CEE 616: Probabilistic Machine Learning

M5 Unsupervised Learning:
L5c: Clustering

Jimi Oke

UMass Amherst

College of Engineering

Tue, Dec 9, 2025

Outline

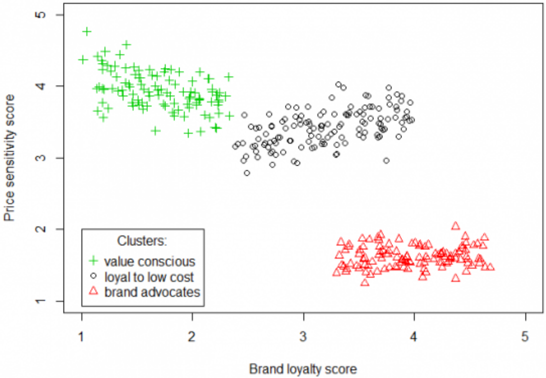
What is Clustering?

- 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
  - low similarity (high separation) between different clusters

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

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

- **Euclidean distance**

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

- Manhattan distance

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

- Minkowski distance

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

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

(12-7)

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

- Care should be taken with multiple dimensions and varying scales
  - Scaling/normalization typically leads to better results
  - E.g. Min-max scaling: $x_{new} = \dfrac{x - x_{min}}{x_{max} - x_{min}}$, $x_{new} \in [0, 1]$
- 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: Number of clusters. Design parameter to decide in advance.
- Cluster centroid: mean of observations assigned to cluster C_k:

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

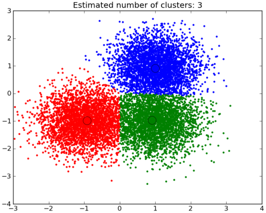
- Within cluster variation of the k-th cluster

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

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

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

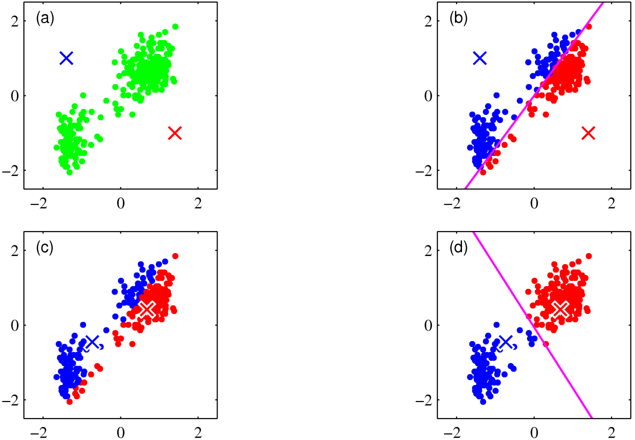
- $\Rightarrow$ Assign $\mathbf{x}$ to C_k with minimum $d(\mathbf{x}, \bar{\mathbf{x}}_k)$



Source: www.scikit-learn.org

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

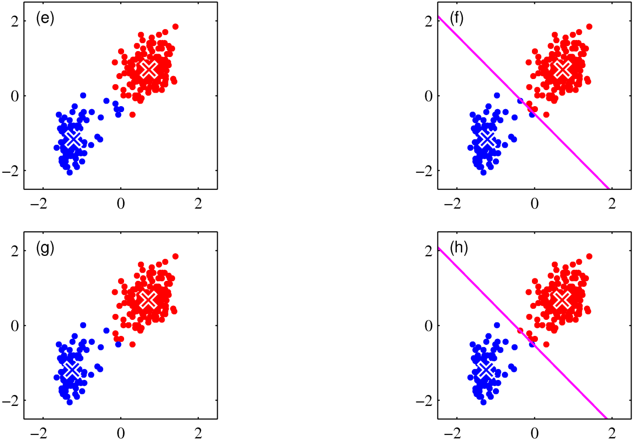
$$\min \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{\mathbf{x} \in C_k} d(\mathbf{x}, \bar{\mathbf{x}}_k)^2$$



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

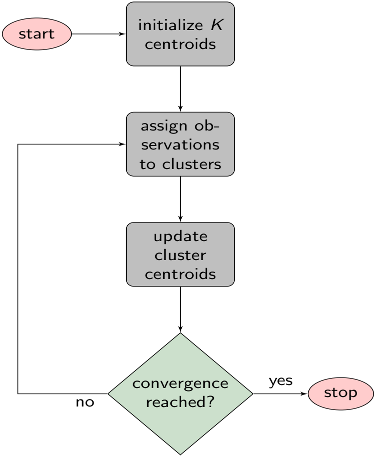
K-means Clustering: Illustration (e) - (h)

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



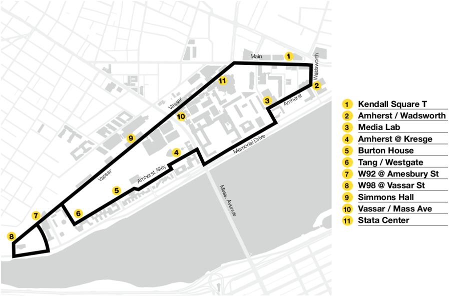
Source: Cristopher M. Bishop, *Pattern Recognition and Machine Learning*

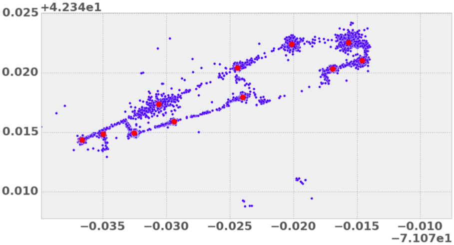
K-means Clustering: Algorithm



- 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

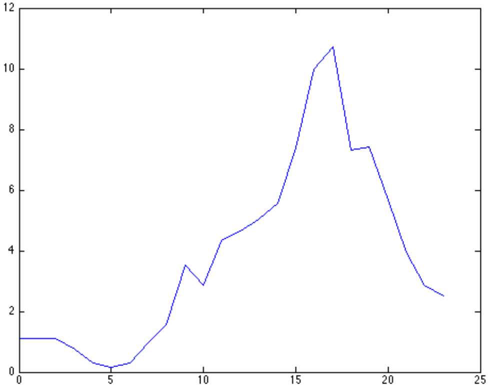




Purple dots: GPS points from buses; Red dots: centroids

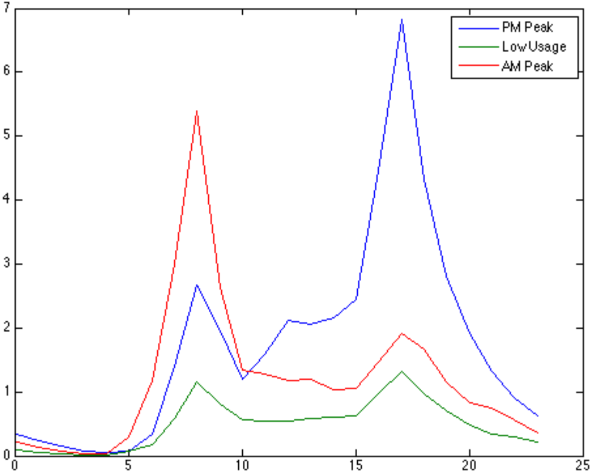
Clustering Hubway rentals

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

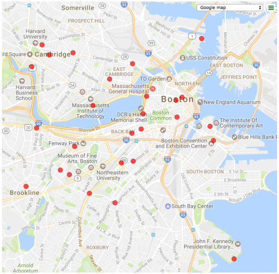


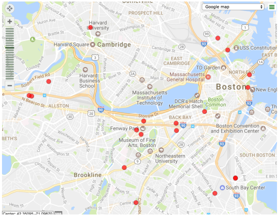
Clustering Hubway rentals

- Month of November 2013
- Consider weekdays only



Clustering Hubway rentals (cont.)





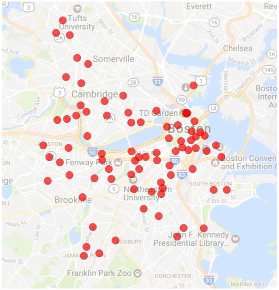
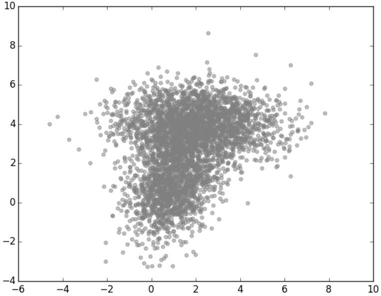


Figure: AM Peak, PM Peak, Low Usage docks

Choice of K

Which K would you choose?



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

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

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

- 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)

A mixture of K Gaussian distributions is given by:

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

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

- **E-step:** Compute responsibilities

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

- **M-step:** Update parameters

$$\pi_k = \frac{N_k}{N} \tag{7}$$

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

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

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 can be seen as a special case of GMM with:

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

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

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

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

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

- Weighted pair-group method with arithmetic means

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

- Single linkage method

$$D(U, V) = \min d(u_i, v_j) \tag{13}$$

- Complete linkage method

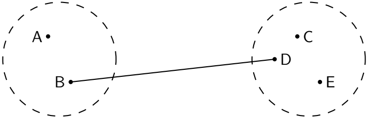
$$D(U, V) = \max d(u_i, v_j) \tag{14}$$

- Ward's method

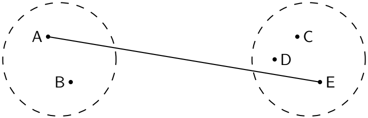
$$D(U, V) = \frac{|U| \cdot |V|}{|U| + |V|} ||\bar{u}_i - \bar{v}_j||^2 \tag{15}$$

Hierarchical Clustering: Linkage

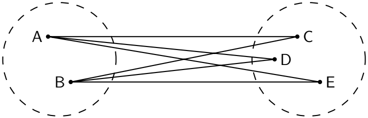
- **Single linkage**:
  minimum distance or nearest neighbor (2 closest border points)



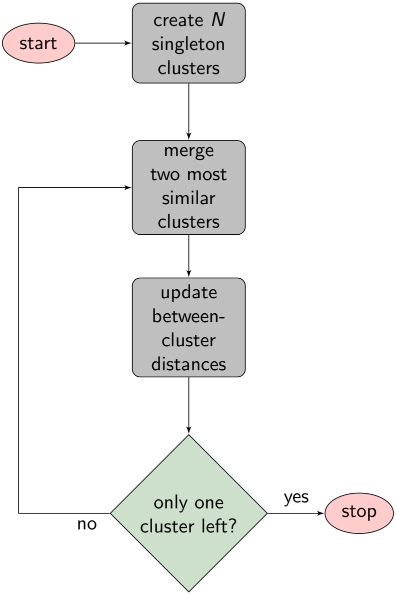
- **Complete linkage**:
  maximum distance or farthest neighbor (2 farthest border points)



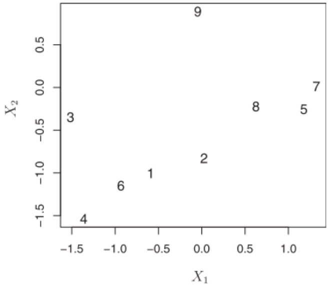
- **Average linkage (unweighted pair-group method)**:
  average distance (all to all)

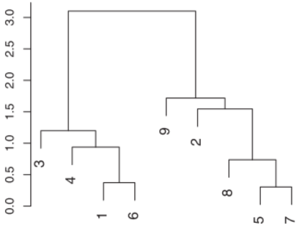


Agglomerative clustering algorithm



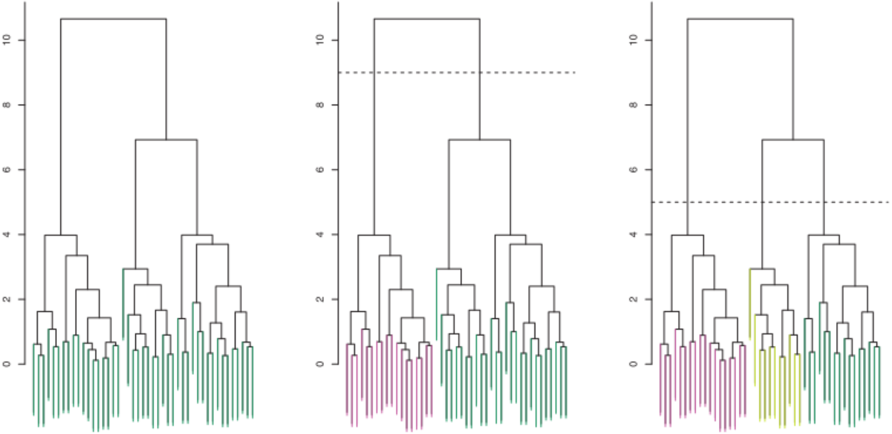
Example:





Choosing the clusters

We decide a Cut



Source: James et Al., Introduction to Statistical Learning

Practical Considerations

Advantages

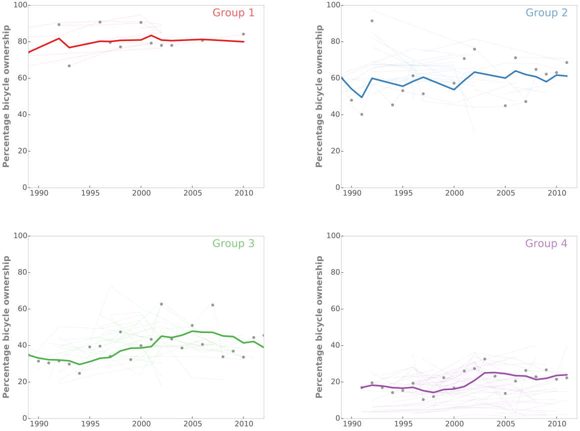
- No apriori number of clusters required
- Simple algorithms
- Self-organized structural view of data

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)

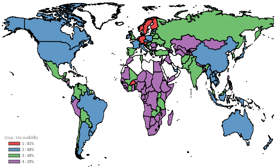
HAC example: Bicycle ownership trends

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



[1] Oke 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

Examples of density-based clustering algorithms:

- DBSCAN
- OPTICS

Density-based spatial clustering of applications with noise

- Introduced in 1996 by Ester, Kriegel, Sander & Xu[2]
- Finds dense regions; recursively expands them to converge at clusters
- Parameters:
  - ε: radius of neighborhood
  - minPoints: minimum number of observations within a neighborhood

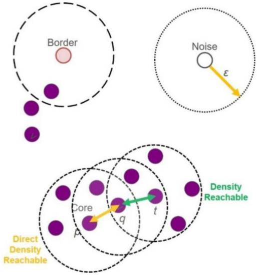
Source: https://www.nature.com/articles/srep34406

Figure: Example of clusters generated by DBSCAN on a dataset

[2] https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.121.9220

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

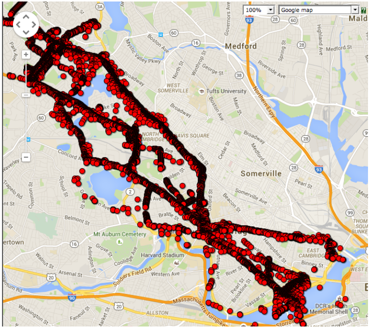
Figure: DBSCAN example with minpoints = 4

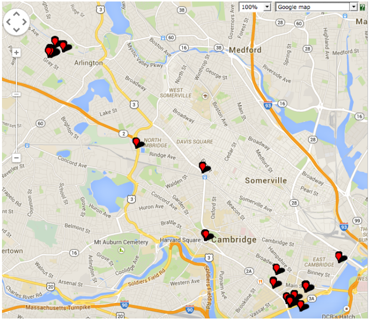
DBSCAN serial algorithm

```
procedure DBSCAN(X, ε, minPoints)
    for each unvisited point x ∈ X do
        mark x as visited
        N ← FindNeighbors(x, ε)
        if |N| < minPoints then
            mark x as noise
        else
            C ← {x}
        end if
        for each point x′ ∈ N do
            N ← N \ x′
            if x′ is not visited then
                mark x′ as visited
                N′ ← FindNeighbors(x′, ε)
                if |N′| ≥ minPoints then
                    N ← N ∪ N′
                end if
            end if
            if x′ is not yet a member of any cluster then
                C ← C ∪ {x′}
            end if
        end for
    end for
end procedure
```

DBSCAN example: stop detection

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





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 $\mathbf{x}_j$, $s(\mathbf{x}_j)$:

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

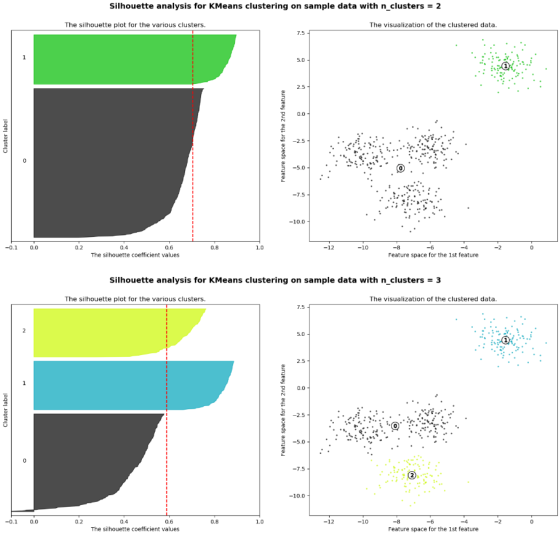
- $a(\mathbf{x}_j)=$ average distance between $\mathbf{x}_j$ and *all* other elements of its cluster (intra-cluster distance)
- $b(\mathbf{x}_j)=$ average distance between $\mathbf{x}_j$ and *all* elements of the second nearest cluster.
- Measures how well an observation fits a cluster

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

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

$$a(\mathbf{x}_j) \ll b(\mathbf{x}_j) \Longrightarrow s(\mathbf{x}_j) \to 1 \tag{18}$$

Silhouette: visualization



source:Scikit-learn: `scikit-learn.org/stable/auto/examples/cluster/plot_kmeans_silhouette_analysis.html`

Outlook

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

Temporary page!

LaTeX was unable to guess the total number of pages correctly. As the
unprocessed data that should have been added to the final page this e
has been added to receive it.
If you rerun the document (without altering it) this surplus page will g
because LaTeX now knows how many pages to expect for this documen