Introduction to Data Science

Clustering

Arijit Mondal

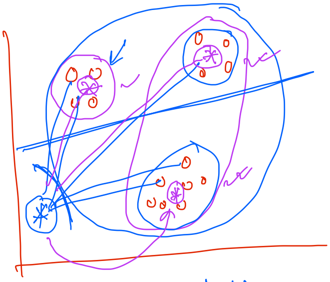
Dept. of Computer Science & Engineering

Indian Institute of Technology Patna

arijit@iitp.ac.in

Introduction

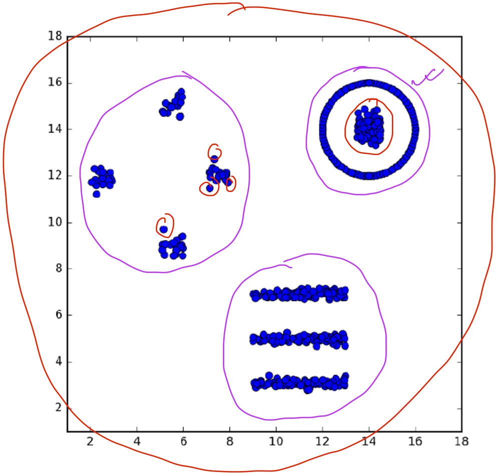
- Clustering is the problem of grouping of points by similarity.
- Splitting of points based on similarity
- Applications
  - Hypothesis development
  - Modeling over smaller subset of data
  - Data reduction
  - Outliers detection



bottom-up

10000 x 100

Example



9 3
8 4

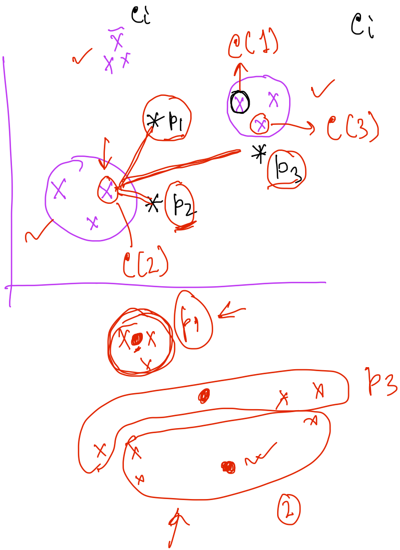
Many

1 — ∞

K

 image source: Data Science Design Manual

k-means Clustering



$k=1$

$k=2$

$\boxed{k=3}$

$$\min\{d(c_i, p_1), d(c_i, p_2), d(c_i, p_3)\}$$

p_1

label c_i with the class of p_j which is closest to c_i

$p_i \leftarrow$ Average of the points in the current cluster

Example

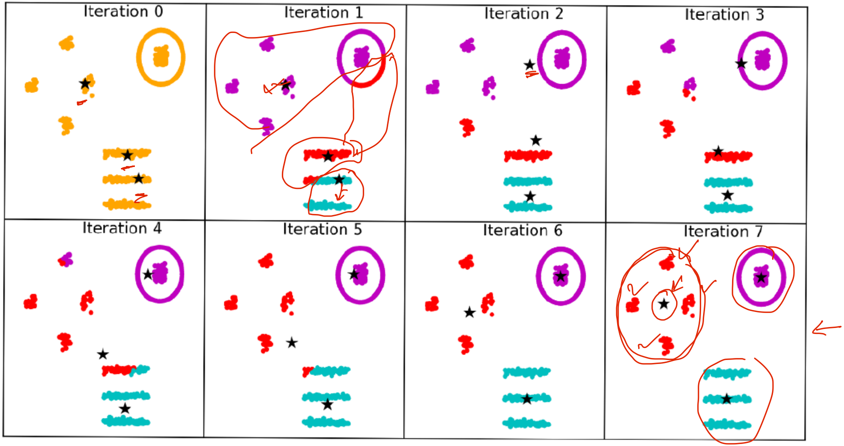
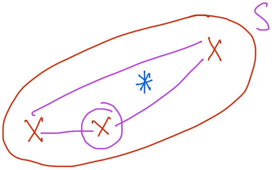


 image source: Data Science Design Manual

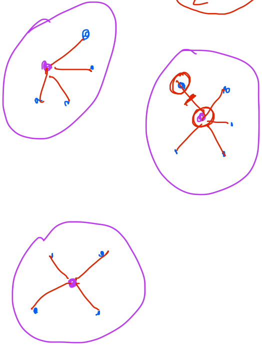
Centers vs Centroids

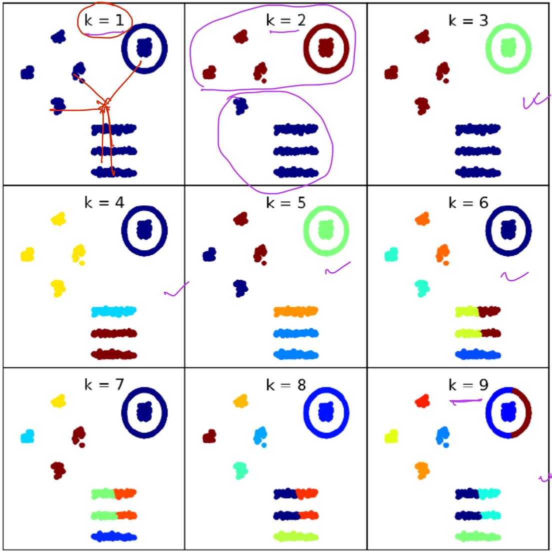
- Centroids: $C_d = \frac{1}{|S|} \sum_{p \in S} p[d]$
- Centers: $\arg\min_{c \in S} \sum_{i=1}^{n} d(c, p_i)$



Number of clusters

$E = \sum d_i$



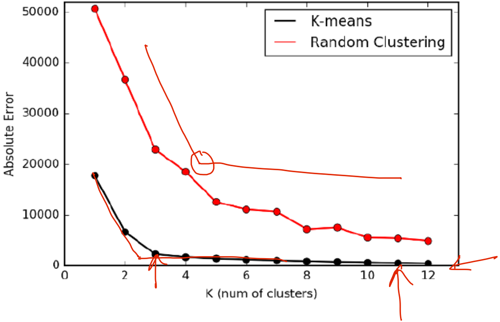


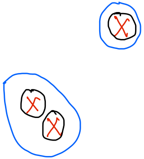
$\boxed{\text{Best}}$?

k

image source: Data Science Design Manual

Number of clusters





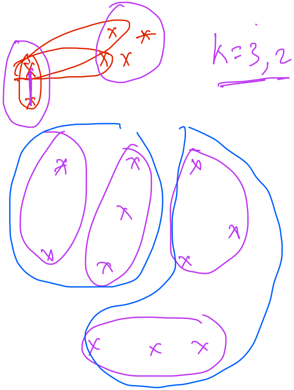
Black → E = 0

~~Red~~ Blue → E > 0

image source: Data Science Design Manual

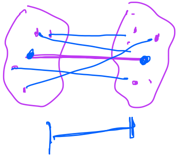
Agglomerative clustering

- A bottom-up approach
- Combining similar items
- Distance measures - C_1, C_2 are some clusters



Agglomerative clustering

- A bottom-up approach
- Combining similar items
- Distance measures - C_1, C_2 are some clusters
  - Nearest neighbor - $d(C_1, C_2) = \min_{x \in C_1, y \in C_2} \|x - y\|$

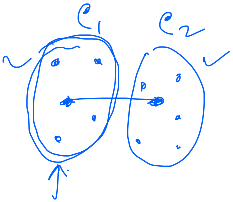


Agglomerative clustering

- A bottom-up approach
- Combining similar items
- Distance measures - C_1, C_2 are some clusters
  - Nearest neighbor - $d(C_1, C_2) = \min_{x \in C_1, y \in C_2} \|x - y\|$
  - Average link - $d(C_1, C_2) = \frac{1}{|C_1| \cdot |C_2|} \sum_{x \in C_1} \sum_{y \in C_2} \|x - y\|$

Agglomerative clustering

- A bottom-up approach
- Combining similar items
- Distance measures - C_1, C_2 are some clusters
  - Nearest neighbor - $d(C_1, C_2) = \min_{x \in C_1, y \in C_2} \|x - y\|$
  - Average link - $d(C_1, C_2) = \frac{1}{|C_1| \cdot |C_2|} \sum_{x \in C_1} \sum_{y \in C_2} \|x - y\|$
  - Nearest centroid - closest centroids, $|C_1| \cdot |C_2|$ point-pairs



$\binom{C_1}{C_2}$

$C_1 \times C_2$

Agglomerative clustering

- A bottom-up approach
- Combining similar items
- Distance measures - C_1, C_2 are some clusters
  - Nearest neighbor - $d(C_1, C_2) = \min_{x \in C_1, y \in C_2} \|x - y\|$
  - Average link - $d(C_1, C_2) = \frac{1}{|C_1| \cdot |C_2|} \sum_{x \in C_1} \sum_{y \in C_2} \|x - y\|$
  - Nearest centroid - closest centroids, $|C_1| \cdot |C_2|$ point-pairs
  - Furthest link - $d(C_1, C_2) = \max_{x \in C_1, y \in C_2} \|x - y\|$

Distance measures

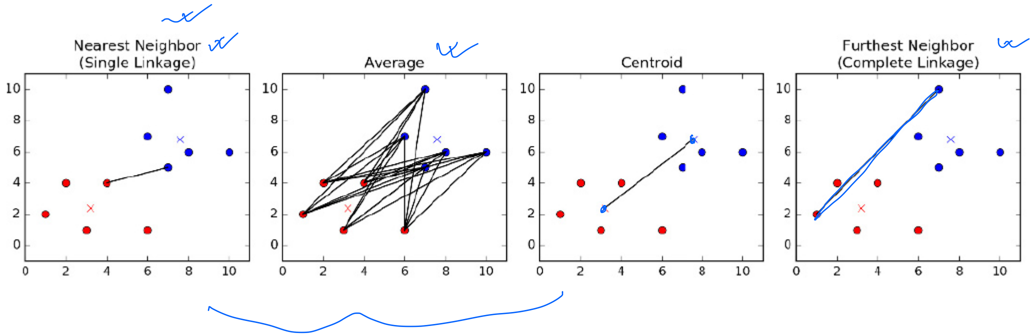
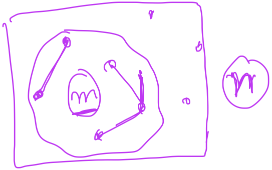
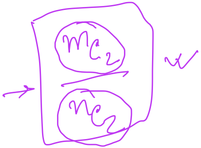


image source: Data Science Design Manual

Comparing clustering

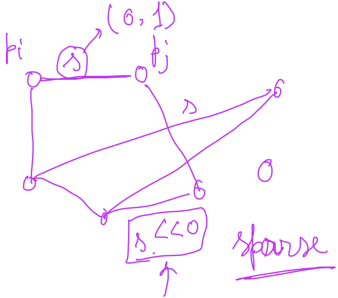
- Jaccard similarity: Similarity between two sets, $J(s_1, s_2) = \frac{|s_1 \cap s_2|}{|s_1 \cup s_2|} \rightarrow \frac{\text{common}}{\text{union}}$
- Jaccard distance: $\boxed{1 - J(s_1, s_2)}$. It is distance metric
- Rand index: ratio of compatible pairs to all possible pairs





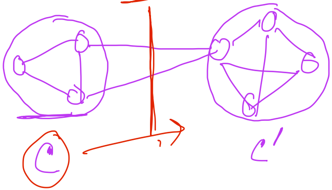
Similarity & Cuts

- Similarity measure - $S[i,j] = e^{-\beta\|p_i - p_j\|}$
- Similarity graph - It is based on similarity measure. Can be made sparse by applying thresholding on similarity values
- Cluster weight - $W(C) = \sum_{x\in C}\sum_{y\in C} S[i,j]$
- Cut weight - $W'(C) = \sum_{x\in C}\sum_{y\in V-C} S[i,j]$
- Conductance of cluster C is $\dfrac{W'(C)}{W(C)}$



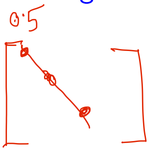
$e^{-x} \to 1$ $\quad x = 0$

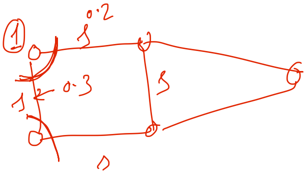
$\to 0$ $\quad x \to \infty$



Spectral clustering

- Construct the Laplacian matrix $L = D - S$
  - S - similarity matrix
  - D - degree-weighted identity matrix, $D[i,i] = \sum_j S[i,j]$
- The most valuable eigenvectors for clustering here turn out to have the smallest non-zero eigenvalues
- Applying k-means clustering in this feature space produces connected clusters





Spectral clustering

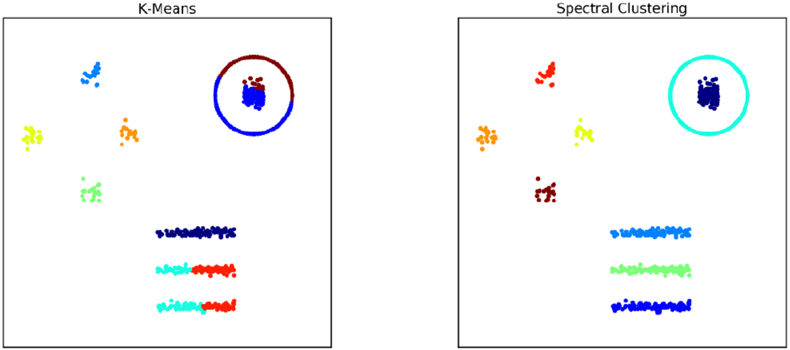


image source: Data Science Design Manual