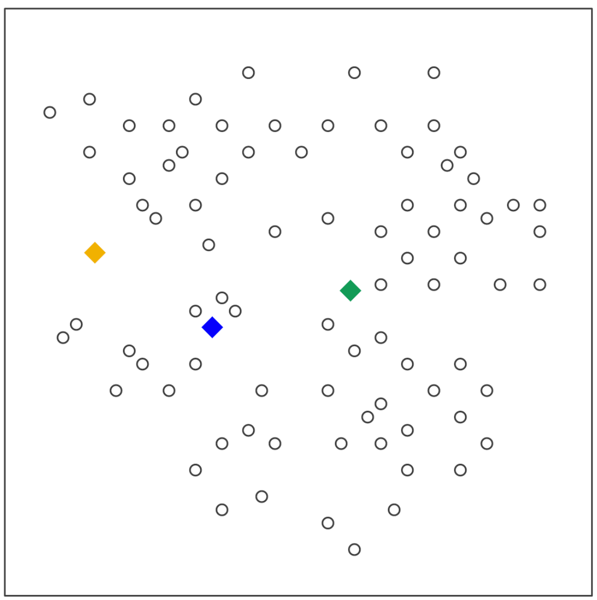
How to minimize the goal?

- Trying a brute-force approach for an optimal solution would be computationally infeasible
  - ✓ i.e. calculating all possible partitions of n observations into K clusters
  - ✓ $n = 25, K = 4$ gives 5×10^{13} possible partitions
- Relaxing our minimization goal
  - ✓ instead of an optimal we can settle with an approximate solution, using an efficient iterative method (Lloyd's algorithm)

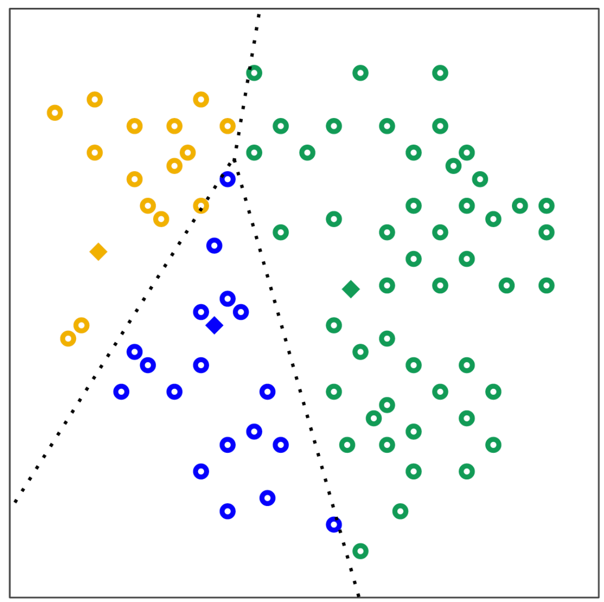
k-Means algorithm

- Randomly **assign** observations to clusters
- Repeat
  - ✓ compute all cluster **centroids**
  - ✓ **assign** each observation to their closest centroid
  - ✓ stop if observations stop changing clusters

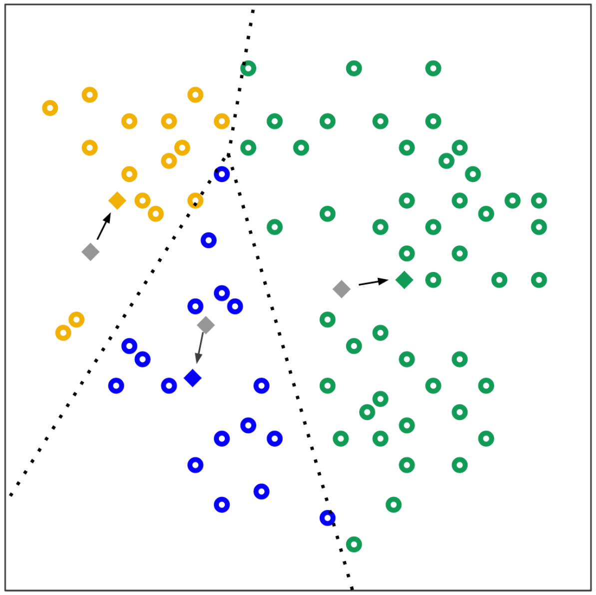
can also use different stopping criteria for large datasets



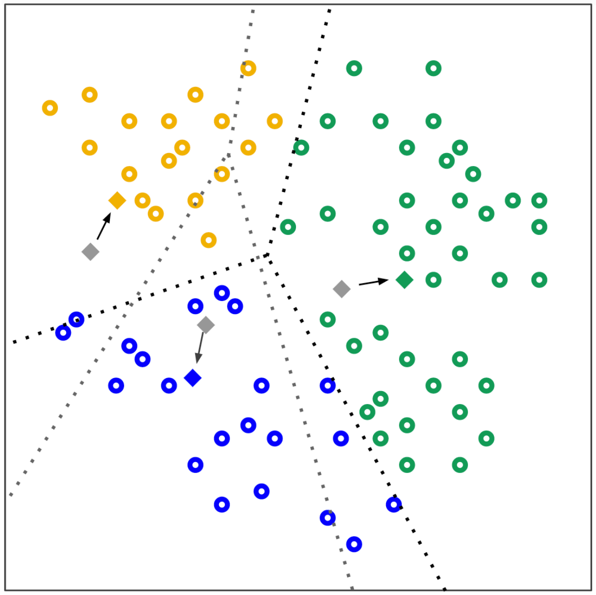
Credit: https://developers.google.com/machine-learning/clustering/clustering-algorithms



Credit: https://developers.google.com/machine-learning/clustering/clustering-algorithms



Credit: https://developers.google.com/machine-learning/clustering/clustering-algorithms



Credit: https://developers.google.com/machine-learning/clustering/clustering-algorithms

Online demo

https://user.ceng.metu.edu.tr/~akifakkus/courses/ceng574/k-means/