Unsupervised methods

Dr. Jakub Kuzilek

Introduction

Today you will learn:

- How clustering works
- Hierarchical clustering algorithm
  - How to compute similarity between samples (clusters)
- k-means algorithm
  - How to determine number of clusters k

Introduction

Small recap: What is Unsupervised learning?

Introduction

Small recap: What is Unsupervised learning?

- We do not know right answers
- Unsupervised learning algorithm is inferring function, which describes hidden structure of unlabelled data. We cannot estimate error of algorithm.

Unsupervised learning

- We do not know right answers.
- Unsupervised learning algorithm is inferring function, which describes hidden structure of **unlabelled** data.
- We cannot estimate error of algorithm

- Input:

$$T = \{x_1, x_2, x_3, \ldots, x_N\}$$

- Output (depending on type):
  - **Labels:** $\{\boldsymbol{y_1}, \boldsymbol{y_2}, \boldsymbol{y_3}, \ldots, \boldsymbol{y_N}\}$
  - Transformed data: $\{\hat{x}_1, \hat{x}_2, \hat{x}_3, \ldots, \hat{x}_N\}$
  - Distribution: $f(x; \theta)$

Clustering

Distance function

Can you tell me properties of distance (metrics) function?

Distance function

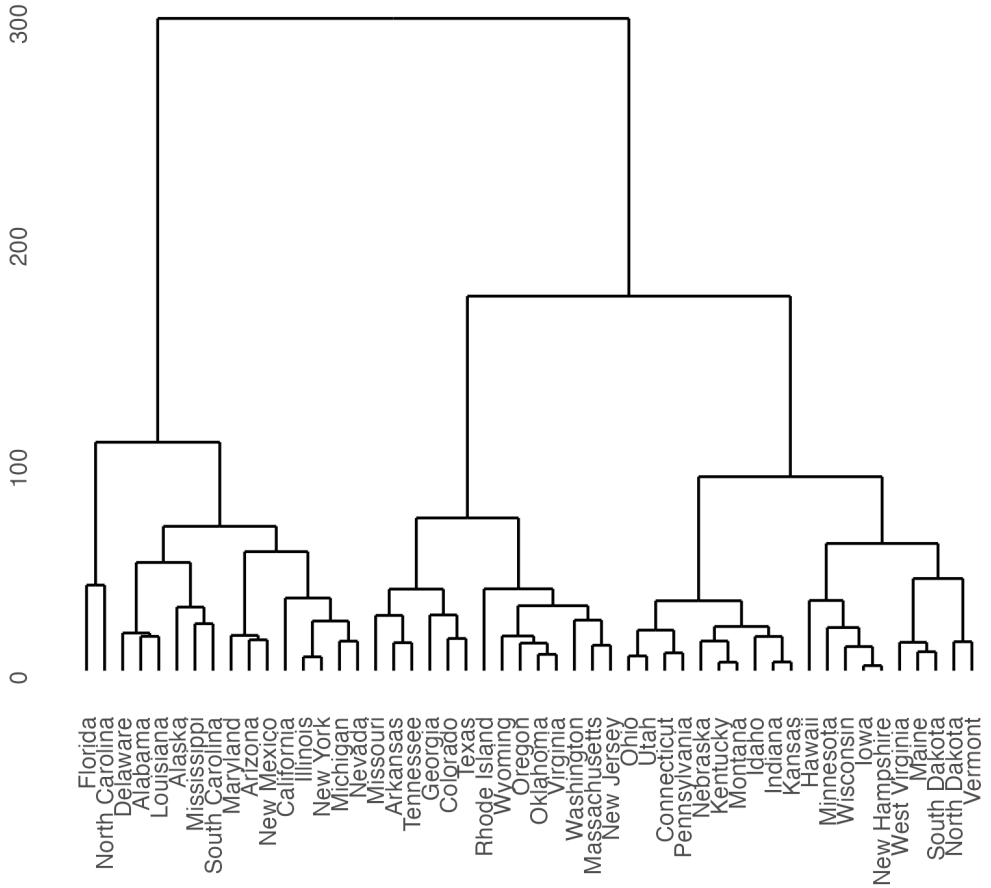
Can you tell me properties of distance (metrics) function?

1. $d(x, x) = 0$
2. $d(x, y) \geq 0$
3. $d(x, y) = d(y, x)$
4. $d(x, z) \leq d(x, y) + d(y, z)$

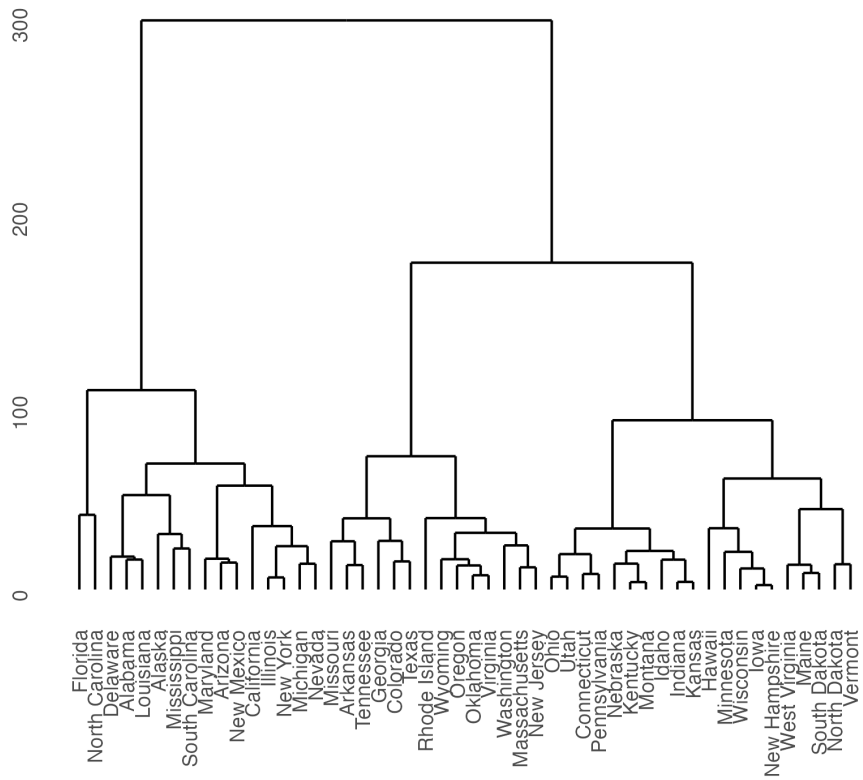
Agglomerative Hierarchical Clustering

Agglomerative hierarchical clustering

- Input: $T = \{x_1, x_2, x_3, \ldots, x_N\}$
- Output: Dendrogram



Agglomerative hierarchical clustering



Algorithm:

- Compute dissimilarity (proximity, distance) matrix $D_{ij} \geq 0$, of size $N \times N$, where N is the number of samples
- Initialize clusters as singletons (each sample is 1 cluster): $C_i \leftarrow \{i\}$
- Initialize set of clusters available for merging $S \leftarrow \{1, \ldots, N\}$
- Repeat:
  1. Select 2 most similar clusters: $(j,k) \leftarrow argmin_{j,k \in S} D_{j,k}$
  2. Create new cluster $C_l \leftarrow C_j \cup C_k$
  3. Remove (j,k) from set of available clusters S
  4. If C_l contains all samples end (no cluster available for merging)
  5. Update dissimilarity matrix D_{il} for all available clusters (set S)

Agglomerative hierarchical clustering

How to compute similarity between two clusters?

Agglomerative hierarchical clustering

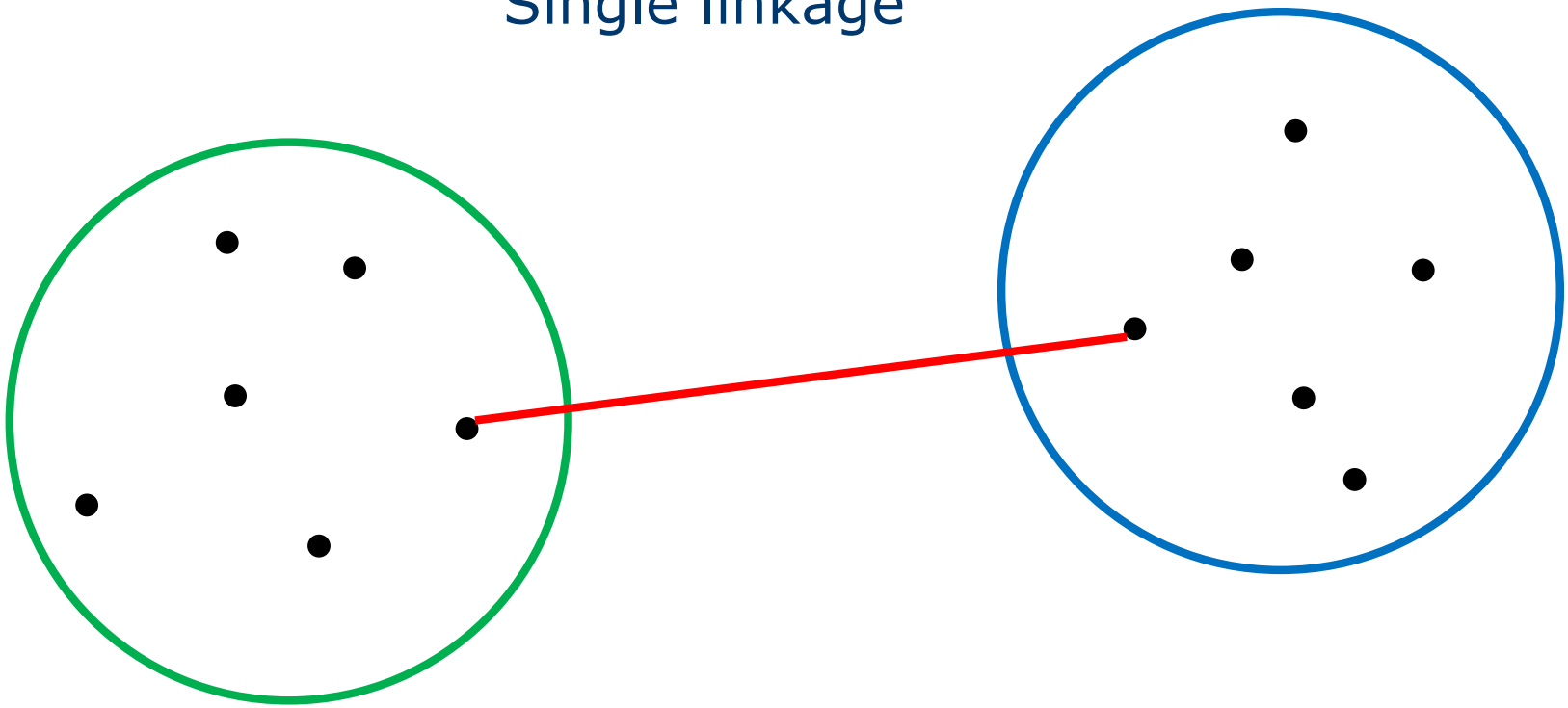
How to compute similarity between two clusters?

There are multiple ways how to do that.

Agglomerative hierarchical clustering

How to compute similarity between two clusters?

Single linkage

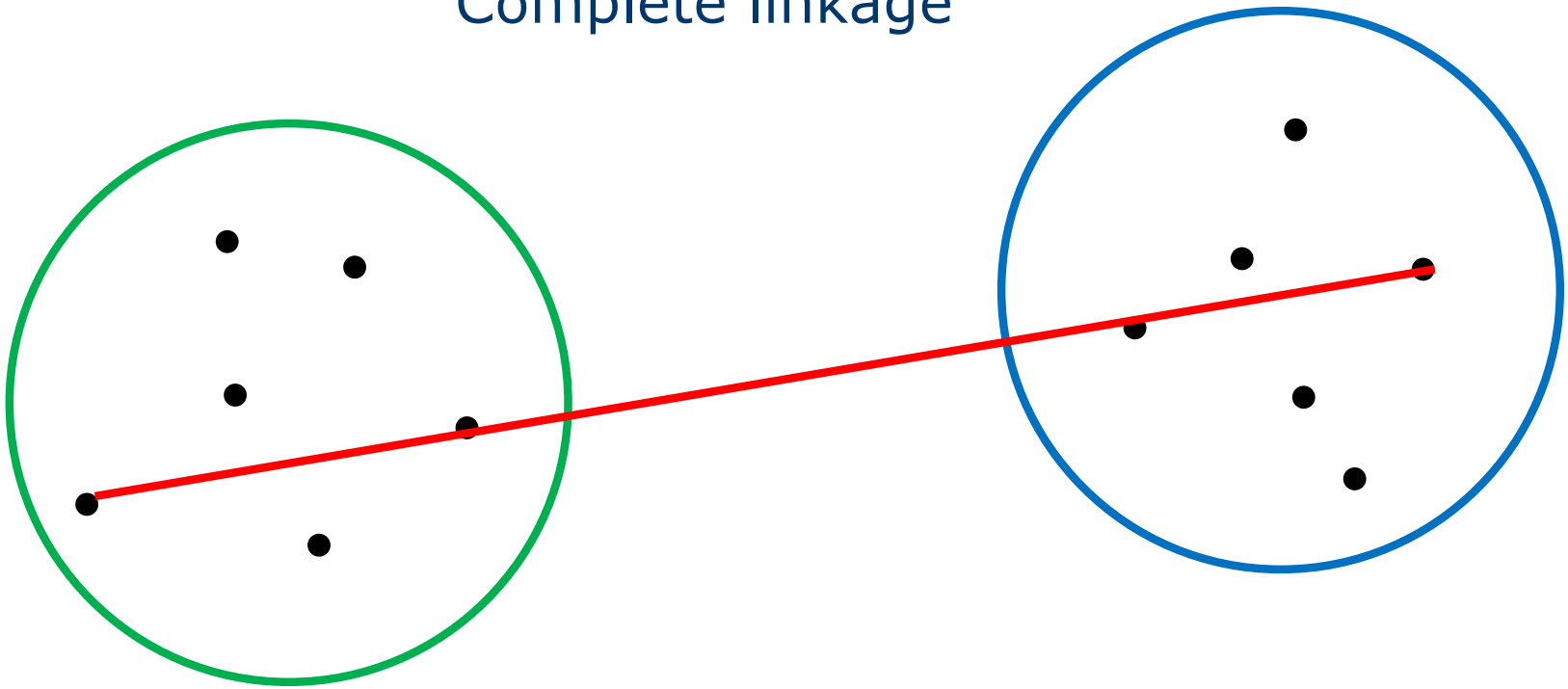


- Can cluster non-elliptical shapes
- Cannot separate clusters if there is noise

Agglomerative hierarchical clustering

How to compute similarity between two clusters?

Complete linkage

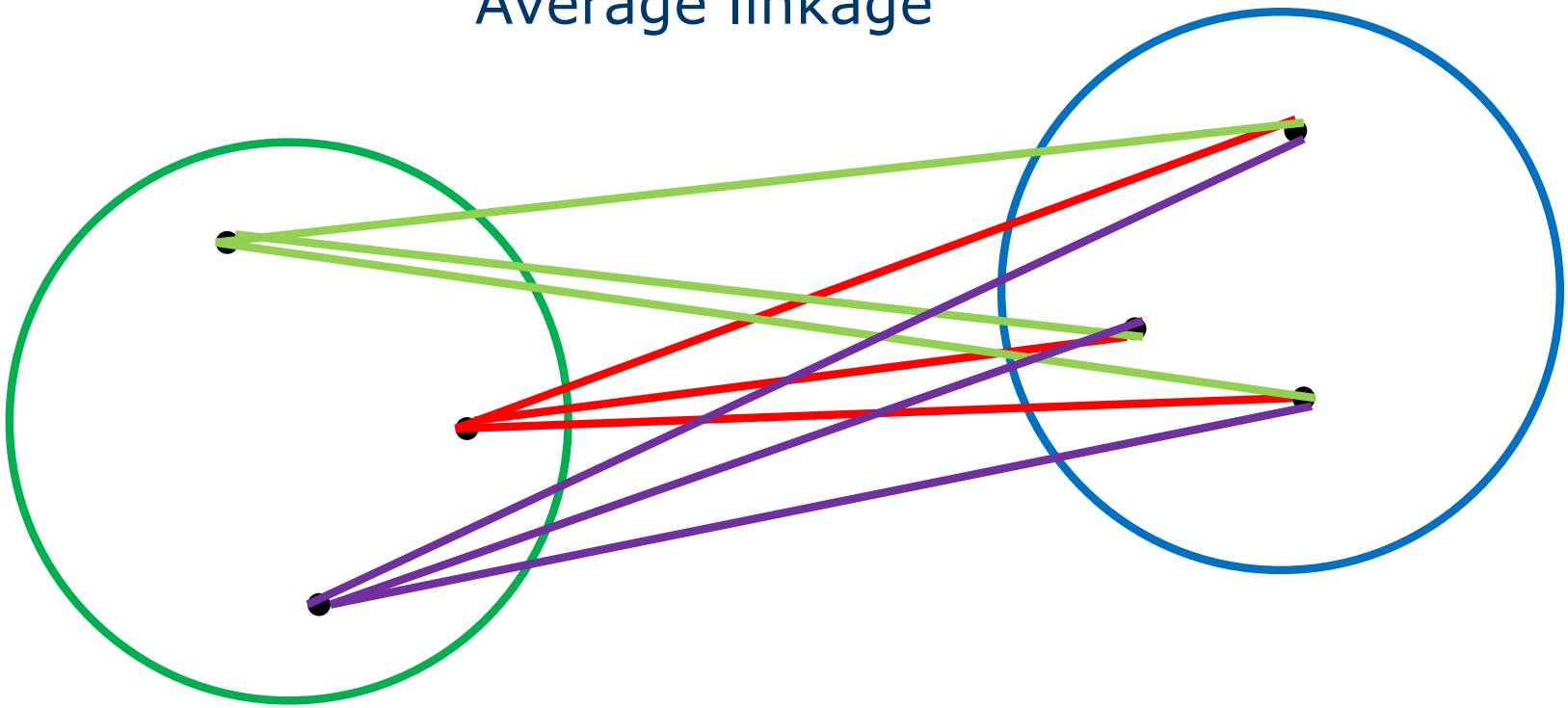


- Does well if there is noise
- Tends to break large clusters
- Bias towards elliptical shape clusters

Agglomerative hierarchical clustering

How to compute similarity between two clusters?

Average linkage

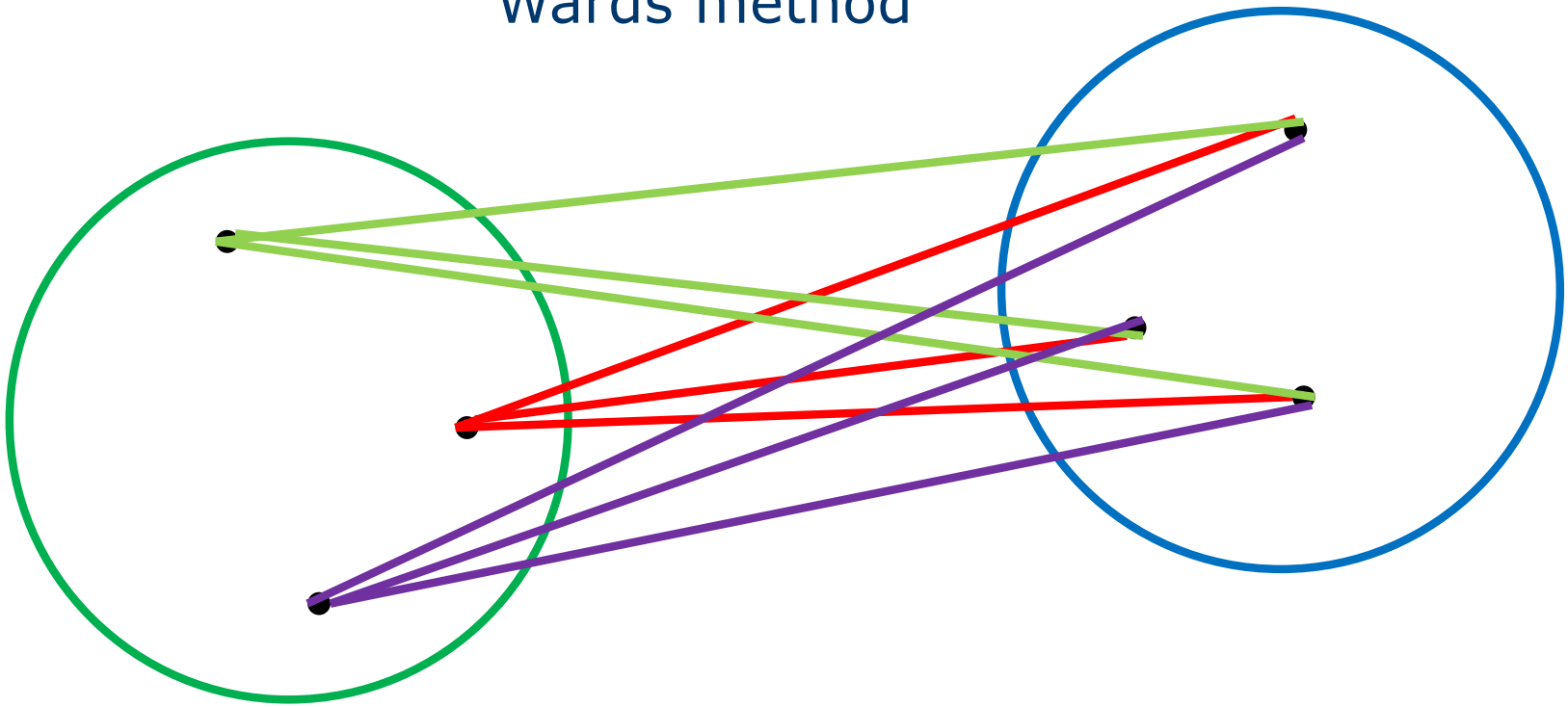


- Does well when there is noise
- Biased towards elliptical shape clusters

Agglomerative hierarchical clustering

How to compute similarity between two clusters?

Wards method

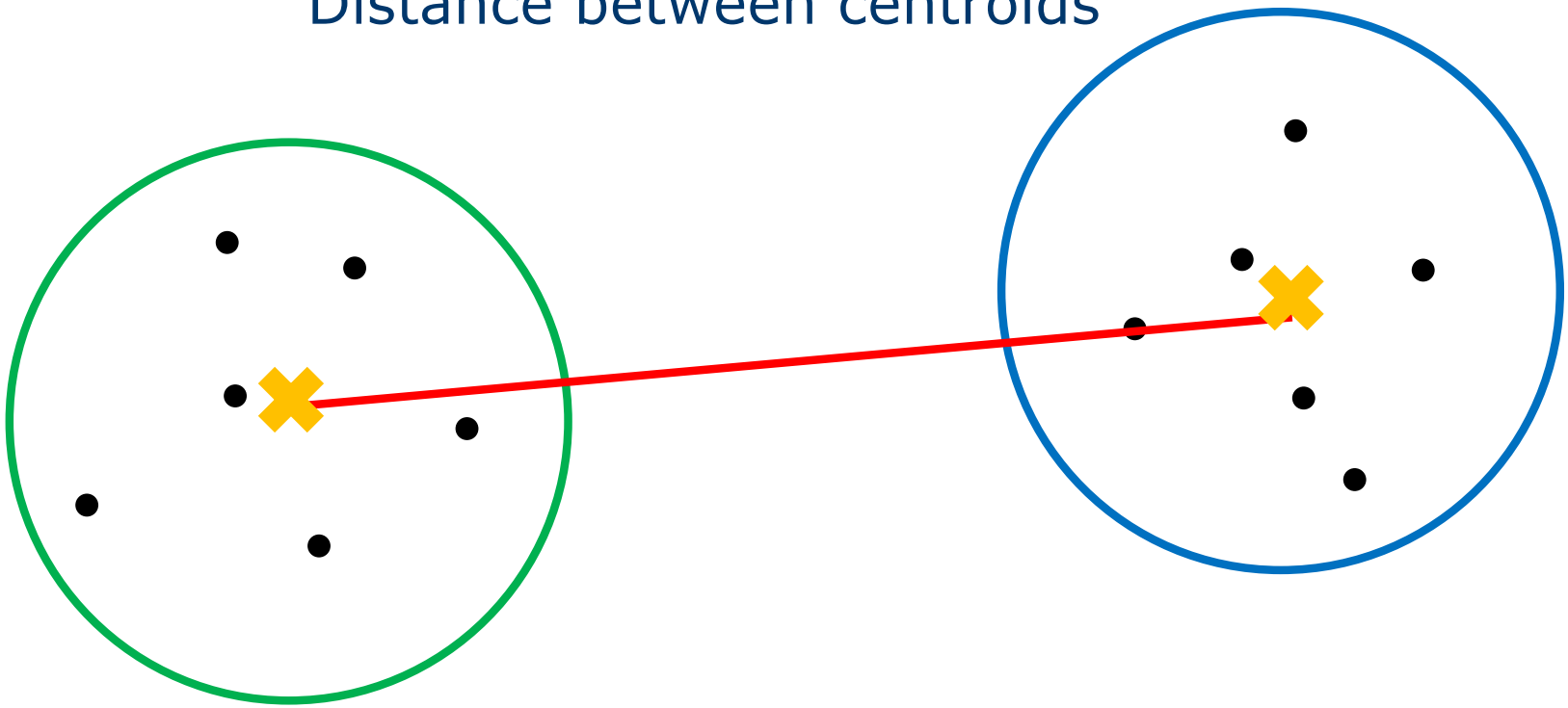


- Like average linkage, but computes the average over squared distances

Agglomerative hierarchical clustering

How to compute similarity between two clusters?

Distance between centroids



Agglomerative hierarchical clustering

Time complexity: $O(N^3)$

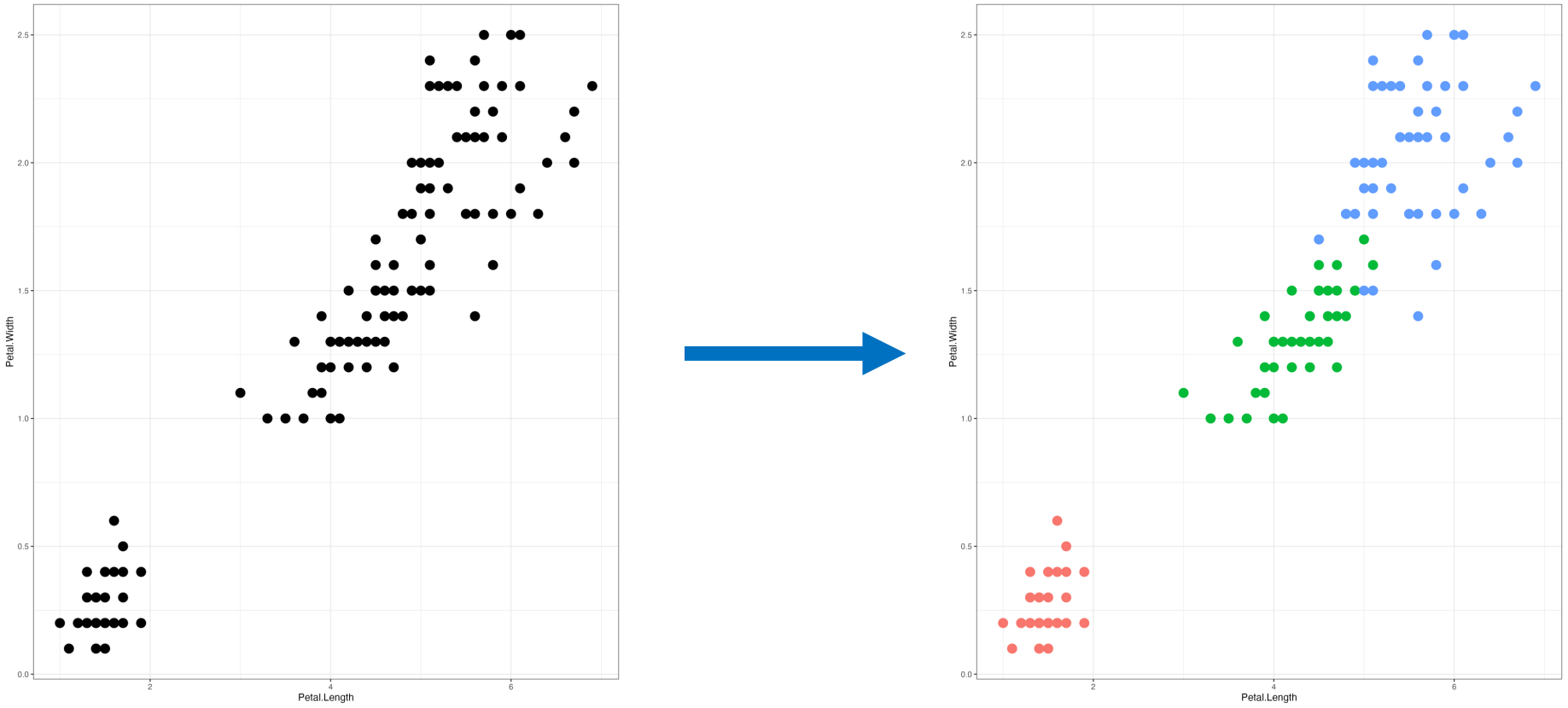
- $O(N^2)$ pick the most similar clusters
- $O(N)$ steps

Space complexity: $O(N^2)$

k-means clustering

k-means clustering

- Input: $T = \{x_1, x_2, x_3, \ldots, x_N\}$; k – number of clusters
- Output: $\{y_1, y_2, y_3, \ldots, y_N\}$



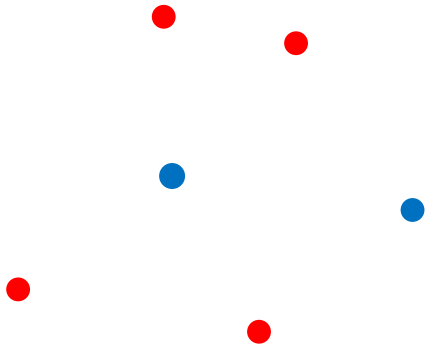
k-means clustering

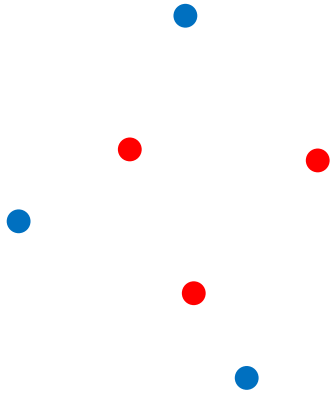
Algorithm:

- Initialize clusters randomly $C_1, C_2, C_3, \dots, C_k$, where $\bigcup_{i=1}^{k} C_i = T$ and $C_i \cup C_j = \emptyset$; $C_i = \{x_{i1}, x_{i2}, x_{i3}, \dots, x_{im_i}\}$, where $m_i \leq N, \sum m_i = N, i \in \{1, \dots, k\}$
- Repeat:
  1. Compute the cluster centres $\mu_i = \frac{1}{N_i} \sum_{j=1}^{m_i} x_{ij}$
  2. Reassign each point to the cluster, to which it has smallest Euclidean distance: $y_n = argmin_k \|x_n - \mu_i\|^2$, where $n \in \{1, \dots, N\}, i \in \{1, \dots, k\}$
  3. If $C_i = C_i'$ for $\forall i$ stop or the change in clusters is insignificant

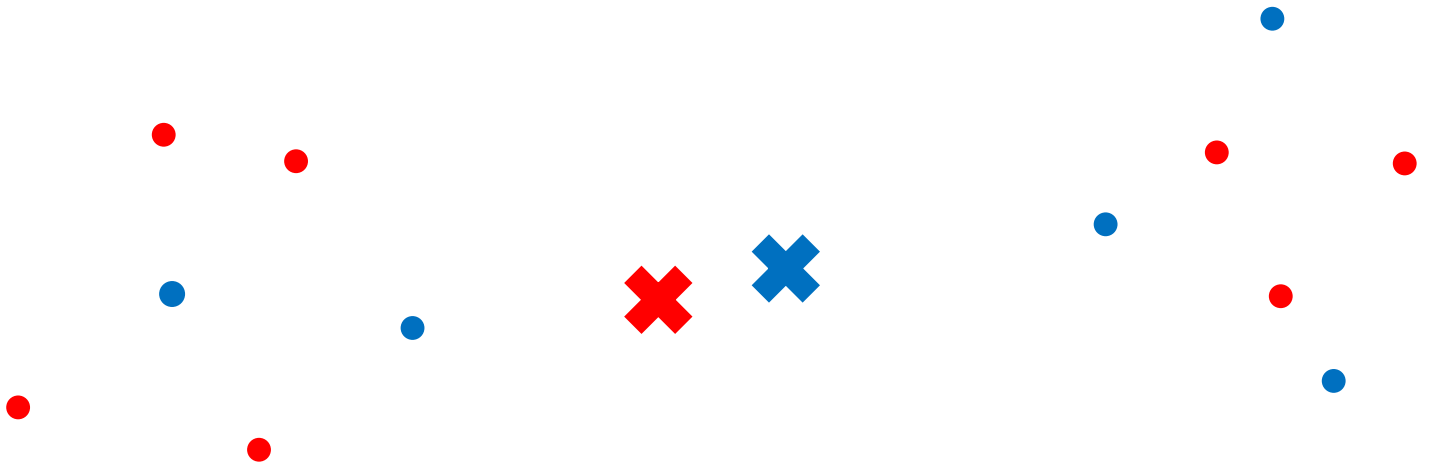
k-means clustering

k-means clustering

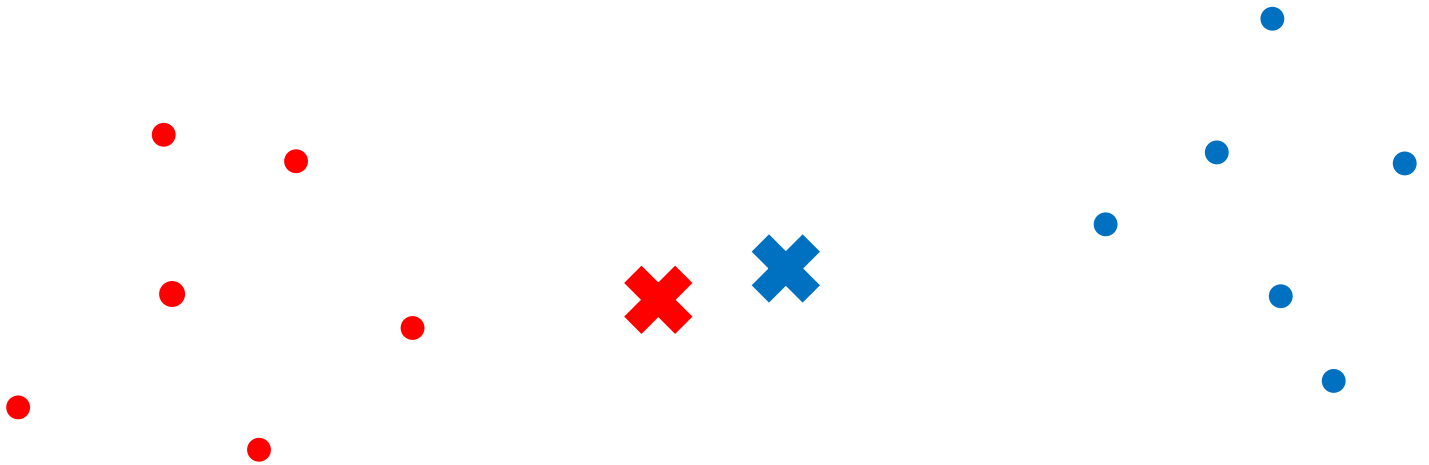




k-means clustering



k-means clustering



k-means clustering

k-means clustering

- Criteria: $\mathrm{J}(\mu_1, \ldots, \mu_k) = \sum_{C_i} \sum_{x_j \in C_i} \left\| x_j - \mu_i \right\|^2$
- The value of J is monotonically non-increasing with each step.
- Each iteration improves the result.
- The algorithm stops after the threshold on J is reached or the maximum number of iterations is reached
- **The algorithm convergence is extremely dependent on initial choice of clusters, noise and outliers**
- Standardization of data is recommended (0 mean, and 1 standard deviation)
- You need to pick k beforehand
- Time complexity: $O(NkI)$, where I is number of iterations

k-means clustering

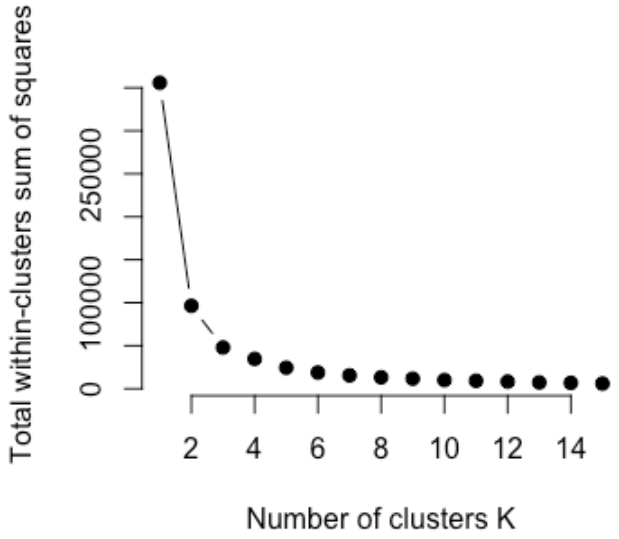
How to select the number of clusters k?

k-means clustering

How to select the number of clusters k?

Elbow Method

- Uses J
- Perform k-means for different number of clusters k
- Select the final k based on the "elbow in graph



k-means clustering

How to select the number of clusters k?

Average silhouette score

- Computes the "homogeneity" of clusters
- Silhouette score for one sample:

$$s(x) = \frac{b(x) - a(x)}{\max\{a(x), b(x)\}}$$

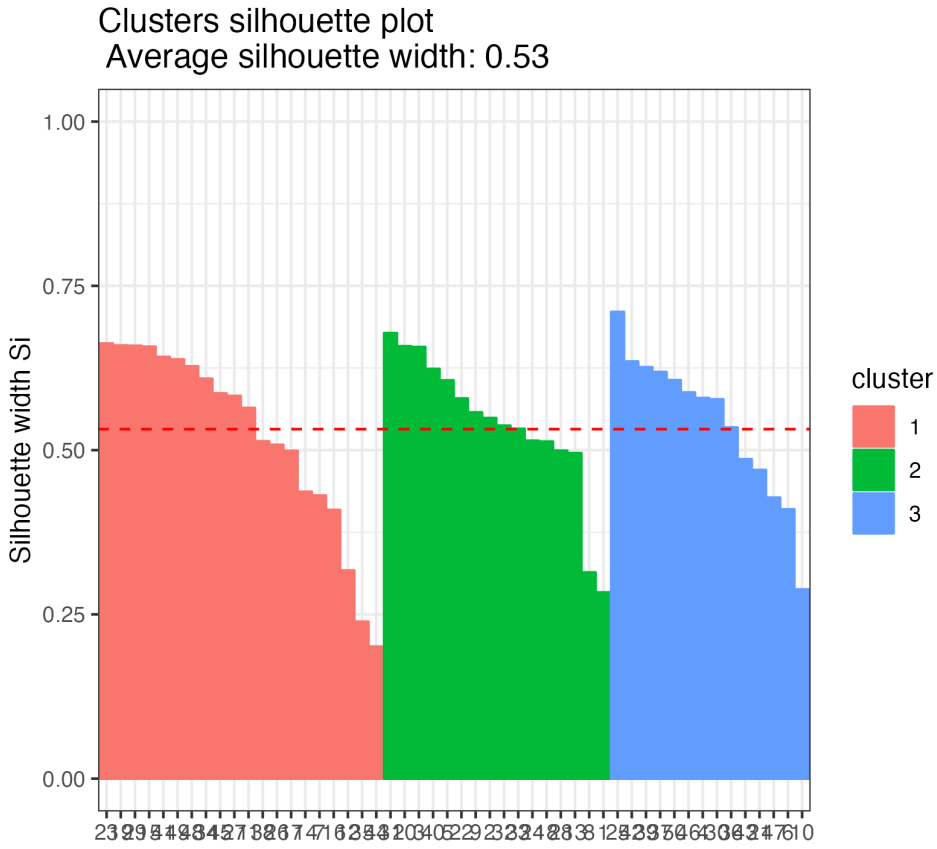
, where $a(x)$ is the average distance between x and all other points in its cluster and $b(x)$ is minimum average distance from x to the points in other clusters

- Range $\langle -1,1 \rangle$ where 1 means compact clusters, -1 is opposite and values around 0 represents overlapping clusters

k-means clustering

How to select the number of clusters k?

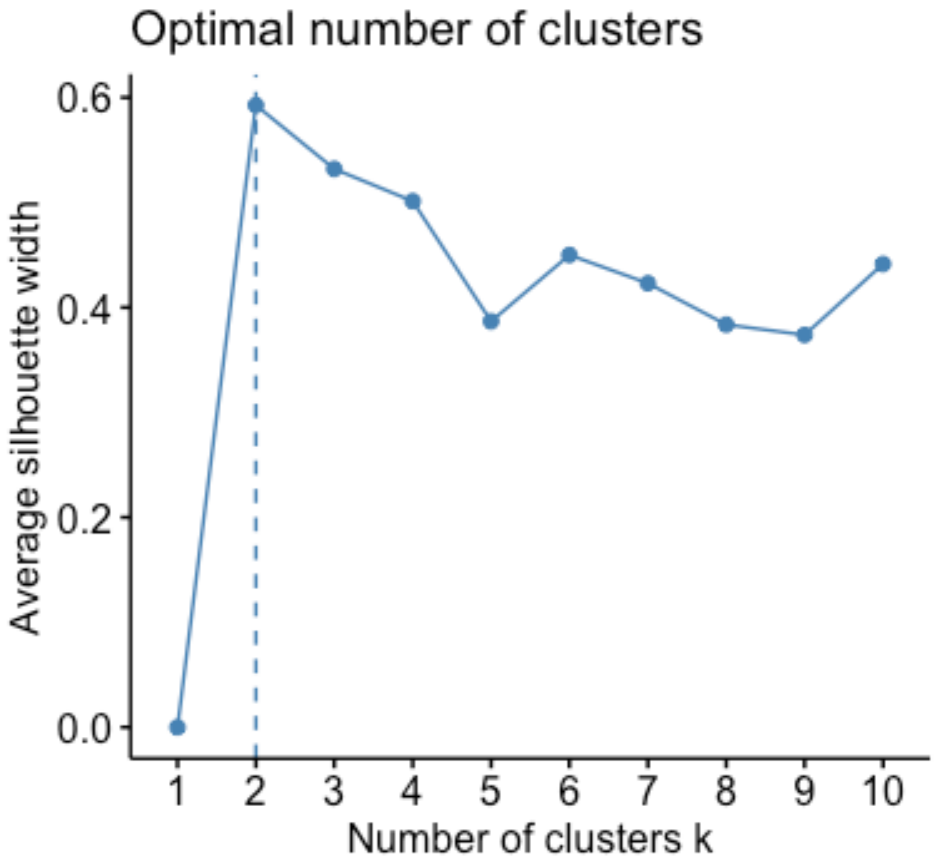
Average silhouette score



k-means clustering

How to select the number of clusters k?

Average silhouette score



Questions?