

CEE 616: Probabilistic Machine Learning
M5 Unsupervised Learning:
L5c: Clustering

Jimi Oke

UMassAmherst

College of Engineering

Tue, Dec 9, 2025

Outline

① Introduction

② Similarity measures

③ K-means clustering

④ Mixture models

⑤ Hierarchical clustering

⑥ DBSCAN

⑦ Silhouette

⑧ Summary

What is Clustering?

What is Clustering?

- Exploratory technique to discover useful relationships in data

What is Clustering?

- Exploratory technique to discover useful relationships in data
- Can also be used for classification

What is Clustering?

- Exploratory technique to discover useful relationships in data
- Can also be used for classification
- Clustering means grouping n observations into homogeneous partitions

What is Clustering?

- Exploratory technique to discover useful relationships in data
- Can also be used for classification
- Clustering means grouping n observations into homogeneous partitions
- There is no dependent variable, y (*unsupervised learning*)

What is Clustering?

- Exploratory technique to discover useful relationships in data
- Can also be used for classification
- Clustering means grouping n observations into homogeneous partitions
- There is no dependent variable, y (*unsupervised learning*)
- Observations \mathbf{x}_j are grouped based on similarity

What is Clustering?

- Exploratory technique to discover useful relationships in data
- Can also be used for classification
- Clustering means grouping n observations into homogeneous partitions
- There is no dependent variable, y (*unsupervised learning*)
- Observations \mathbf{x}_j are grouped based on similarity

What is Clustering?

- Exploratory technique to discover useful relationships in data
- Can also be used for classification
- Clustering means grouping n observations into homogeneous partitions
- There is no dependent variable, y (*unsupervised learning*)
- Observations \mathbf{x}_j are grouped based on similarity
- Objective:

What is Clustering?

- Exploratory technique to discover useful relationships in data
 - Can also be used for classification
 - Clustering means grouping n observations into homogeneous partitions
 - There is no dependent variable, y (*unsupervised learning*)
 - Observations \mathbf{x}_j are grouped based on similarity
-
- Objective:
 - high similarity between items that belong to the same cluster

What is Clustering?

- Exploratory technique to discover useful relationships in data
 - Can also be used for classification
 - Clustering means grouping n observations into homogeneous partitions
 - There is no dependent variable, y (*unsupervised learning*)
 - Observations x_j are grouped based on similarity
-
- Objective:
 - high similarity between items that belong to the same cluster
 - low similarity (high separation) between different clusters

What is Clustering?

- Exploratory technique to discover useful relationships in data
 - Can also be used for classification
 - Clustering means grouping n observations into homogeneous partitions
 - There is no dependent variable, y (*unsupervised learning*)
 - Observations x_j are grouped based on similarity
-
- Objective:
 - high similarity between items that belong to the same cluster
 - low similarity (high separation) between different clusters

Clustering Application: Marketing

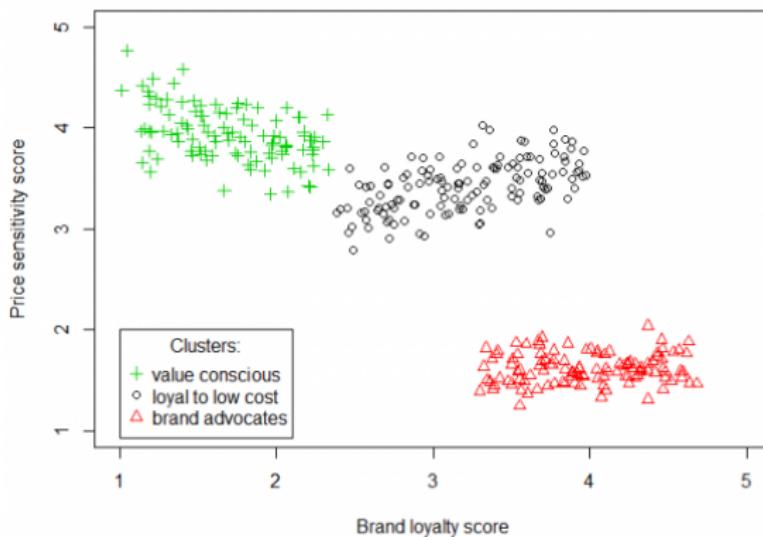
Clustering Application: Marketing

- Customer segmentation based on brand loyalty and price sensitivity scores.

Source: <http://www.select-statistics.co.uk/>

Clustering Application: Marketing

- Customer segmentation based on brand loyalty and price sensitivity scores.



Source: <http://www.select-statistics.co.uk/>

Similarity Measures

- How similar are two observations?
 - Geographical distance
 - Vehicle color
 - Vehicle type
 - Vehicle brand
 - Engine type
 - Engine power
 - ...



Figure: Vehicles as items for cluster analysis

Similarity measures: numerical/quantitative data

Similarity measures: numerical/quantitative data

Comparing two vectors, \mathbf{x}_i and \mathbf{x}_k , with p variables/features:

Similarity measures: numerical/quantitative data

Comparing two vectors, \mathbf{x}_i and \mathbf{x}_k , with p variables/features:

- **Euclidean distance**

Similarity measures: numerical/quantitative data

Comparing two vectors, \mathbf{x}_i and \mathbf{x}_k , with p variables/features:

- **Euclidean distance**

$$d(\mathbf{x}_i, \mathbf{x}_k) =$$

Similarity measures: numerical/quantitative data

Comparing two vectors, \mathbf{x}_i and \mathbf{x}_k , with p variables/features:

- **Euclidean distance**

$$d(\mathbf{x}_i, \mathbf{x}_k) = \sqrt{\sum_{j=1}^p (x_{ij} - x_{kj})^2} \quad (1)$$

Similarity measures: numerical/quantitative data

Comparing two vectors, \mathbf{x}_i and \mathbf{x}_k , with p variables/features:

- **Euclidean distance**

$$d(\mathbf{x}_i, \mathbf{x}_k) = \sqrt{\sum_{j=1}^p (x_{ij} - x_{kj})^2} \quad (1)$$

- **Manhattan distance**

Similarity measures: numerical/quantitative data

Comparing two vectors, \mathbf{x}_i and \mathbf{x}_k , with p variables/features:

- **Euclidean distance**

$$d(\mathbf{x}_i, \mathbf{x}_k) = \sqrt{\sum_{j=1}^p (x_{ij} - x_{kj})^2} \quad (1)$$

- **Manhattan distance**

$$d(\mathbf{x}_i, \mathbf{x}_k) =$$

Similarity measures: numerical/quantitative data

Comparing two vectors, \mathbf{x}_i and \mathbf{x}_k , with p variables/features:

- **Euclidean distance**

$$d(\mathbf{x}_i, \mathbf{x}_k) = \sqrt{\sum_{j=1}^p (x_{ij} - x_{kj})^2} \quad (1)$$

- **Manhattan distance**

$$d(\mathbf{x}_i, \mathbf{x}_k) = \sum_{j=1}^p |x_{ij} - x_{kj}| \quad (2)$$

Similarity measures: numerical/quantitative data

Comparing two vectors, \mathbf{x}_i and \mathbf{x}_k , with p variables/features:

- **Euclidean distance**

$$d(\mathbf{x}_i, \mathbf{x}_k) = \sqrt{\sum_{j=1}^p (x_{ij} - x_{kj})^2} \quad (1)$$

- **Manhattan distance**

$$d(\mathbf{x}_i, \mathbf{x}_k) = \sum_{j=1}^p |x_{ij} - x_{kj}| \quad (2)$$

- **Minkowski distance**

Similarity measures: numerical/quantitative data

Comparing two vectors, \mathbf{x}_i and \mathbf{x}_k , with p variables/features:

- **Euclidean distance**

$$d(\mathbf{x}_i, \mathbf{x}_k) = \sqrt{\sum_{j=1}^p (x_{ij} - x_{kj})^2} \quad (1)$$

- **Manhattan distance**

$$d(\mathbf{x}_i, \mathbf{x}_k) = \sum_{j=1}^p |x_{ij} - x_{kj}| \quad (2)$$

- **Minkowski distance**

$$d(\mathbf{x}_i, \mathbf{x}_k) =$$

Similarity measures: numerical/quantitative data

Comparing two vectors, \mathbf{x}_i and \mathbf{x}_k , with p variables/features:

- **Euclidean distance**

$$d(\mathbf{x}_i, \mathbf{x}_k) = \sqrt{\sum_{j=1}^p (x_{ij} - x_{kj})^2} \quad (1)$$

- **Manhattan distance**

$$d(\mathbf{x}_i, \mathbf{x}_k) = \sum_{j=1}^p |x_{ij} - x_{kj}| \quad (2)$$

- **Minkowski distance**

$$d(\mathbf{x}_i, \mathbf{x}_k) = \left[\sum_{j=1}^p |x_{ij} - x_{kj}|^m \right]^{1/m} \quad (3)$$

Similarity measures: Categorical data

Similarity measures: Categorical data

- Based on **presence or absence** of certain characteristics (*binary variables*).

Similarity measures: Categorical data

- Based on **presence or absence** of certain characteristics (*binary variables*).

Similarity measures: Categorical data

- Based on **presence or absence** of certain characteristics (*binary variables*).

		Variables				
		1	2	3	4	5
Item i	1	0	0	1	1	
	Item k	1	1	0	1	0

Similarity measures: Categorical data

- Based on **presence or absence** of certain characteristics (*binary variables*).

		Variables				
		1	2	3	4	5
Item i	1	0	0	1	1	
	Item k	1	1	0	1	0

- Contingency table: variable matches and mismatches between observations (items) i and k .

Similarity measures: Categorical data

- Based on **presence or absence** of certain characteristics (*binary variables*).

		Variables				
		1	2	3	4	5
Item i	1	0	0	1	1	
	Item k	1	1	0	1	0

- Contingency table: variable matches and mismatches between observations (items) i and k .

Source: Johnson & Wichern

Similarity measures: Categorical data

- Based on **presence or absence** of certain characteristics (*binary variables*).

		Variables				
		1	2	3	4	5
Item i	1	0	0	1	1	
	Item k	1	1	0	1	0

- Contingency table: variable matches and mismatches between observations (items) i and k .

		Item k		Totals	
		1	0		
Item i	1	a	b	$a + b$	
	0	c	d	$c + d$	
Totals		$a + c$	$b + d$	$p = a + b + c + d$	

Source: Johnson & Wichern

Similarity measures: Categorical data

- Based on **presence or absence** of certain characteristics (*binary variables*).

		Variables				
		1	2	3	4	5
Item i	1	0	0	1	1	
	Item k	1	1	0	1	0

- Contingency table: variable matches and mismatches between observations (items) i and k .

		Item k		Totals	
		1	0		
Item i	1	a	b	$a + b$	
	0	c	d	$c + d$	
Totals		$a + c$	$b + d$	$p = a + b + c + d$	

Source: Johnson & Wichern

- Various **similarity coefficients** can be calculated from these frequencies:

Similarity measures: Categorical data

- Based on **presence or absence** of certain characteristics (*binary variables*).

		Variables				
		1	2	3	4	5
Item i	1	0	0	1	1	
	Item k	1	1	0	1	0

- Contingency table: variable matches and mismatches between observations (items) i and k .

		Item k		Totals	
		1	0		
Item i	1	a	b	$a + b$	
	0	c	d	$c + d$	
Totals		$a + c$	$b + d$	$p = a + b + c + d$	

Source: Johnson & Wichern

- Various **similarity coefficients** can be calculated from these frequencies:

- examples $\frac{a+d}{p}$, $\frac{a}{p}$

Similarity measures: Categorical data

- Based on **presence or absence** of certain characteristics (*binary variables*).

		Variables				
		1	2	3	4	5
Item <i>i</i>	1	0	0	1	1	
	Item <i>k</i>	1	1	0	1	0

- Contingency table: variable matches and mismatches between observations (items) *i* and *k*.

		Item <i>k</i>		Totals		(12-7)	
		1	0				
Item <i>i</i>	1	<i>a</i>	<i>b</i>	<i>a</i> + <i>b</i>			
	0	<i>c</i>	<i>d</i>	<i>c</i> + <i>d</i>			
Totals		<i>a</i> + <i>c</i>	<i>b</i> + <i>d</i>	<i>p</i> = <i>a</i> + <i>b</i> + <i>c</i> + <i>d</i>			

Source: Johnson & Wichern

- Various **similarity coefficients** can be calculated from these frequencies:
 - examples $\frac{a+d}{p}$, $\frac{a}{p}$
 - Distance can be constructed from similarity measures. Under some hypothesis, $d_{ik} = \sqrt{2(1 - s_{ik})}$, where s_{ik} is the similarity between samples *i* and *k*.

Similarity measures: Categorical data

- Based on **presence or absence** of certain characteristics (*binary variables*).

		Variables				
		1	2	3	4	5
Item <i>i</i>	1	0	0	1	1	
	Item <i>k</i>	1	1	0	1	0

- Contingency table: variable matches and mismatches between observations (items) *i* and *k*.

		Item <i>k</i>		Totals		(12-7)
		1	0	$a + b$	$c + d$	
Item <i>i</i>	1	a	b	$a + b$	$c + d$	
	0	c	d			
Totals		$a + c$	$b + d$	$p = a + b + c + d$		

Source: Johnson & Wichern

- Various **similarity coefficients** can be calculated from these frequencies:
 - examples $\frac{a+d}{p}$, $\frac{a}{p}$
 - Distance can be constructed from similarity measures. Under some hypothesis, $d_{ik} = \sqrt{2(1 - s_{ik})}$, where s_{ik} is the similarity between samples *i* and *k*.

Some notes about distance metrics

Some notes about distance metrics

- Care should be taken with multiple dimensions and varying scales

Some notes about distance metrics

- Care should be taken with multiple dimensions and varying scales
 - Scaling/normalization typically leads to better results

Some notes about distance metrics

- Care should be taken with multiple dimensions and varying scales
 - Scaling/normalization typically leads to better results
 - E.g. Min-max scaling:

Some notes about distance metrics

- Care should be taken with multiple dimensions and varying scales
 - Scaling/normalization typically leads to better results
 - E.g. Min-max scaling: $x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$, $x_{new} \in [0, 1]$

Some notes about distance metrics

- Care should be taken with multiple dimensions and varying scales
 - Scaling/normalization typically leads to better results
 - E.g. Min-max scaling: $x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$, $x_{new} \in [0, 1]$
- Choice of similarity measure:

Some notes about distance metrics

- Care should be taken with multiple dimensions and varying scales
 - Scaling/normalization typically leads to better results
 - E.g. Min-max scaling: $x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$, $x_{new} \in [0, 1]$
- Choice of similarity measure:
 - May lead to **different groupings**

Some notes about distance metrics

- Care should be taken with multiple dimensions and varying scales
 - Scaling/normalization typically leads to better results
 - E.g. Min-max scaling: $x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$, $x_{new} \in [0, 1]$
- Choice of similarity measure:
 - May lead to **different groupings**
 - Subjective and domain-dependent

Some notes about distance metrics

- Care should be taken with multiple dimensions and varying scales
 - Scaling/normalization typically leads to better results
 - E.g. Min-max scaling: $x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$, $x_{new} \in [0, 1]$
- Choice of similarity measure:
 - May lead to **different groupings**
 - Subjective and domain-dependent
 - Dependent on the **variable type** (discrete, continuous, binary)

Some notes about distance metrics

- Care should be taken with multiple dimensions and varying scales
 - Scaling/normalization typically leads to better results
 - E.g. Min-max scaling: $x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$, $x_{new} \in [0, 1]$
- Choice of similarity measure:
 - May lead to **different groupings**
 - Subjective and domain-dependent
 - Dependent on the **variable type** (discrete, continuous, binary)
 - Dependent on the **scale of measurement**

Some notes about distance metrics

- Care should be taken with multiple dimensions and varying scales
 - Scaling/normalization typically leads to better results
 - E.g. Min-max scaling: $x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$, $x_{new} \in [0, 1]$
- Choice of similarity measure:
 - May lead to **different groupings**
 - Subjective and domain-dependent
 - Dependent on the **variable type** (discrete, continuous, binary)
 - Dependent on the **scale of measurement**
- For items/entities, similarity is typically based on some measure of distance

Some notes about distance metrics

- Care should be taken with multiple dimensions and varying scales
 - Scaling/normalization typically leads to better results
 - E.g. Min-max scaling: $x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$, $x_{new} \in [0, 1]$
- Choice of similarity measure:
 - May lead to **different groupings**
 - Subjective and domain-dependent
 - Dependent on the **variable type** (discrete, continuous, binary)
 - Dependent on the **scale of measurement**
- For items/entities, similarity is typically based on some measure of distance
- For variables, similarity is based on statistical correlation

Some notes about distance metrics

- Care should be taken with multiple dimensions and varying scales
 - Scaling/normalization typically leads to better results
 - E.g. Min-max scaling: $x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$, $x_{new} \in [0, 1]$
- Choice of similarity measure:
 - May lead to **different groupings**
 - Subjective and domain-dependent
 - Dependent on the **variable type** (discrete, continuous, binary)
 - Dependent on the **scale of measurement**
- For items/entities, similarity is typically based on some measure of distance
- For variables, similarity is based on statistical correlation

K-means Clustering: Definitions

K-means Clustering: Definitions

K-means Clustering: Definitions

- K : Number of clusters. Design parameter to decide in advance.

K-means Clustering: Definitions

- K : Number of clusters. Design parameter to decide in advance.
- Cluster **centroid**: mean of observations assigned to cluster C_k :

$$\bar{\mathbf{x}}_k \triangleq \frac{1}{|C_k|} \sum_{\mathbf{x} \in C_k} \mathbf{x}$$

K-means Clustering: Definitions

- K : Number of clusters. Design parameter to decide in advance.
- Cluster **centroid**: mean of observations assigned to cluster C_k :

$$\bar{\mathbf{x}}_k \triangleq \frac{1}{|C_k|} \sum_{\mathbf{x} \in C_k} \mathbf{x}$$

- Within cluster variation of the k -th cluster

$$W(C_k) \triangleq \frac{1}{|C_k|} \sum_{\mathbf{x} \in C_k} d(\mathbf{x}, \bar{\mathbf{x}}_k)^2$$

K-means Clustering: Definitions

- K : Number of clusters. Design parameter to decide in advance.
- Cluster **centroid**: mean of observations assigned to cluster C_k :

$$\bar{\mathbf{x}}_k \triangleq \frac{1}{|C_k|} \sum_{\mathbf{x} \in C_k} \mathbf{x}$$

- Within cluster variation of the k -th cluster

$$W(C_k) \triangleq \frac{1}{|C_k|} \sum_{\mathbf{x} \in C_k} d(\mathbf{x}, \bar{\mathbf{x}}_k)^2$$

K-means Clustering: Definitions

- K : Number of clusters. Design parameter to decide in advance.
- Cluster **centroid**: mean of observations assigned to cluster C_k :

$$\bar{\mathbf{x}}_k \triangleq \frac{1}{|C_k|} \sum_{\mathbf{x} \in C_k} \mathbf{x}$$

- Within cluster variation of the k -th cluster

$$W(C_k) \triangleq \frac{1}{|C_k|} \sum_{\mathbf{x} \in C_k} d(\mathbf{x}, \bar{\mathbf{x}}_k)^2$$

K-means Clustering: Definitions

- K : Number of clusters. Design parameter to decide in advance.
- Cluster **centroid**: mean of observations assigned to cluster C_k :

$$\bar{\mathbf{x}}_k \triangleq \frac{1}{|C_k|} \sum_{\mathbf{x} \in C_k} \mathbf{x}$$

- Within cluster variation of the k -th cluster

$$W(C_k) \triangleq \frac{1}{|C_k|} \sum_{\mathbf{x} \in C_k} d(\mathbf{x}, \bar{\mathbf{x}}_k)^2$$

K-means Clustering: Definitions

- K : Number of clusters. Design parameter to decide in advance.
- Cluster **centroid**: mean of observations assigned to cluster C_k :

$$\bar{\mathbf{x}}_k \triangleq \frac{1}{|C_k|} \sum_{\mathbf{x} \in C_k} \mathbf{x}$$

- Within cluster variation of the k -th cluster

$$W(C_k) \triangleq \frac{1}{|C_k|} \sum_{\mathbf{x} \in C_k} d(\mathbf{x}, \bar{\mathbf{x}}_k)^2$$

- Usually $d(\mathbf{x}, \bar{\mathbf{x}}_k)^2 = \sum_{j=1}^r (x_j - \bar{x}_{kj})^2$
- Goal: minimize total variation

$$\min \sum_{k=1}^K W(C_k)$$

K-means Clustering: Definitions

- K : Number of clusters. Design parameter to decide in advance.
- Cluster **centroid**: mean of observations assigned to cluster C_k :

$$\bar{\mathbf{x}}_k \triangleq \frac{1}{|C_k|} \sum_{\mathbf{x} \in C_k} \mathbf{x}$$

- Within cluster variation of the k -th cluster

$$W(C_k) \triangleq \frac{1}{|C_k|} \sum_{\mathbf{x} \in C_k} d(\mathbf{x}, \bar{\mathbf{x}}_k)^2$$

- Usually $d(\mathbf{x}, \bar{\mathbf{x}}_k)^2 = \sum_{j=1}^r (x_j - \bar{x}_{kj})^2$
- Goal: minimize total variation

$$\min \sum_{k=1}^K W(C_k)$$

- \Rightarrow Assign \mathbf{x} to C_k with minimum $d(\mathbf{x}, \bar{\mathbf{x}}_k)$

K-means Clustering: Definitions

- K : Number of clusters. Design parameter to decide in advance.
- Cluster **centroid**: mean of observations assigned to cluster C_k :

$$\bar{\mathbf{x}}_k \triangleq \frac{1}{|C_k|} \sum_{\mathbf{x} \in C_k} \mathbf{x}$$

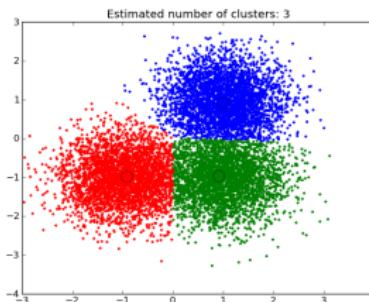
- Within cluster variation of the k -th cluster

$$W(C_k) \triangleq \frac{1}{|C_k|} \sum_{\mathbf{x} \in C_k} d(\mathbf{x}, \bar{\mathbf{x}}_k)^2$$

- Usually $d(\mathbf{x}, \bar{\mathbf{x}}_k)^2 = \sum_{j=1}^r (x_j - \bar{x}_{kj})^2$
- Goal: minimize total variation

$$\min \sum_{k=1}^K W(C_k)$$

- \Rightarrow Assign \mathbf{x} to C_k with minimum $d(\mathbf{x}, \bar{\mathbf{x}}_k)$



K-means Clustering: Definitions

- K : Number of clusters. Design parameter to decide in advance.
- Cluster **centroid**: mean of observations assigned to cluster C_k :

$$\bar{\mathbf{x}}_k \triangleq \frac{1}{|C_k|} \sum_{\mathbf{x} \in C_k} \mathbf{x}$$

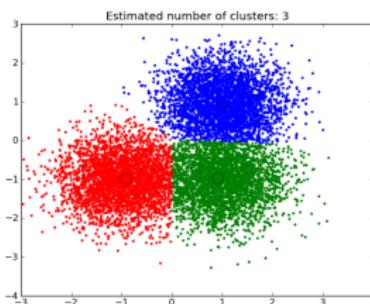
- Within cluster variation of the k -th cluster

$$W(C_k) \triangleq \frac{1}{|C_k|} \sum_{\mathbf{x} \in C_k} d(\mathbf{x}, \bar{\mathbf{x}}_k)^2$$

- Usually $d(\mathbf{x}, \bar{\mathbf{x}}_k)^2 = \sum_{j=1}^r (x_j - \bar{x}_{kj})^2$
- Goal: minimize total variation

$$\min \sum_{k=1}^K W(C_k)$$

- \Rightarrow Assign \mathbf{x} to C_k with minimum $d(\mathbf{x}, \bar{\mathbf{x}}_k)$



Source: www.scikit-learn.org

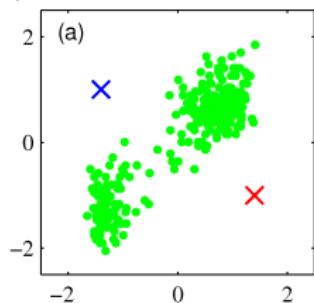
K-means Clustering: Illustration (a) - (d)

K-means Clustering: Illustration (a) - (d)

$$\min \sum_{k=1}^K \frac{1}{|C_k|} \sum_{x \in C_k} d(x, \bar{x}_k)^2$$

K-means Clustering: Illustration (a) - (d)

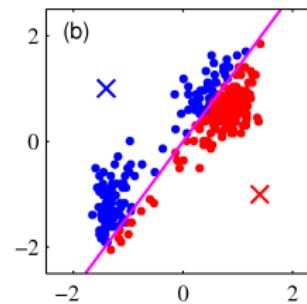
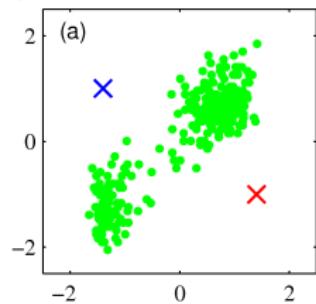
$$\min \sum_{k=1}^K \frac{1}{|C_k|} \sum_{x \in C_k} d(x, \bar{x}_k)^2$$



Source: Christopher M. Bishop, *Pattern Recognition and Machine Learning*

K-means Clustering: Illustration (a) - (d)

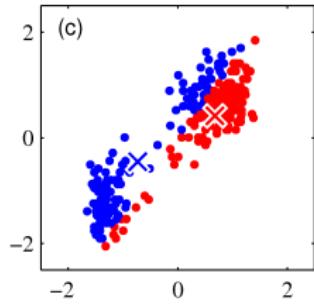
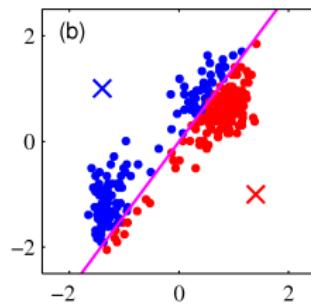
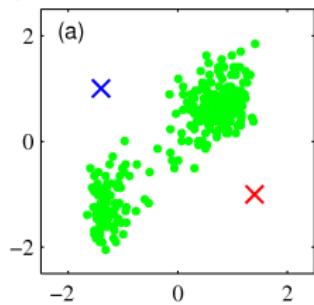
$$\min \sum_{k=1}^K \frac{1}{|C_k|} \sum_{x \in C_k} d(x, \bar{x}_k)^2$$



Source: Christopher M. Bishop, *Pattern Recognition and Machine Learning*

K-means Clustering: Illustration (a) - (d)

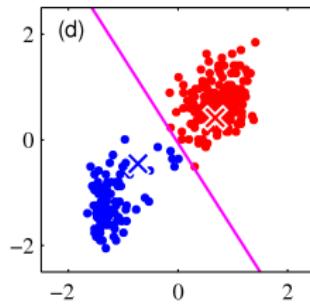
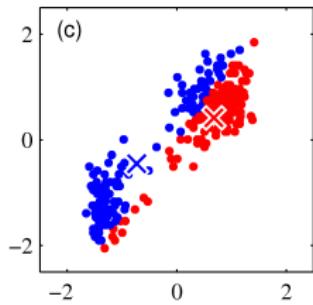
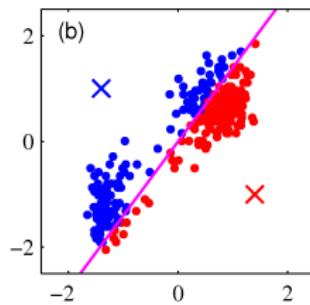
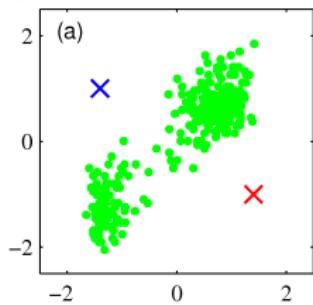
$$\min \sum_{k=1}^K \frac{1}{|C_k|} \sum_{x \in C_k} d(x, \bar{x}_k)^2$$



Source: Christopher M. Bishop, *Pattern Recognition and Machine Learning*

K-means Clustering: Illustration (a) - (d)

$$\min \sum_{k=1}^K \frac{1}{|C_k|} \sum_{x \in C_k} d(x, \bar{x}_k)^2$$



Source: Cristopher M. Bishop, *Pattern Recognition and Machine Learning*

K-means Clustering: Illustration (e) - (h)

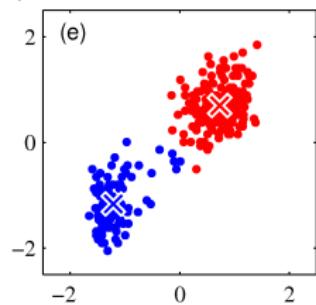
K-means Clustering: Illustration (e) - (h)

$$\min \sum_{k=1}^K \frac{1}{|C_k|} \sum_{x \in C_k} d(x, \bar{x}_k)^2$$

Source: Christopher M. Bishop, *Pattern Recognition and Machine Learning*

K-means Clustering: Illustration (e) - (h)

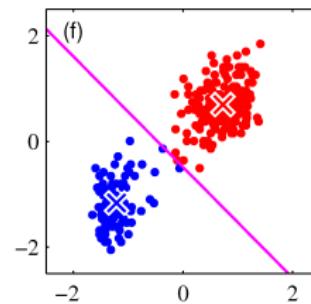
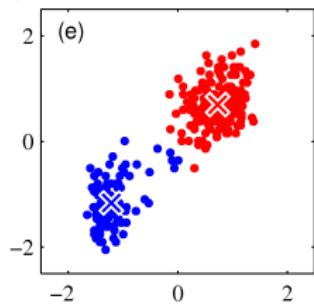
$$\min \sum_{k=1}^K \frac{1}{|C_k|} \sum_{x \in C_k} d(x, \bar{x}_k)^2$$



Source: Christopher M. Bishop, *Pattern Recognition and Machine Learning*

K-means Clustering: Illustration (e) - (h)

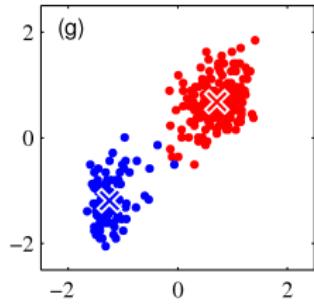
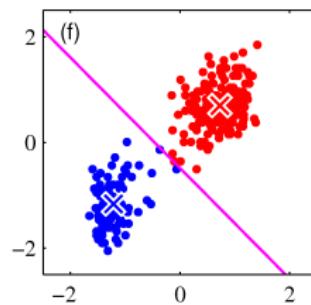
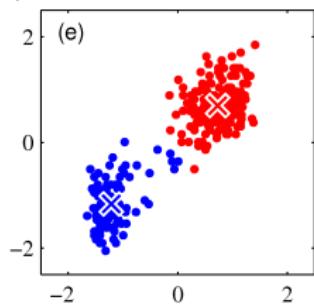
$$\min \sum_{k=1}^K \frac{1}{|C_k|} \sum_{x \in C_k} d(x, \bar{x}_k)^2$$



Source: Christopher M. Bishop, *Pattern Recognition and Machine Learning*

K-means Clustering: Illustration (e) - (h)

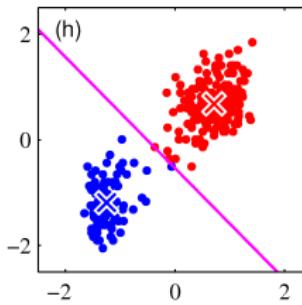
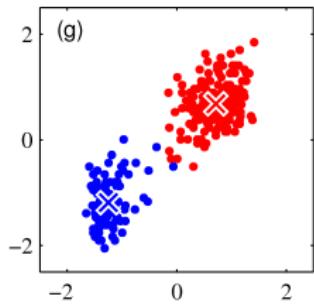
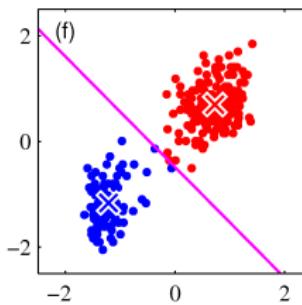
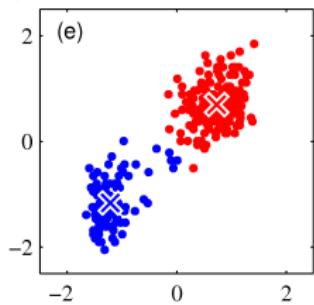
$$\min \sum_{k=1}^K \frac{1}{|C_k|} \sum_{x \in C_k} d(x, \bar{x}_k)^2$$



Source: Christopher M. Bishop, *Pattern Recognition and Machine Learning*

K-means Clustering: Illustration (e) - (h)

$$\min \sum_{k=1}^K \frac{1}{|C_k|} \sum_{x \in C_k} d(x, \bar{x}_k)^2$$



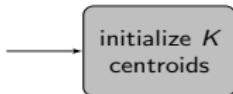
Source: Cristopher M. Bishop, *Pattern Recognition and Machine Learning*

K-means Clustering: Algorithm

K-means Clustering: Algorithm

initialize K
centroids

K-means Clustering: Algorithm



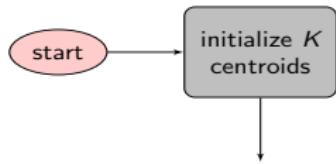
- At each Assign and Update, the total W decreases until *convergence*.
- The W at convergence depends on the initial centroids chosen (local minimum).

K-means Clustering: Algorithm



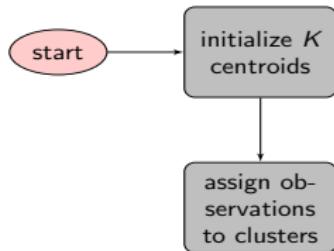
- At each Assign and Update, the total W decreases until *convergence*.
- The W at convergence depends on the initial centroids chosen (local minimum).
- Repeat the algorithm with different random initial centroids multiple times, and choose the clustering with the lowest W .

K-means Clustering: Algorithm



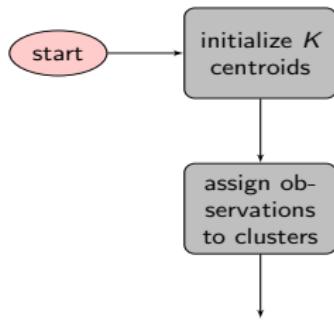
- At each Assign and Update, the total W decreases until *convergence*.
- The W at convergence depends on the initial centroids chosen (local minimum).
- Repeat the algorithm with different random initial centroids multiple times, and choose the clustering with the lowest W .

K-means Clustering: Algorithm



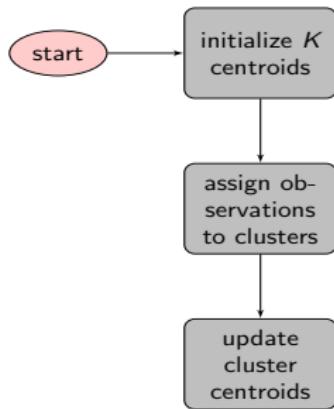
- At each Assign and Update, the total W decreases until *convergence*.
- The W at convergence depends on the initial centroids chosen (local minimum).
- Repeat the algorithm with different random initial centroids multiple times, and choose the clustering with the lowest W .

K-means Clustering: Algorithm



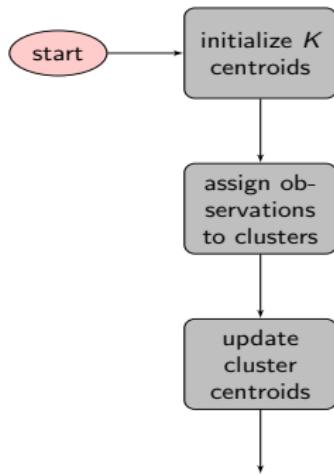
- At each Assign and Update, the total W decreases until *convergence*.
- The W at convergence depends on the initial centroids chosen (local minimum).
- Repeat the algorithm with different random initial centroids multiple times, and choose the clustering with the lowest W .

K-means Clustering: Algorithm



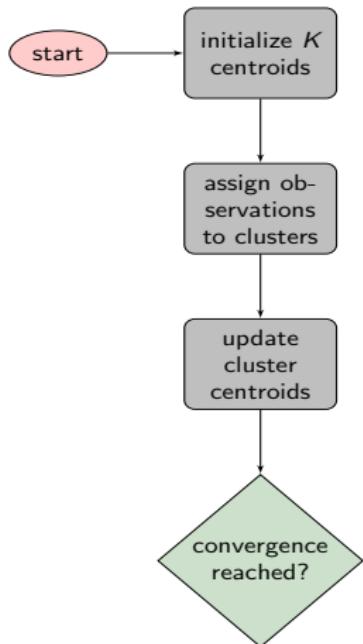
- At each Assign and Update, the total W decreases until *convergence*.
- The W at convergence depends on the initial centroids chosen (local minimum).
- Repeat the algorithm with different random initial centroids multiple times, and choose the clustering with the lowest W .

K-means Clustering: Algorithm



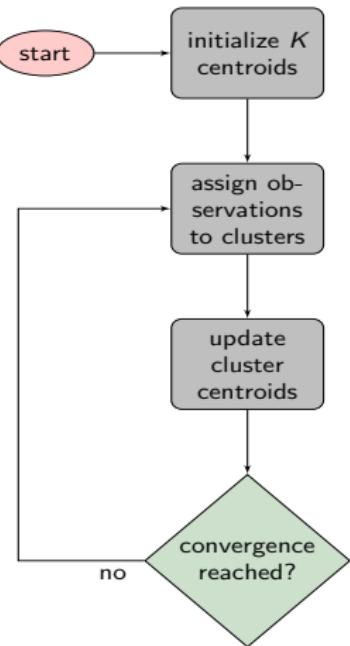
- At each Assign and Update, the total W decreases until *convergence*.
- The W at convergence depends on the initial centroids chosen (local minimum).
- Repeat the algorithm with different random initial centroids multiple times, and choose the clustering with the lowest W .

K-means Clustering: Algorithm



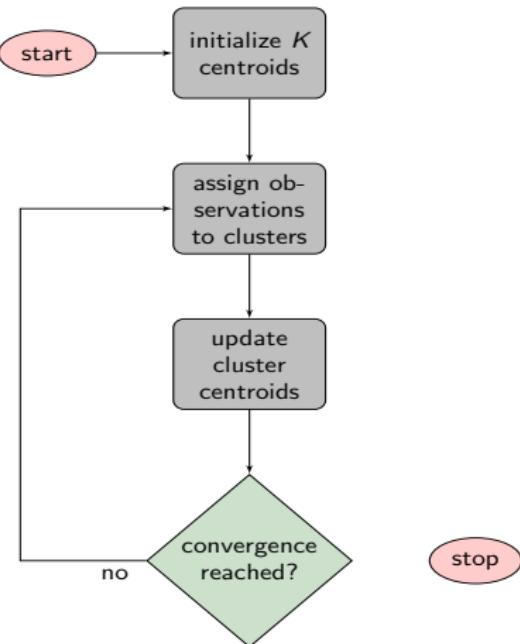
- At each Assign and Update, the total W decreases until *convergence*.
- The W at convergence depends on the initial centroids chosen (local minimum).
- Repeat the algorithm with different random initial centroids multiple times, and choose the clustering with the lowest W .

K-means Clustering: Algorithm



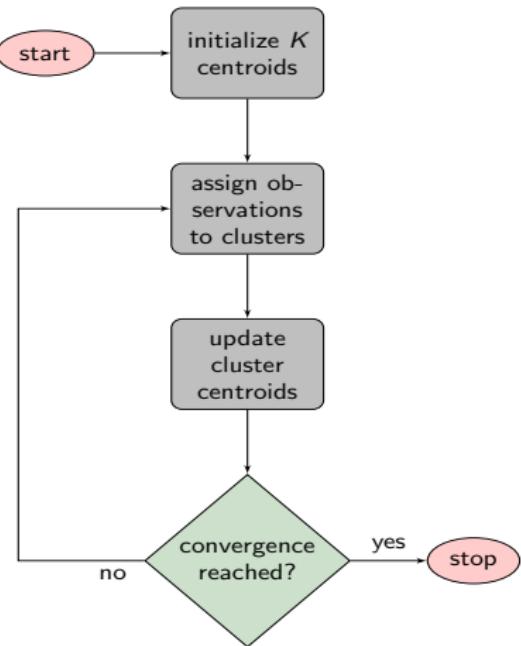
- At each Assign and Update, the total W decreases until *convergence*.
- The W at convergence depends on the initial centroids chosen (local minimum).
- Repeat the algorithm with different random initial centroids multiple times, and choose the clustering with the lowest W .

K-means Clustering: Algorithm



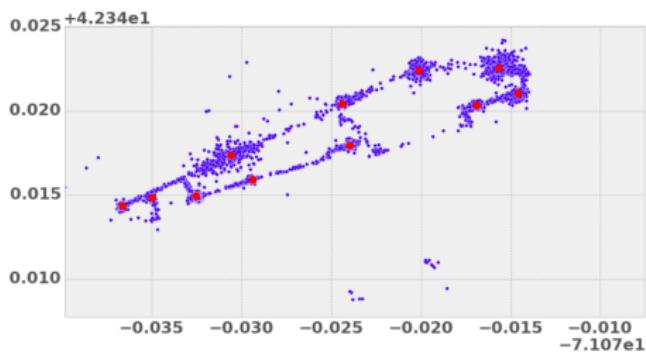
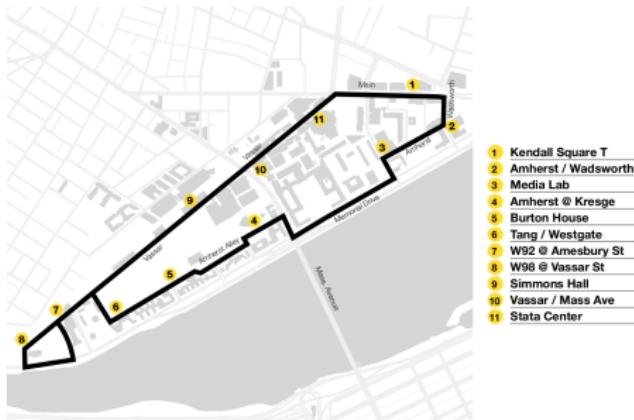
- At each Assign and Update, the total W decreases until *convergence*.
- The W at convergence depends on the initial centroids chosen (local minimum).
- Repeat the algorithm with different random initial centroids multiple times, and choose the clustering with the lowest W .

K-means Clustering: Algorithm



- At each Assign and Update, the total W decreases until *convergence*.
- The W at convergence depends on the initial centroids chosen (local minimum).
- Repeat the algorithm with different random initial centroids multiple times, and choose the clustering with the lowest W .

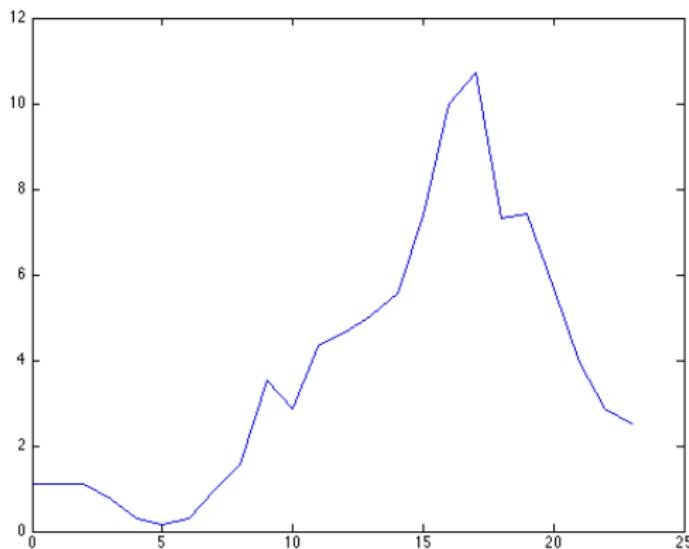
KMeans example: MIT tech shuttle



Purple dots: GPS points from buses; Red dots: centroids

Clustering Hubway rentals

- Comparing patterns
- Challenge: group stations according to similar demand patterns



Clustering Hubway rentals

Clustering Hubway rentals

Clustering Hubway rentals

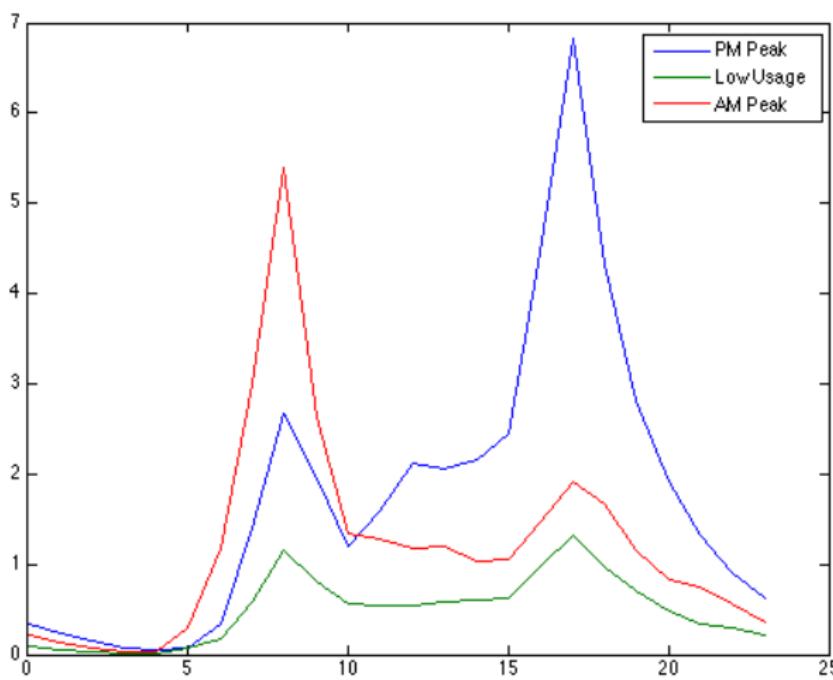
- Month of November 2013

Clustering Hubway rentals

- Month of November 2013
- Consider weekdays only

Clustering Hubway rentals

- Month of November 2013
- Consider weekdays only



Clustering Hubway rentals (cont.)

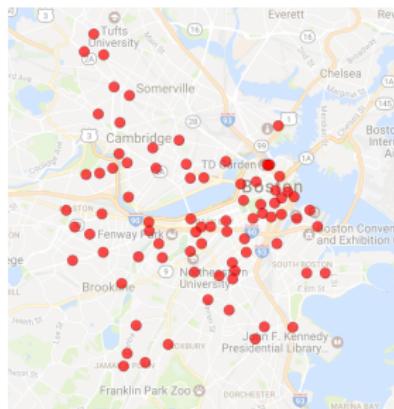
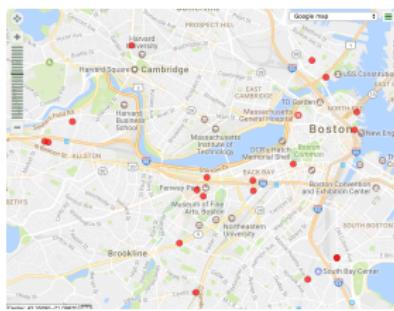
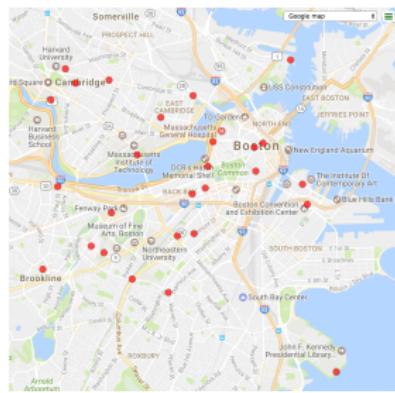


Figure: AM Peak, PM Peak, Low Usage docks

Choice of K

Which K would you choose?

- Having to pre-specify K is one limitation of the K -means approach

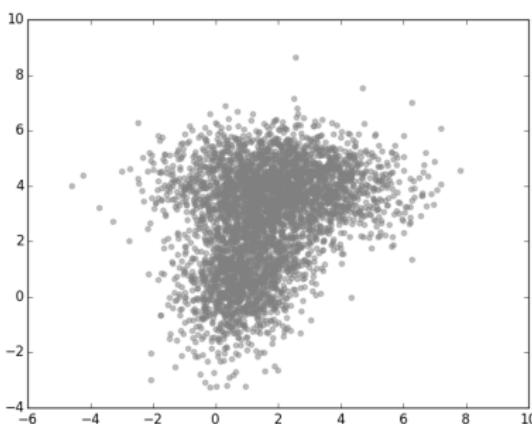
Choice of K

Which K would you choose?

- Having to pre-specify K is one limitation of the K -means approach
- However, several statistics can be used to choose the best K , e.g. gap statistic (Tibshirani et al., 2001; ESL p. 519)

Choice of K

Which K would you choose?



- Having to pre-specify K is one limitation of the K -means approach
- However, several statistics can be used to choose the best K , e.g. gap statistic (Tibshirani et al., 2001; ESL p. 519)
- An alternative is hierarchical [agglomerative] clustering (HAC), which gives a tree-based representation of the dataset

Mixture models for clustering

Mixture models for clustering

- Assume data generated from a mixture of K distributions

Mixture models for clustering

- Assume data generated from a mixture of K distributions

Mixture models for clustering

- Assume data generated from a mixture of K distributions
- Each cluster corresponds to one component of the mixture

Mixture models for clustering

- Assume data generated from a mixture of K distributions
- Each cluster corresponds to one component of the mixture

Mixture models for clustering

- Assume data generated from a mixture of K distributions
- Each cluster corresponds to one component of the mixture
- E.g. Gaussian Mixture Model (GMM):

$$p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (4)$$

Mixture models for clustering

- Assume data generated from a mixture of K distributions
- Each cluster corresponds to one component of the mixture
- E.g. Gaussian Mixture Model (GMM):

$$p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (4)$$

Mixture models for clustering

- Assume data generated from a mixture of K distributions
- Each cluster corresponds to one component of the mixture
- E.g. Gaussian Mixture Model (GMM):

$$p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (4)$$

- π_k : mixing coefficient (prior probability of cluster k), $\boldsymbol{\mu}_k$: mean, $\boldsymbol{\Sigma}_k$: covariance matrix

Mixture models for clustering

- Assume data generated from a mixture of K distributions
- Each cluster corresponds to one component of the mixture
- E.g. Gaussian Mixture Model (GMM):

$$p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (4)$$

- π_k : mixing coefficient (prior probability of cluster k), $\boldsymbol{\mu}_k$: mean, $\boldsymbol{\Sigma}_k$: covariance matrix

Mixture models for clustering

- Assume data generated from a mixture of K distributions
- Each cluster corresponds to one component of the mixture
- E.g. Gaussian Mixture Model (GMM):

$$p(\mathbf{x}|\theta) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (4)$$

- π_k : mixing coefficient (prior probability of cluster k), $\boldsymbol{\mu}_k$: mean, $\boldsymbol{\Sigma}_k$: covariance matrix
- Parameters can be estimated using Expectation-Maximization (EM) algorithm

Mixture models for clustering

- Assume data generated from a mixture of K distributions
- Each cluster corresponds to one component of the mixture
- E.g. Gaussian Mixture Model (GMM):

$$p(\mathbf{x}|\theta) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (4)$$

- π_k : mixing coefficient (prior probability of cluster k), $\boldsymbol{\mu}_k$: mean, $\boldsymbol{\Sigma}_k$: covariance matrix
- Parameters can be estimated using Expectation-Maximization (EM) algorithm

Mixture models for clustering

- Assume data generated from a mixture of K distributions
- Each cluster corresponds to one component of the mixture
- E.g. Gaussian Mixture Model (GMM):

$$p(\mathbf{x}|\theta) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (4)$$

- π_k : mixing coefficient (prior probability of cluster k), $\boldsymbol{\mu}_k$: mean, $\boldsymbol{\Sigma}_k$: covariance matrix
- Parameters can be estimated using Expectation-Maximization (EM) algorithm
- Soft clustering: each observation has a probability of belonging to each cluster

Mixture models for clustering

- Assume data generated from a mixture of K distributions
- Each cluster corresponds to one component of the mixture
- E.g. Gaussian Mixture Model (GMM):

$$p(\mathbf{x}|\theta) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (4)$$

- π_k : mixing coefficient (prior probability of cluster k), $\boldsymbol{\mu}_k$: mean, $\boldsymbol{\Sigma}_k$: covariance matrix
- Parameters can be estimated using Expectation-Maximization (EM) algorithm
- Soft clustering: each observation has a probability of belonging to each cluster

Mixture models for clustering

- Assume data generated from a mixture of K distributions
- Each cluster corresponds to one component of the mixture
- E.g. Gaussian Mixture Model (GMM):

$$p(\mathbf{x}|\theta) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (4)$$

- π_k : mixing coefficient (prior probability of cluster k), $\boldsymbol{\mu}_k$: mean, $\boldsymbol{\Sigma}_k$: covariance matrix
- Parameters can be estimated using Expectation-Maximization (EM) algorithm
- Soft clustering: each observation has a probability of belonging to each cluster
- Hard clustering: assign each observation to the cluster with the highest probability

Gaussian mixture modeling (GMM)

Gaussian mixture modeling (GMM)

A mixture of K Gaussian distributions is given by:

Gaussian mixture modeling (GMM)

A mixture of K Gaussian distributions is given by:

$$p(\mathbf{x}|\theta) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (5)$$

Gaussian mixture modeling (GMM)

A mixture of K Gaussian distributions is given by:

$$p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (5)$$

where $\boldsymbol{\theta} = (\boldsymbol{\pi}, \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\})$.

Gaussian mixture modeling (GMM)

A mixture of K Gaussian distributions is given by:

$$p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (5)$$

where $\boldsymbol{\theta} = (\boldsymbol{\pi}, \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\})$.

These parameters are estimated typically via the EM algorithm.

Gaussian mixture modeling (GMM)

A mixture of K Gaussian distributions is given by:

$$p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (5)$$

where $\boldsymbol{\theta} = (\boldsymbol{\pi}, \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\})$.

These parameters are estimated typically via the EM algorithm.

EM algorithm for GMM clustering

EM algorithm for GMM clustering

- **E-step:** Compute responsibilities

$$r_{ik} = \frac{\pi_k \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \quad (6)$$

EM algorithm for GMM clustering

- **E-step:** Compute responsibilities

$$r_{ik} = \frac{\pi_k \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \quad (6)$$

EM algorithm for GMM clustering

- **E-step:** Compute responsibilities

$$r_{ik} = \frac{\pi_k \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \quad (6)$$

- **M-step:** Update parameters

$$\pi_k = \frac{N_k}{N} \quad (7)$$

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{i=1}^N r_{ik} \mathbf{x}_i \quad (8)$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{i=1}^N r_{ik} (\mathbf{x}_i - \boldsymbol{\mu}_k) (\mathbf{x}_i - \boldsymbol{\mu}_k)^\top \quad (9)$$

EM algorithm for GMM clustering

- **E-step:** Compute responsibilities

$$r_{ik} = \frac{\pi_k \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \quad (6)$$

- **M-step:** Update parameters

$$\pi_k = \frac{N_k}{N} \quad (7)$$

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{i=1}^N r_{ik} \mathbf{x}_i \quad (8)$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{i=1}^N r_{ik} (\mathbf{x}_i - \boldsymbol{\mu}_k) (\mathbf{x}_i - \boldsymbol{\mu}_k)^\top \quad (9)$$

EM algorithm for GMM clustering

- **E-step:** Compute responsibilities

$$r_{ik} = \frac{\pi_k \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \quad (6)$$

- **M-step:** Update parameters

$$\pi_k = \frac{N_k}{N} \quad (7)$$

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{i=1}^N r_{ik} \mathbf{x}_i \quad (8)$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{i=1}^N r_{ik} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^\top \quad (9)$$

where $N_k = \sum_{i=1}^N r_{ik}$ is the effective number of points assigned to cluster k .

KMeans as special case of GMM

KMeans as special case of GMM

KMeans can be seen as a special case of GMM with:

KMeans as special case of GMM

KMeans can be seen as a special case of GMM with:

- All components are spherical Gaussians with identical covariance $\Sigma_k = \sigma^2 I$.

KMeans as special case of GMM

KMeans can be seen as a special case of GMM with:

- All components are spherical Gaussians with identical covariance $\Sigma_k = \sigma^2 I$.
- Each cluster has equal prior probability $\pi_k = \frac{1}{K}$.

KMeans as special case of GMM

KMeans can be seen as a special case of GMM with:

- All components are spherical Gaussians with identical covariance $\Sigma_k = \sigma^2 I$.
- Each cluster has equal prior probability $\pi_k = \frac{1}{K}$.
- Ultimately,

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{i=1}^N r_{ik} \mathbf{x}_i \quad (10)$$

where $r_{ik} \in \{0, 1\}$ indicates hard assignment of point i to cluster k .

Hierarchical agglomerative clustering methods

Hierarchical agglomerative clustering methods

Iteratively group clusters U and V by pairing the **closest** based on cluster separation metric $D(U, V)$

Hierarchical agglomerative clustering methods

Iteratively group clusters U and V by pairing the **closest** based on cluster separation metric $D(U, V)$

- Unweighted pair-group method with arithmetic means

$$D(U, V) = \sum_{ij} \frac{d(u_i, v_j)}{|U| \cdot |V|} \quad (11)$$

Hierarchical agglomerative clustering methods

Iteratively group clusters U and V by pairing the **closest** based on cluster separation metric $D(U, V)$

- Unweighted pair-group method with arithmetic means

$$D(U, V) = \sum_{ij} \frac{d(u_i, v_j)}{|U| \cdot |V|} \quad (11)$$

- Weighted pair-group method with arithmetic means

$$D(U, V) = \frac{d(S, V) + d(T, V)}{2}, \quad U = S \cup T \quad (12)$$

Hierarchical agglomerative clustering methods

Iteratively group clusters U and V by pairing the **closest** based on cluster separation metric $D(U, V)$

- Unweighted pair-group method with arithmetic means

$$D(U, V) = \sum_{ij} \frac{d(u_i, v_j)}{|U| \cdot |V|} \quad (11)$$

- Weighted pair-group method with arithmetic means

$$D(U, V) = \frac{d(S, V) + d(T, V)}{2}, \quad U = S \cup T \quad (12)$$

- Single linkage method

$$D(U, V) = \min d(u_i, v_j) \quad (13)$$

Hierarchical agglomerative clustering methods

Iteratively group clusters U and V by pairing the **closest** based on cluster separation metric $D(U, V)$

- Unweighted pair-group method with arithmetic means

$$D(U, V) = \sum_{ij} \frac{d(u_i, v_j)}{|U| \cdot |V|} \quad (11)$$

- Weighted pair-group method with arithmetic means

$$D(U, V) = \frac{d(S, V) + d(T, V)}{2}, \quad U = S \cup T \quad (12)$$

- Single linkage method

$$D(U, V) = \min d(u_i, v_j) \quad (13)$$

- Complete linkage method

$$D(U, V) = \max d(u_i, v_j) \quad (14)$$

Hierarchical agglomerative clustering methods

Iteratively group clusters U and V by pairing the **closest** based on cluster separation metric $D(U, V)$

- Unweighted pair-group method with arithmetic means

$$D(U, V) = \sum_{ij} \frac{d(u_i, v_j)}{|U| \cdot |V|} \quad (11)$$

- Weighted pair-group method with arithmetic means

$$D(U, V) = \frac{d(S, V) + d(T, V)}{2}, \quad U = S \cup T \quad (12)$$

- Single linkage method

$$D(U, V) = \min d(u_i, v_j) \quad (13)$$

- Complete linkage method

$$D(U, V) = \max d(u_i, v_j) \quad (14)$$

- Ward's method

$$D(U, V) = \frac{|U| \cdot |V|}{|U| + |V|} \|\bar{u}_i - \bar{v}_j\|^2 \quad (15)$$

Hierarchical Clustering: Linkage

Hierarchical Clustering: Linkage

- **Single linkage:**

minimum distance or nearest neighbor (2 closest border points)

Hierarchical Clustering: Linkage

- **Single linkage:**

minimum distance or nearest neighbor (2 closest border points)

Hierarchical Clustering: Linkage

- **Single linkage:**

minimum distance or nearest neighbor (2 closest border points)



Hierarchical Clustering: Linkage

- **Single linkage:**

minimum distance or nearest neighbor (2 closest border points)



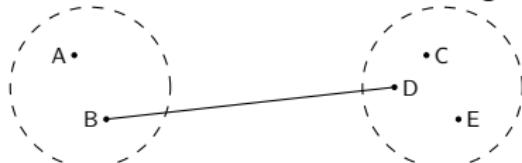
- **Complete linkage:**

maximum distance or farthest neighbor (2 farthest border points)

Hierarchical Clustering: Linkage

- **Single linkage:**

minimum distance or nearest neighbor (2 closest border points)



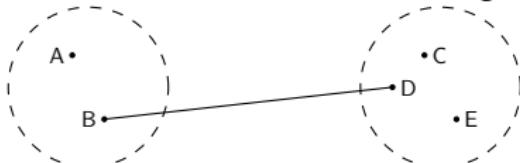
- **Complete linkage:**

maximum distance or farthest neighbor (2 farthest border points)

Hierarchical Clustering: Linkage

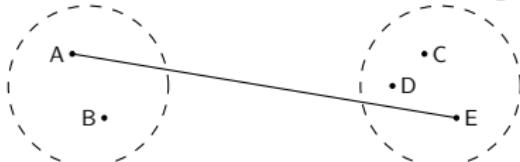
- **Single linkage:**

minimum distance or nearest neighbor (2 closest border points)



- **Complete linkage:**

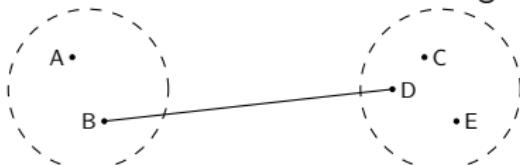
maximum distance or farthest neighbor (2 farthest border points)



Hierarchical Clustering: Linkage

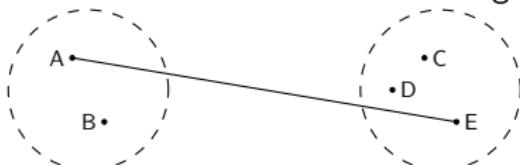
- **Single linkage:**

minimum distance or nearest neighbor (2 closest border points)



- **Complete linkage:**

maximum distance or farthest neighbor (2 farthest border points)



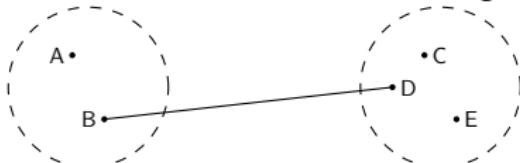
- **Average linkage (unweighted pair-group method):**

average distance (all to all)

Hierarchical Clustering: Linkage

- **Single linkage:**

minimum distance or nearest neighbor (2 closest border points)



- **Complete linkage:**

maximum distance or farthest neighbor (2 farthest border points)



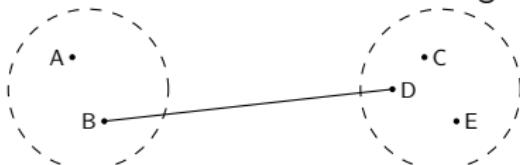
- **Average linkage (unweighted pair-group method):**

average distance (all to all)

Hierarchical Clustering: Linkage

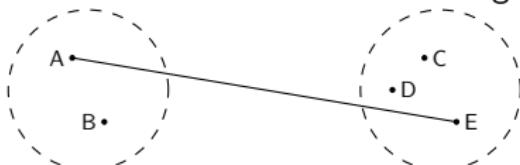
- **Single linkage:**

minimum distance or nearest neighbor (2 closest border points)



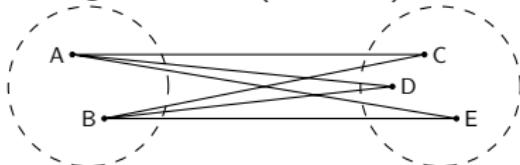
- **Complete linkage:**

maximum distance or farthest neighbor (2 farthest border points)



- **Average linkage (unweighted pair-group method):**

average distance (all to all)



Agglomerative clustering algorithm

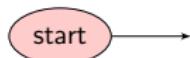
Agglomerative clustering algorithm

Example:

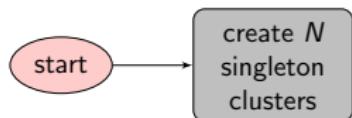


Agglomerative clustering algorithm

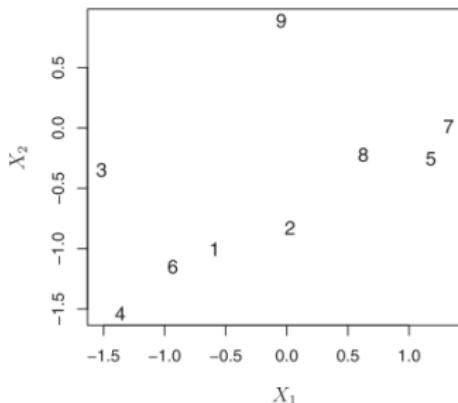
Example:



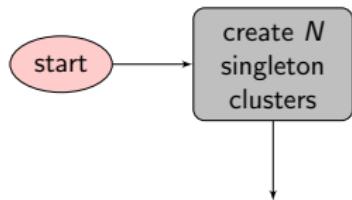
Agglomerative clustering algorithm



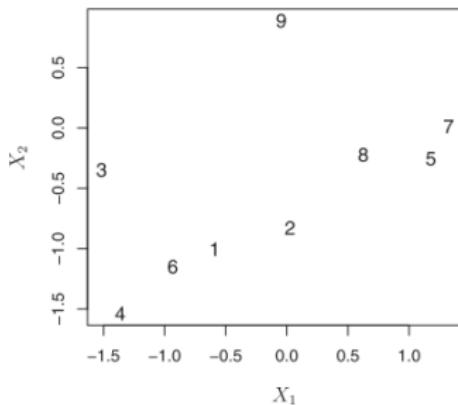
Example:



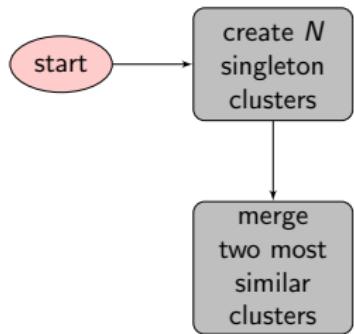
Agglomerative clustering algorithm



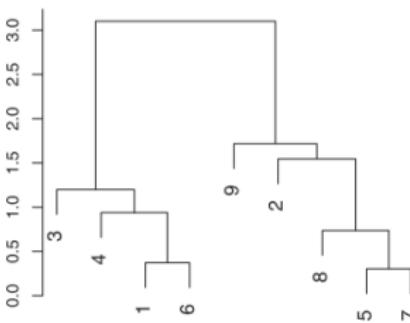
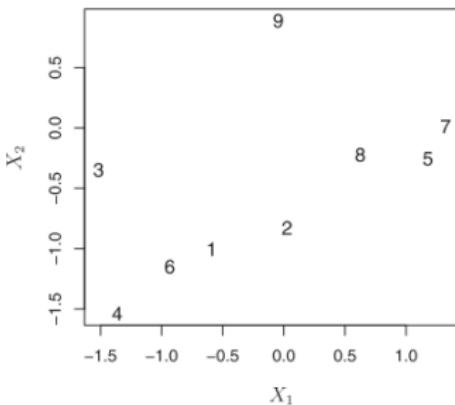
Example:



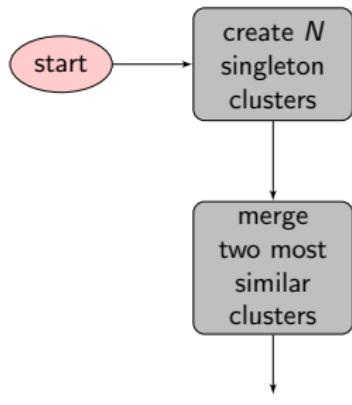
Agglomerative clustering algorithm



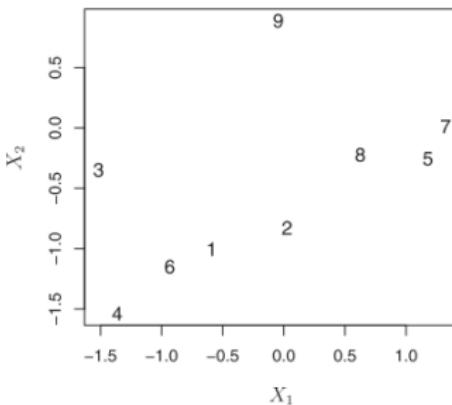
Example:



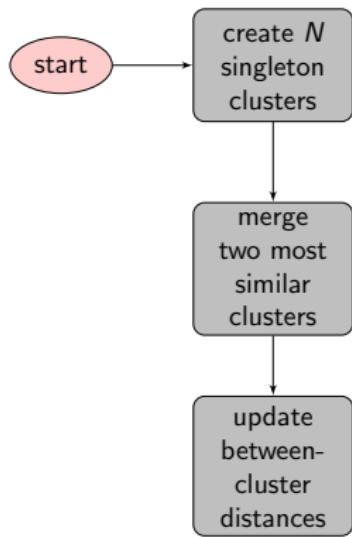
Agglomerative clustering algorithm



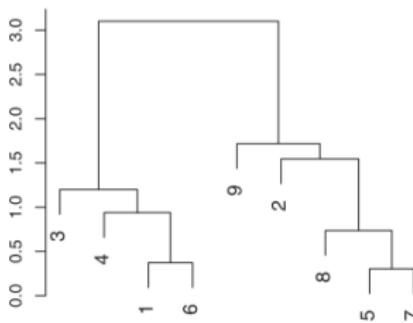
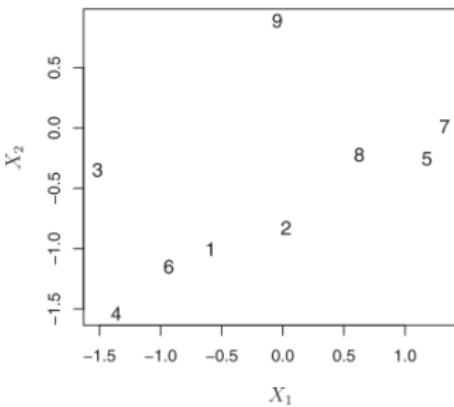
Example:



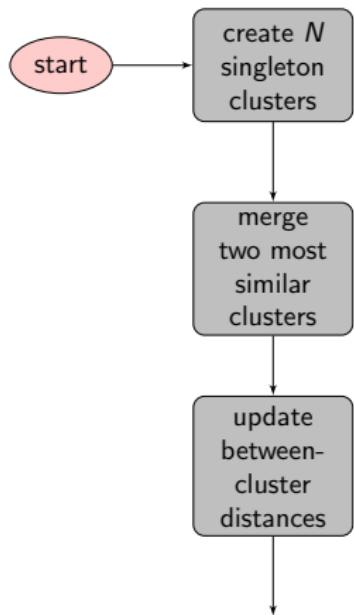
Agglomerative clustering algorithm



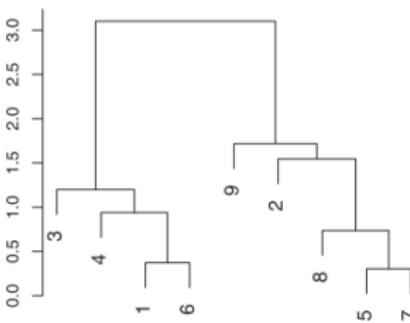
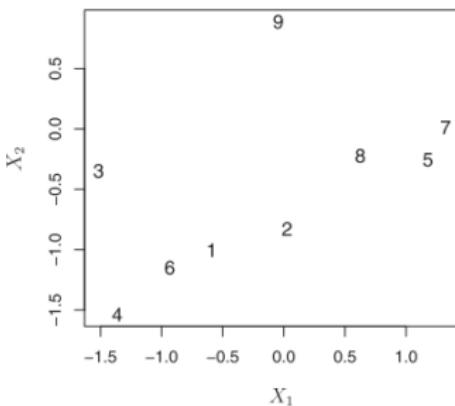
Example:



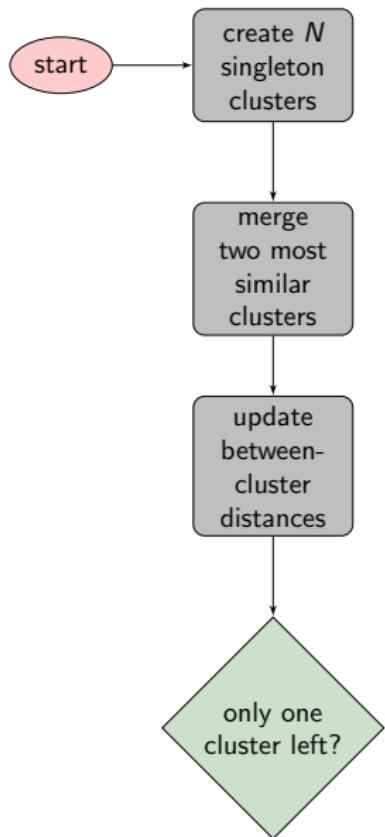
Agglomerative clustering algorithm



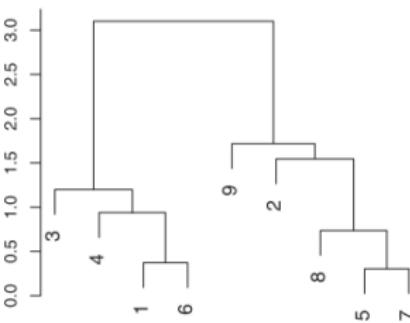
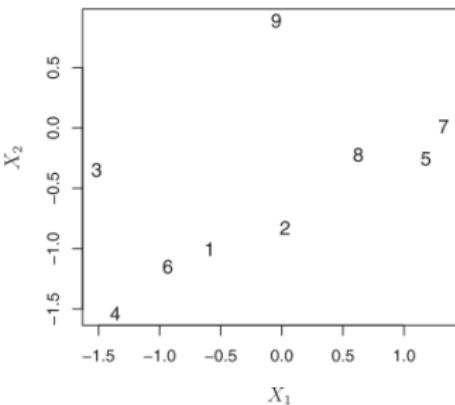
Example:



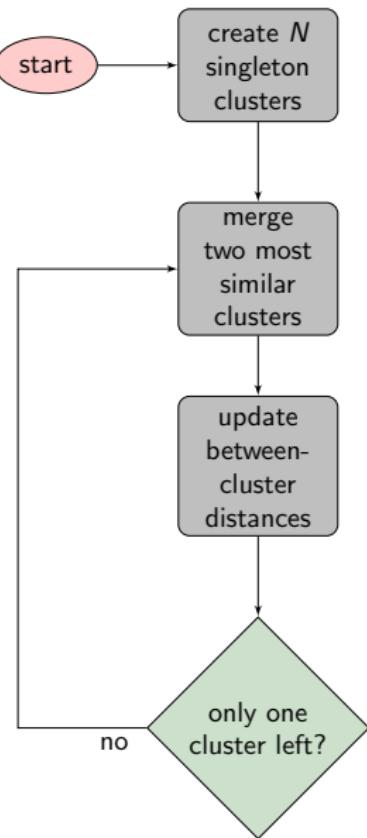
Agglomerative clustering algorithm



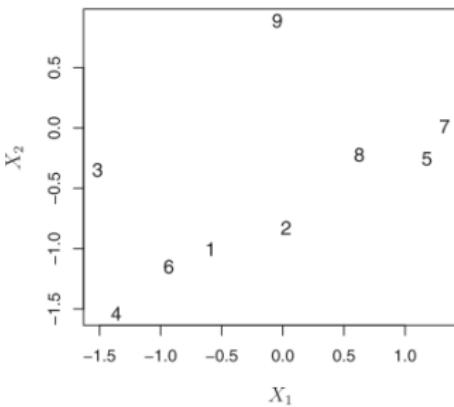
Example:



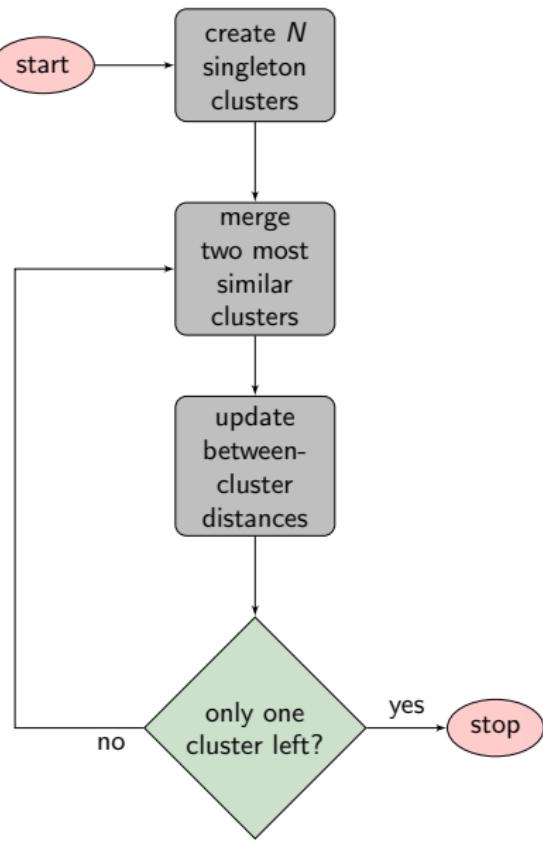
Agglomerative clustering algorithm



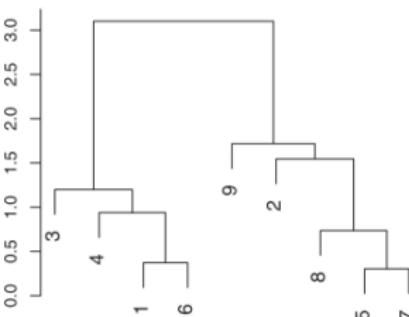
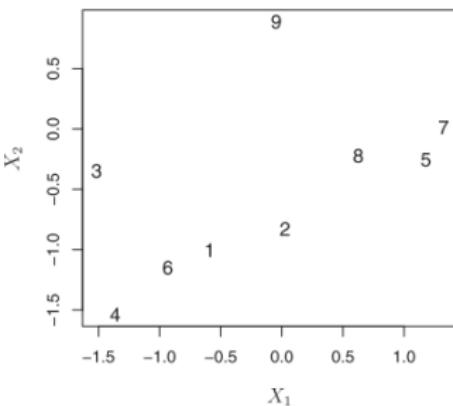
Example:



Agglomerative clustering algorithm

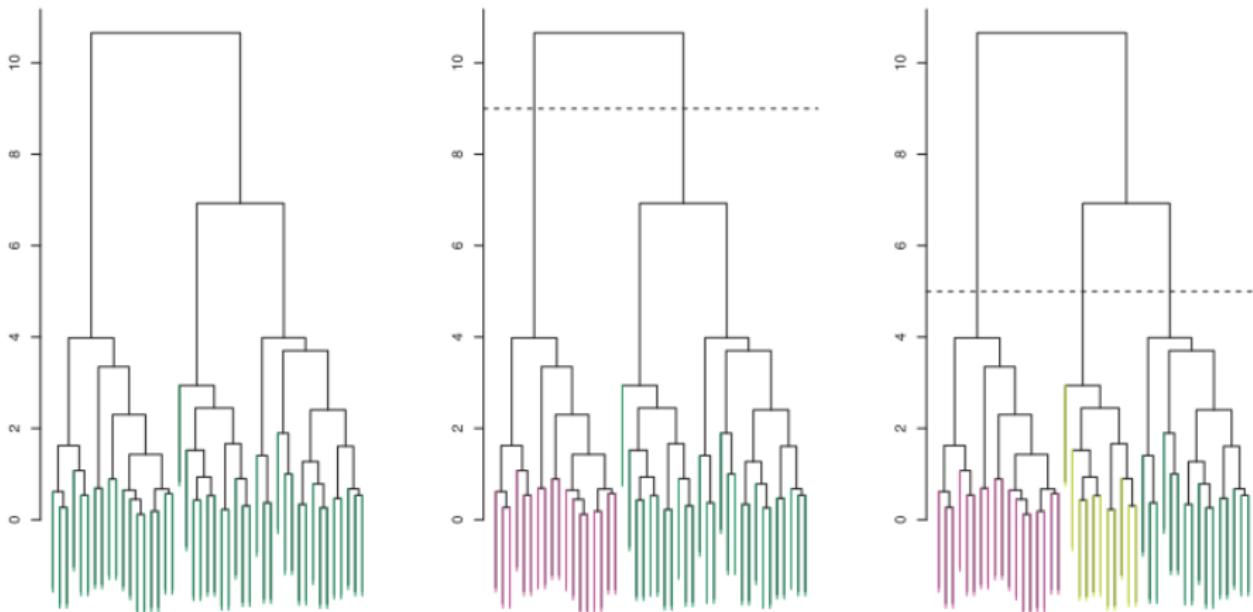


Example:



Choosing the clusters

We decide a **Cut**



Source: James et Al., Introduction to Statistical Learning

Practical Considerations

Advantages

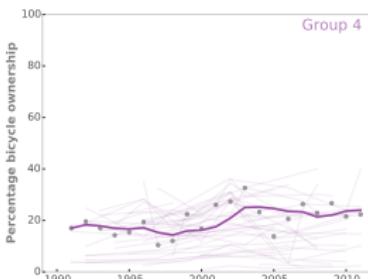
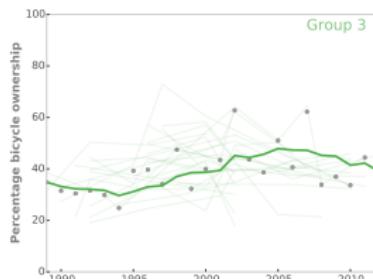
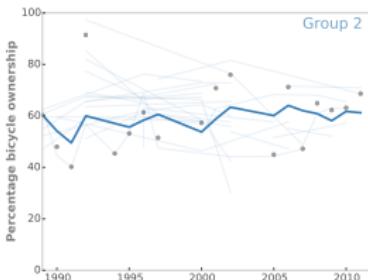
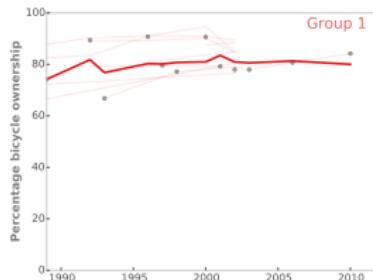
- No apriori number of clusters required
- Simple algorithms
- Self-organized structural view of data

Disadvantages

- Dendrogram often difficult to visualize
- Sometimes the inherent clusters in our data are not hierarchical by nature
(K-means performs better in these cases)

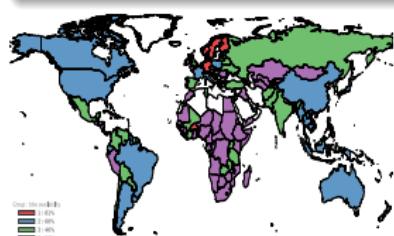
HAC example: Bicycle ownership trends

Pattern discovery from survey data in 150 countries spanning 30 years ¹



Key findings

- To cluster time-series data of varying lengths, the dynamic time warping (DTW) algorithm can be used to compute the dissimilarity matrix



¹Oke et al., 2015

<https://www.sciencedirect.com/science/article/abs/pii/S2214140515006787>

Density-based clustering

Density-based clustering approaches are based on these hypotheses:

- Clusters are dense spatial regions
- Clusters are separated by low-density regions
- The density of points in a cluster are greater than a given minimum

Density-based clustering

Density-based clustering approaches are based on these hypotheses:

- Clusters are dense spatial regions
- Clusters are separated by low-density regions
- The density of points in a cluster are greater than a given minimum

Examples of density-based clustering algorithms:

Density-based clustering

Density-based clustering approaches are based on these hypotheses:

- Clusters are dense spatial regions
- Clusters are separated by low-density regions
- The density of points in a cluster are greater than a given minimum

Examples of density-based clustering algorithms:

- DBSCAN
- OPTICS

Density-based spatial clustering of applications with noise

²<https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.121.9220>

Density-based spatial clustering of applications with noise

²<https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.121.9220>

Density-based spatial clustering of applications with noise

²<https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.121.9220>

Density-based spatial clustering of applications with noise

- Introduced in 1996 by Ester, Kriegel, Sander & Xu²
- Finds dense regions; recursively expands them to converge at clusters

²<https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.121.9220>

Density-based spatial clustering of applications with noise

- Introduced in 1996 by Ester, Kriegel, Sander & Xu²
- Finds dense regions; recursively expands them to converge at clusters
- Parameters:

²<https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.121.9220>

Density-based spatial clustering of applications with noise

- Introduced in 1996 by Ester, Kriegel, Sander & Xu²
- Finds dense regions; recursively expands them to converge at clusters
- Parameters:
 - ε : radius of neighborhood

²<https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.121.9220>

Density-based spatial clustering of applications with noise

- Introduced in 1996 by Ester, Kriegel, Sander & Xu²
- Finds dense regions; recursively expands them to converge at clusters
- Parameters:
 - ε : radius of neighborhood
 - minPoints: minimum number of observations within a neighborhood

²<https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.121.9220>

Density-based spatial clustering of applications with noise

- Introduced in 1996 by Ester, Kriegel, Sander & Xu²
- Finds dense regions; recursively expands them to converge at clusters
- Parameters:
 - ε : radius of neighborhood
 - minPoints: minimum number of observations within a neighborhood

²<https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.121.9220>

Density-based spatial clustering of applications with noise

- Introduced in 1996 by Ester, Kriegel, Sander & Xu²
- Finds dense regions; recursively expands them to converge at clusters
- Parameters:
 - ε : radius of neighborhood
 - minPoints: minimum number of observations within a neighborhood

Source: <https://www.nature.com/articles/srep34406>

Figure: Example of clusters generated by DBSCAN on a dataset

²<https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.121.9220>

Density-based spatial clustering of applications with noise

- Introduced in 1996 by Ester, Kriegel, Sander & Xu²
- Finds dense regions; recursively expands them to converge at clusters
- Parameters:
 - ε : radius of neighborhood
 - minPoints: minimum number of observations within a neighborhood



Source: <https://www.nature.com/articles/srep34406>

Figure: Example of clusters generated by DBSCAN on a dataset

²<https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.121.9220>

Introduction
oo

Similarity measures
oooo

K-means clustering
oooooooo

Mixture models
oooo

Hierarchical clustering
oooooo

DBSCAN
oo●ooo

Silhouette
ooo

Summary
o

DBSCAN: key definitions

DBSCAN: key definitions

- Epsilon neighborhood, N_ε : set of all observations within distance ε
- Core point: has at least `minPoint` observations within its N_ε
- DDR: An observation j is **directly density reachable** from a core point i if $j \in N_\varepsilon$
- DR: Two observations are **density reachable** if there exists a chain of DDR observations linking them
- Boundary/border points: these are DDR but not core points
- Noise/outlier points: do not belong to any observations N_ε

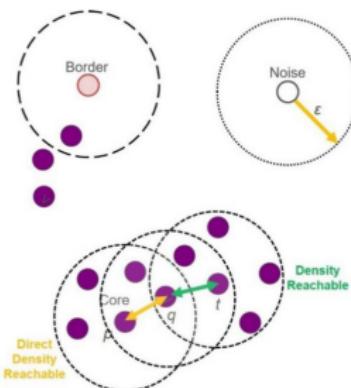
Source: Giacoumidis et al. (2019)

<https://www.mdpi.com/2076-3417/9/20/4398/htm>

Figure: DBSCAN example with `minpoints = 4`

DBSCAN: key definitions

- Epsilon neighborhood, N_ε : set of all observations within distance ε
- Core point: has at least `minPoint` observations within its N_ε
- DDR: An observation j is **directly density reachable** from a core point i if $j \in N_\varepsilon$
- DR: Two observations are **density reachable** if there exists a chain of DDR observations linking them
- Boundary/border points: these are DDR but not core points
- Noise/outlier points: do not belong to any observations N_ε



Source: Giacoumidis et al. (2019)

<https://www.mdpi.com/2076-3417/9/20/4398/htm>

Figure: DBSCAN example with `minpoints = 4`

Introduction
oo

Similarity measures
oooo

K-means clustering
oooooooo

Mixture models
oooo

Hierarchical clustering
oooooo

DBSCAN
oooo●oo

Silhouette
ooo

Summary
o

DBSCAN serial algorithm

DBSCAN serial algorithm

```
1: procedure DBSCAN( $X, \varepsilon, \text{minPoints}$ )
2:   for each unvisited point  $x \in X$  do
3:     mark  $x$  as visited
4:      $N \leftarrow \text{FINDNEIGHBORS}(x, \varepsilon)$ 
5:     if  $|N| < \text{minPoints}$  then
6:       mark  $x$  as noise
7:     else
8:        $C \leftarrow \{x\}$ 
9:     end if
10:    for each point  $x' \in N$  do
11:       $N \leftarrow N \setminus x'$ 
12:      if  $x'$  is not visited then
13:        mark  $x'$  as visited
14:         $N' \leftarrow \text{FINDNEIGHBORS}(x', \varepsilon)$ 
15:        if  $|N'| \geq \text{minPoints}$  then
16:           $N \leftarrow N \cup N'$ 
17:        end if
18:      end if
19:      if  $x'$  is not yet a member of any cluster then
20:         $C \leftarrow C \cup \{x'\}$ 
21:      end if
22:    end for
23:  end for
24: end procedure
```

Introduction
oo

Similarity measures
oooo

K-means clustering
oooooooo

Mixture models
oooo

Hierarchical clustering
oooooo

DBSCAN
oooo●○

Silhouette
ooo

Summary
o

DBSCAN example: stop detection

DBSCAN example: stop detection

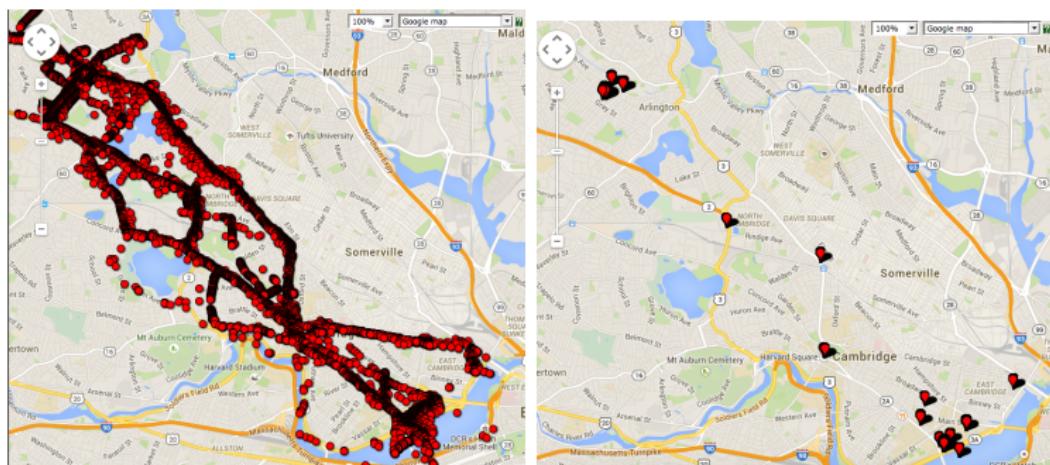
- Stop detection using smartphone data.

DBSCAN example: stop detection

- Stop detection using smartphone data.
- Challenges: GPS data is noisy. Data gaps (e.g. no GPS inside buildings).

DBSCAN example: stop detection

- Stop detection using smartphone data.
- Challenges: GPS data is noisy. Data gaps (e.g. no GPS inside buildings).



Introduction
oo

Similarity measures
oooo

K-means clustering
oooooooo

Mixture models
oooo

Hierarchical clustering
oooooo

DBSCAN
ooooo●

Silhouette
ooo

Summary
o

DBSCAN considerations

DBSCAN considerations

- Performs well on geographical data
- Requires careful selection of two parameters (can be computationally intensive)
- Several improvements and updates to the original DBSCAN algorithm have been made (e.g. OPTICS: “Ordering points to identify the clustering structure”)

Fitness of clustering solution

Good clustering should:

- Minimize **within-cluster** (inter-cluster) variability (**W**)
- Maximize the **silhouette** (Rousseeuw, 1987)
- Several other goodness-of-fit measures can be used:
 - Krzanowski-Lai (KL) index
 - Gap statistic (Tibshirani et al., 2001)
- We consider the silhouette metric in detail

Silhouette

- Silhouette of observation \mathbf{x}_j , $s(\mathbf{x}_j)$:

$$s(\mathbf{x}_j) = \frac{b(\mathbf{x}_j) - a(\mathbf{x}_j)}{\max\{a(\mathbf{x}_j), b(\mathbf{x}_j)\}} \quad (16)$$

Silhouette

- Silhouette of observation \mathbf{x}_j , $s(\mathbf{x}_j)$:

$$s(\mathbf{x}_j) = \frac{b(\mathbf{x}_j) - a(\mathbf{x}_j)}{\max\{a(\mathbf{x}_j), b(\mathbf{x}_j)\}} \quad (16)$$

- $a(\mathbf{x}_j)$ = average distance between \mathbf{x}_j and *all* other elements of its cluster (intra-cluster distance)

Silhouette

- Silhouette of observation \mathbf{x}_j , $s(\mathbf{x}_j)$:

$$s(\mathbf{x}_j) = \frac{b(\mathbf{x}_j) - a(\mathbf{x}_j)}{\max\{a(\mathbf{x}_j), b(\mathbf{x}_j)\}} \quad (16)$$

- $a(\mathbf{x}_j)$ = average distance between \mathbf{x}_j and *all* other elements of its cluster (intra-cluster distance)
- $b(\mathbf{x}_j)$ = average distance between \mathbf{x}_j and *all* elements of the second nearest cluster.

Silhouette

- Silhouette of observation \mathbf{x}_j , $s(\mathbf{x}_j)$:

$$s(\mathbf{x}_j) = \frac{b(\mathbf{x}_j) - a(\mathbf{x}_j)}{\max\{a(\mathbf{x}_j), b(\mathbf{x}_j)\}} \quad (16)$$

- $a(\mathbf{x}_j)$ = average distance between \mathbf{x}_j and *all* other elements of its cluster (intra-cluster distance)
- $b(\mathbf{x}_j)$ = average distance between \mathbf{x}_j and *all* elements of the second nearest cluster.
- Measures how well an observation fits a cluster

$$-1 < s(\mathbf{x}_j) < 1 \quad (17)$$

Silhouette

- Silhouette of observation \mathbf{x}_j , $s(\mathbf{x}_j)$:

$$s(\mathbf{x}_j) = \frac{b(\mathbf{x}_j) - a(\mathbf{x}_j)}{\max\{a(\mathbf{x}_j), b(\mathbf{x}_j)\}} \quad (16)$$

- $a(\mathbf{x}_j)$ = average distance between \mathbf{x}_j and *all* other elements of its cluster (intra-cluster distance)
- $b(\mathbf{x}_j)$ = average distance between \mathbf{x}_j and *all* elements of the second nearest cluster.
- Measures how well an observation fits a cluster

$$-1 < s(\mathbf{x}_j) < 1 \quad (17)$$

- We want $a(\mathbf{x}_j)$ to be small and $b(\mathbf{x}_j)$ to be large:

$$a(\mathbf{x}_j) \ll b(\mathbf{x}_j)$$

Silhouette

- Silhouette of observation \mathbf{x}_j , $s(\mathbf{x}_j)$:

$$s(\mathbf{x}_j) = \frac{b(\mathbf{x}_j) - a(\mathbf{x}_j)}{\max\{a(\mathbf{x}_j), b(\mathbf{x}_j)\}} \quad (16)$$

- $a(\mathbf{x}_j)$ = average distance between \mathbf{x}_j and *all* other elements of its cluster (intra-cluster distance)
- $b(\mathbf{x}_j)$ = average distance between \mathbf{x}_j and *all* elements of the second nearest cluster.
- Measures how well an observation fits a cluster

$$-1 < s(\mathbf{x}_j) < 1 \quad (17)$$

- We want $a(\mathbf{x}_j)$ to be small and $b(\mathbf{x}_j)$ to be large:

$$a(\mathbf{x}_j) \ll b(\mathbf{x}_j)$$

Silhouette

- Silhouette of observation \mathbf{x}_j , $s(\mathbf{x}_j)$:

$$s(\mathbf{x}_j) = \frac{b(\mathbf{x}_j) - a(\mathbf{x}_j)}{\max\{a(\mathbf{x}_j), b(\mathbf{x}_j)\}} \quad (16)$$

- $a(\mathbf{x}_j)$ = average distance between \mathbf{x}_j and *all* other elements of its cluster (intra-cluster distance)
- $b(\mathbf{x}_j)$ = average distance between \mathbf{x}_j and *all* elements of the second nearest cluster.
- Measures how well an observation fits a cluster

$$-1 < s(\mathbf{x}_j) < 1 \quad (17)$$

- We want $a(\mathbf{x}_j)$ to be small and $b(\mathbf{x}_j)$ to be large:

$$a(\mathbf{x}_j) \ll b(\mathbf{x}_j) \implies$$

Silhouette

- Silhouette of observation \mathbf{x}_j , $s(\mathbf{x}_j)$:

$$s(\mathbf{x}_j) = \frac{b(\mathbf{x}_j) - a(\mathbf{x}_j)}{\max\{a(\mathbf{x}_j), b(\mathbf{x}_j)\}} \quad (16)$$

- $a(\mathbf{x}_j)$ = average distance between \mathbf{x}_j and *all* other elements of its cluster (intra-cluster distance)
- $b(\mathbf{x}_j)$ = average distance between \mathbf{x}_j and *all* elements of the second nearest cluster.
- Measures how well an observation fits a cluster

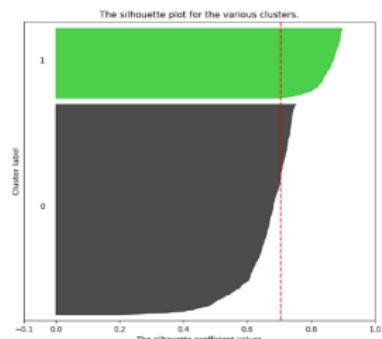
$$-1 < s(\mathbf{x}_j) < 1 \quad (17)$$

- We want $a(\mathbf{x}_j)$ to be small and $b(\mathbf{x}_j)$ to be large:

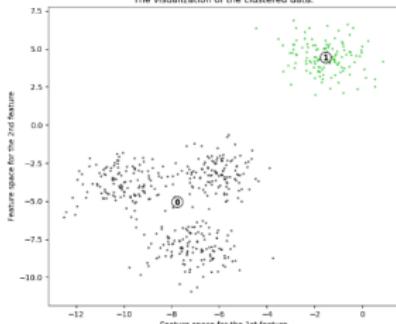
$$a(\mathbf{x}_j) \ll b(\mathbf{x}_j) \implies s(\mathbf{x}_j) \rightarrow 1 \quad (18)$$

Silhouette: visualization

Silhouette analysis for KMeans clustering on sample data with n_clusters = 2

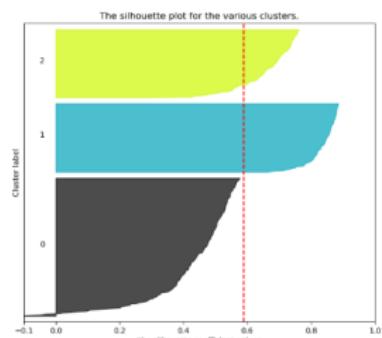


The visualization of the clustered data.

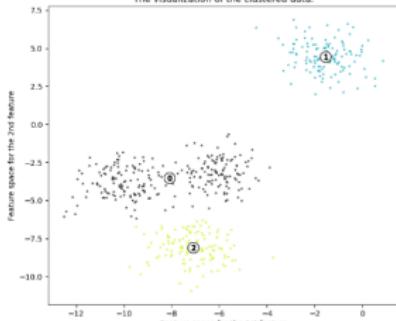


source: Scikit-learn: scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

Silhouette analysis for KMeans clustering on sample data with n_clusters = 3



The visualization of the clustered data.



Outlook

Outlook

- Assigned reading: ISLR 10.3, 10.4

Outlook

- Assigned reading: ISLR 10.3, 10.4
- Further recommended reading: ESL 14.3