# CS456: Machine Learning

## Unsupervised learning: Clustering

Jakramate Bootkrajang

Department of Computer Science
Chiang Mai University

# Objectives

- To understand unsupervised learning
- To understand basic concept of data clustering
- To understand the working of k-means clustering algorithm

# Outlines

- Unsupervised learning
- Data clustering
- K-means algorithm

# Unsupervised learning

- A task of inferring a function $f(\mathbf{x})$ which maps input $\mathbf{x}$ to output $y$
- Unlike supervised learning, there is no definitive answer to what the value of $y$ should be
- Data is available to the algorithm in the form of $\{\mathbf{x}_i\}_{i=1}^N$

  (no label or target)

# Examples

- Data clustering
  - input=feature vector $\mathbf{x}$, output=cluster label $y \in \{1, \ldots, K\}$
- Dimensionality reduction, auto-encoder
  - input=feature vector $\mathbf{x} \in R^m$, output=$\mathbf{x} \in R^k$, $k < m$
- Independent Component Analysis
  - input=feature vector $\mathbf{x}$, output=$\mathbf{z}_1 + \mathbf{z}_2 + \cdots + \mathbf{z}_k = \mathbf{x}$

# Data clustering

- the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters) [1]
- The notion of a "cluster" cannot be precisely defined
- Current clustering algorithms employ different definition of clustering *heuristic*

[1] https://en.wikipedia.org/wiki/Cluster_analysis

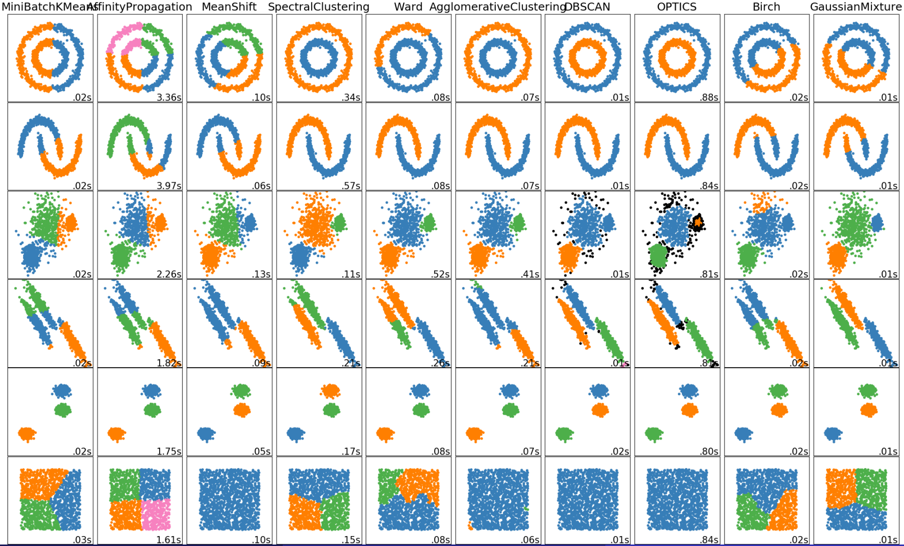
# Data clustering heuristic (1/2)

- Connectivity-based: objects is more related to nearby objects than to objects further away
  - hierarchical clustering
- Centroid-based: clusters are represented by a central vector
  - k-means algorithm

# Data clustering heuristic (1/2)

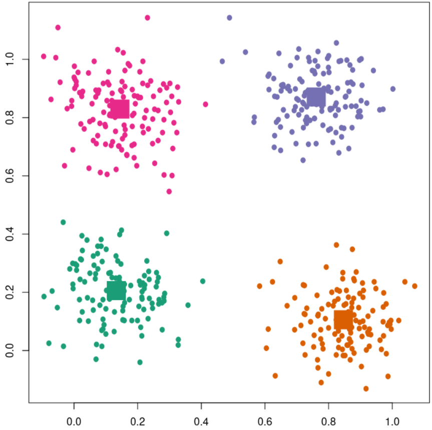
- Distribution-based: Clusters can be defined as objects belonging most likely to the same distribution
  - Gaussian Mixture Model
- Density-based: clusters are defined as areas of higher density than the remainder of the data set
  - DBSCAN

# Data clustering algorithms



# Ideas

- The basic idea is to describe each cluster by its mean value
- Assign data point to its nearest cluster



# K-means algorithm: notation

- Let $\{\mathbf{x}_i\}_{i=1}^{N}$ denotes a set of $m$ dimensional data points
- Let $d(\mathbf{x}_i, \mathbf{x}_j)$ denotes distance measure (e.g., Euclidean distance) between $\mathbf{x}_i$ and $\mathbf{x}_j$: $d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(x_i^1 - x_j^1)^2 + \cdots + (x_i^m - x_j^m)^2}$
- Let $\mathbf{z}$ denotes a vector of cluster assignment of length $N$
  - For example, if $\mathbf{x}_i$ belongs to cluster 3, $z_i = 3$

# K-means algorithm: objective

- In k-means, the success of clustering is measured by the sum of the squared distances of each point to each assigned mean

$$f(\mathbf{z}, \mu_1, \mu_2, \ldots, \mu_k) = \frac{1}{2} \sum_{i=1}^{N} ||\mathbf{x}_i - \mu_{z_i}||^2$$

- k-means will minimise this objective function

# K-means algorithm (1/2)

$$f(\mathbf{z}, \mu_1, \mu_2, \ldots, \mu_k) = \frac{1}{2} \sum_{i=1}^{N} ||\mathbf{x}_i - \mu_{z_i}||^2$$

- If we fix $\mu_1, \mu_2, \ldots, \mu_k$, it is easy to see that to minimise the objective we must assign $\mathbf{x}$ to the nearest cluster

$$z_i = \arg \min_{j=\{1,\ldots,k\}} d(\mathbf{x}_i, \mu_j)$$

# K-means algorithm (2/2)

$$f(\mathbf{z}, \mu_1, \mu_2, \ldots, \mu_k) = \frac{1}{2} \sum_{i=1}^{N} ||\mathbf{x}_i - \mu_{z_i}||^2$$

- Now if we fix **z** (not allow **x** to move), we see that to further minimise the objective we must update the mean.
- Taking derivative of the objective w.r.t $\mu_j$ we will get a closed-form solution

$$\mu_j = \frac{1}{N_j} \sum_{i:z_i=j} \mathbf{x}_i$$

# K-means algorithm summary

1. Initialise: choose initial cluster $\mu_{1:k}$ arbitrarily
2. Repeat
   - assign $\mathbf{x}$ to the nearest cluster

$$z_i = \arg \min_{j=\{1,\ldots,k\}} d(\mathbf{x}_i, \mu_j)$$

   - update the mean

$$\mu_j = \frac{1}{N_j} \sum_{i:z_i=j} \mathbf{x}_i$$

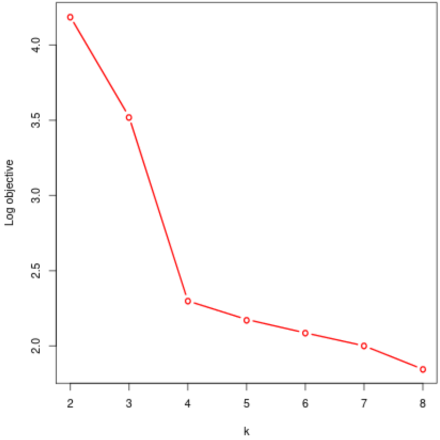
3. Until $\mathbf{z}$ do not change (data points no longer move between cluster)

# Remarks

- In k-means, there is no correct answer to guide the algorithm (unsupervised)
- What the algorithm does is to optimise the heuristic criteria defined by user
- Due to random initialisation, result of each run might not be the same
- Choosing $k$ can be done by plotting objective function values versus $k$ and pick $k$ which exhibit a 'kink'

# Remarks

The kink occurs at $k = 4$ so it is highly possible that there are 4 natural clusters



# Objectives: revisited

- To understand unsupervised learning
- To understand basic concept of data clustering
- To understand the working of k-means clustering algorithm