Lecture 5

Unsupervised learning
Clustering ||

Golnaz Taheri, PhD
Senior Lecturer, Stockholm University

How **good** is a clustering?

- Several metrics for assessing the quality of a cluster
- External evaluation
- Internal evaluation

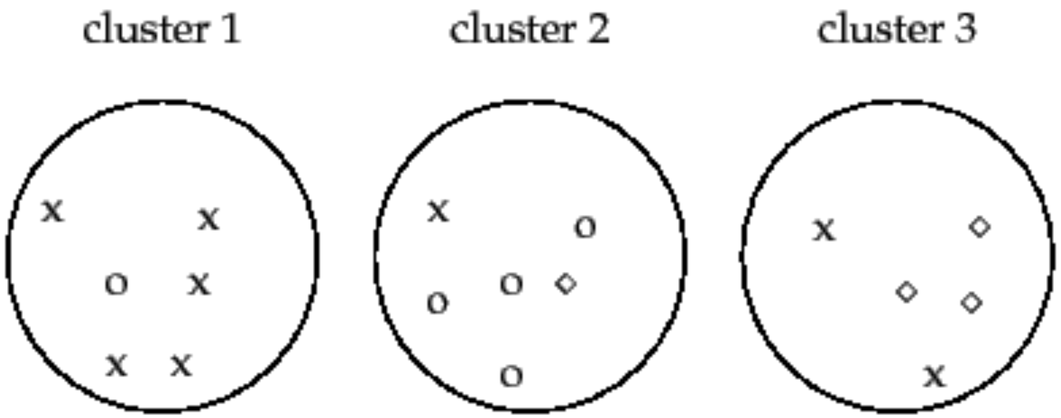
External Evaluation

- Clustering is evaluated based on external information, such as class labels of the clustered objects
- Metrics in this category typically assess how close is the clustering to the predefined classes
- Example of such measures:
  - Purity
  - Rand Index
  - Jaccard Index
  - Mutual Information

Cluster Purity

- Each cluster is assigned to the class label that is *most frequent* in the cluster

- Purity is measured by counting the *number of correctly assigned objects* and dividing by the *total number of objects*



Purity = (5 + 4 + 3) / 17 = 12 / 17 = 0.71

How good is purity?

- Cluster purity
  - Simple!
  - Intuitive
- Trade-off between number of clusters and purity?
  - If each object belongs to its own cluster, then purity is 1.0
  - Is that desirable?
- Cannot trade-off the quality of clusters against the number of clusters
- Alternatives:
  - Rand Index
  - Normalized Mutual Information

Rand Index

- Computes how similar are the clusters to a set of given class labels
- Measures the percentage of correct decisions taken by the clustering algorithm
- Decision: given a pair of objects, is it assigned to the same cluster?

	Same Cluster	Different Clusters
Same Class	TP	FN
Different Classes	FP	TN

TP: all pairs of objects assigned to the same cluster and also belong to the same class

TN: all pairs of objects assigned to different clusters and also belong different classes

FP: all pairs of objects assigned to the same cluster but belong to different classes

FN: all pairs of objects assigned to different clusters but belong to the same class

Rand Index

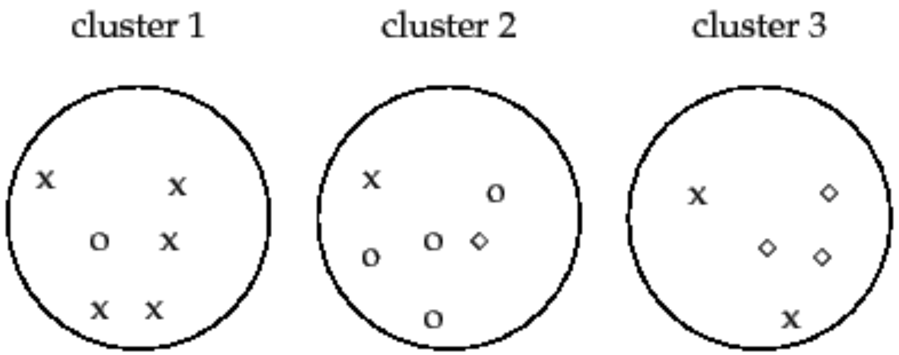
- Computes how similar are the clusters to a set of given class labels
- Measures the percentage of correct decisions taken by the clustering algorithm
- Decision: given a pair of objects, is it assigned to the same cluster?

	Same Cluster	Different Clusters
Same Class	TP	FN
Different Classes	FP	TN

$$\boldsymbol{Rand\ Index} = (\boldsymbol{TP} + \boldsymbol{TN}) \ / \ (\boldsymbol{TP} + \boldsymbol{FP} + \boldsymbol{FN} + \boldsymbol{TN})$$

Rand Index

	Same Cluster	Different Clusters
Same Class	TP	FN
Different Classes	FP	TN



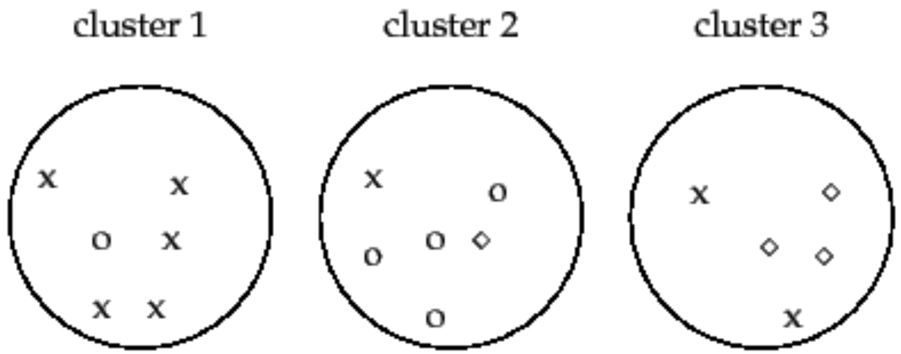
- **What are all the positives? TP + FP?**
- *all pairs that are correctly or incorrectly placed in the same cluster?*

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

$$\mathrm{TP} + \mathrm{FP} = \binom{6}{2} + \binom{6}{2} + \binom{5}{2} = 40$$

Rand Index

	Same Cluster	Different Clusters
Same Class	TP	FN
Different Classes	FP	TN



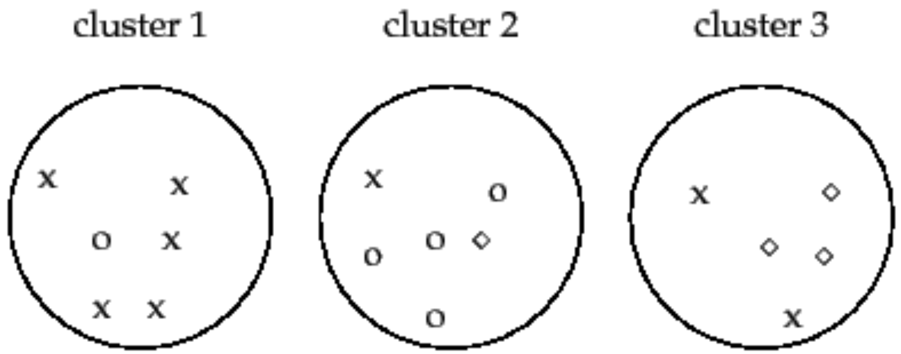
- **What are the TP?**
- *all pairs that are correctly placed in the same cluster?*

$$\mathrm{TP} = \binom{5}{2} + \binom{4}{2} + \binom{3}{2} + \binom{2}{2} = 20$$

Stockholms universitet

Rand Index

	Same Cluster	Different Clusters
Same Class	TP	FN
Different Classes	FP	TN



So:

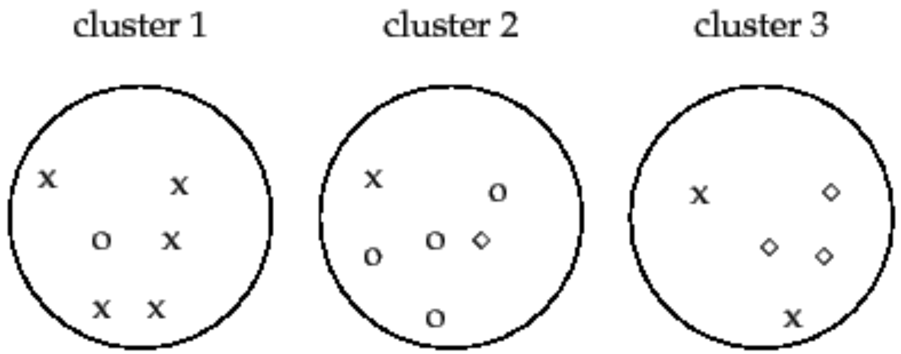
$$TP = 20$$

$$FP = 20$$

Stockholms universitet

Rand Index

	Same Cluster	Different Clusters
Same Class	TP	FN
Different Classes	FP	TN



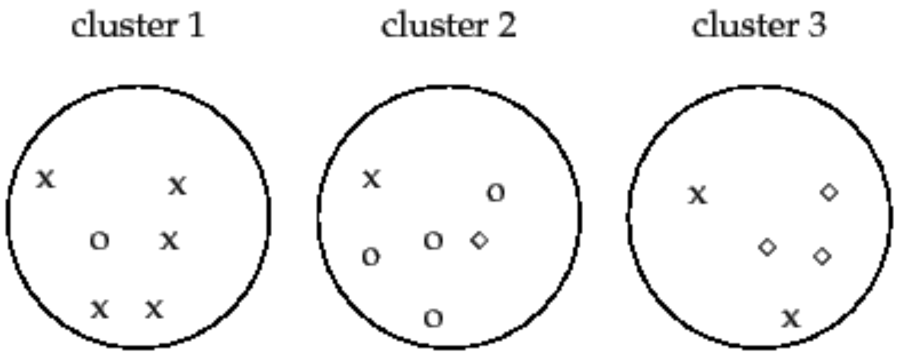
- **What are all the negatives? TN + FN?**

  all pairs that are correctly or incorrectly placed in different clusters?

$$TN + FN = 6 * 6 + 6 * 5 + 6 * 5 = 96$$

Rand Index

	Same Cluster	Different Clusters
Same Class	TP	FN
Different Classes	FP	TN



- **What are the FN?**

  all pairs that are incorrectly placed in a different cluster?

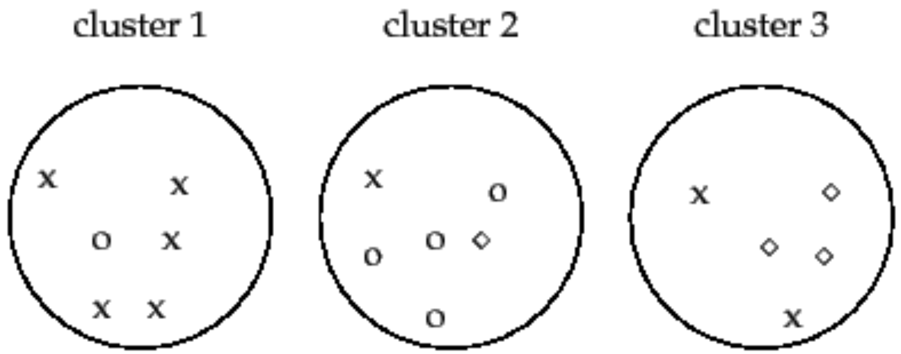
$$FN = (5+4) + (2+3) + (2*5) = 24$$

Hence: $TN = 96 - FN = 96 - 24 = 72$

Stockholms universitet

Rand Index

	Same Cluster	Different Clusters
Same Class	TP	FN
Different Classes	FP	TN



	Same cluster	Different clusters
Same class	TP = 20	FN = 24
Different classes	FP = 20	TN = 72

$$RI = (20 + 72) / (20 + 72 + 20 + 24) = 92/136 = 0.68$$

Internal Evaluation

- The clustering is evaluated based on merely the data that was used for the clustering
- Metrics in this category typically assess the intra-cluster and inter-cluster similarities
- Example of such measures:
  - Dunn Index
  - Silhouette coefficient

Stockholms universitet

The Dunn Index

- Does not assume any class distribution or assignment to the objects
- It is based on the principle objective of clustering:
  - Intra-cluster distance (or spread) is minimized
  - Inter-cluster distance is maximized
- Many ways of defining the spread (or diameter) of a cluster
  - Maximum distance between the objects
  - Mean distance between the objects
  - Total distance of all objects from their mean

The Dunn Index

- Many ways of quantifying the inter-cluster distance
  - Distance of the two closest objects
  - Distance of the two farthest objects
  - Distance between the centroids of the two clusters
- The family of Dunn Indices DI_m
  - Given a set of m clusters
  - Let $\delta(C_i, C_j)$ be a chosen inter-cluster metric for two clusters C_i and C_j
  - Let Δ_i be a chosen intra-cluster metric for cluster C_i

$$DI_m = \frac{\min\limits_{1 \leqslant i < j \leqslant m} \delta(C_i, C_j)}{\max\limits_{1 \leqslant k \leqslant m} \Delta_k}.$$

The Dunn Index

- In other words:
  - the Dunn Index is the minimum inter-cluster distance divided by the maximum intra-cluster distance
- Since the demoninator contains a ”max” term:
  - if all clusters are compact enough except for one
  - then DI will be relatively small
- Hence, it is a worst-case indicator and should be used with caution

$$DI_m = \frac{\min\limits_{1 \leqslant i < j \leqslant m} \delta(C_i, C_j)}{\max\limits_{1 \leqslant k \leqslant m} \Delta_k}.$$

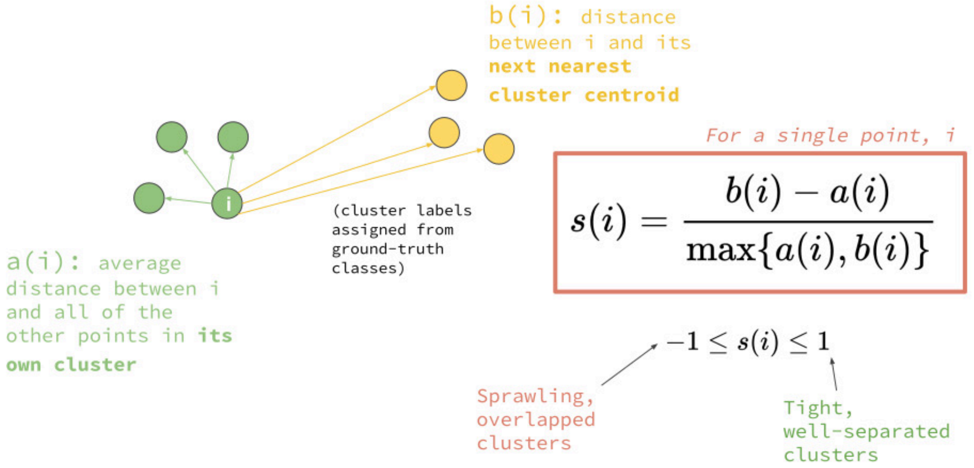
What if we don't have class labels?

- $a(i)$: *average dissimilarity (distance)* of object i with all other objects from the same cluster
- $b(i)$: the *lowest average dissimilarity (distance)* of object i to any other object in all clusters where i is not a member

- Takes values between -1 and 1

  if $b(i) \gg a(i)$ then $s(i) = 1$

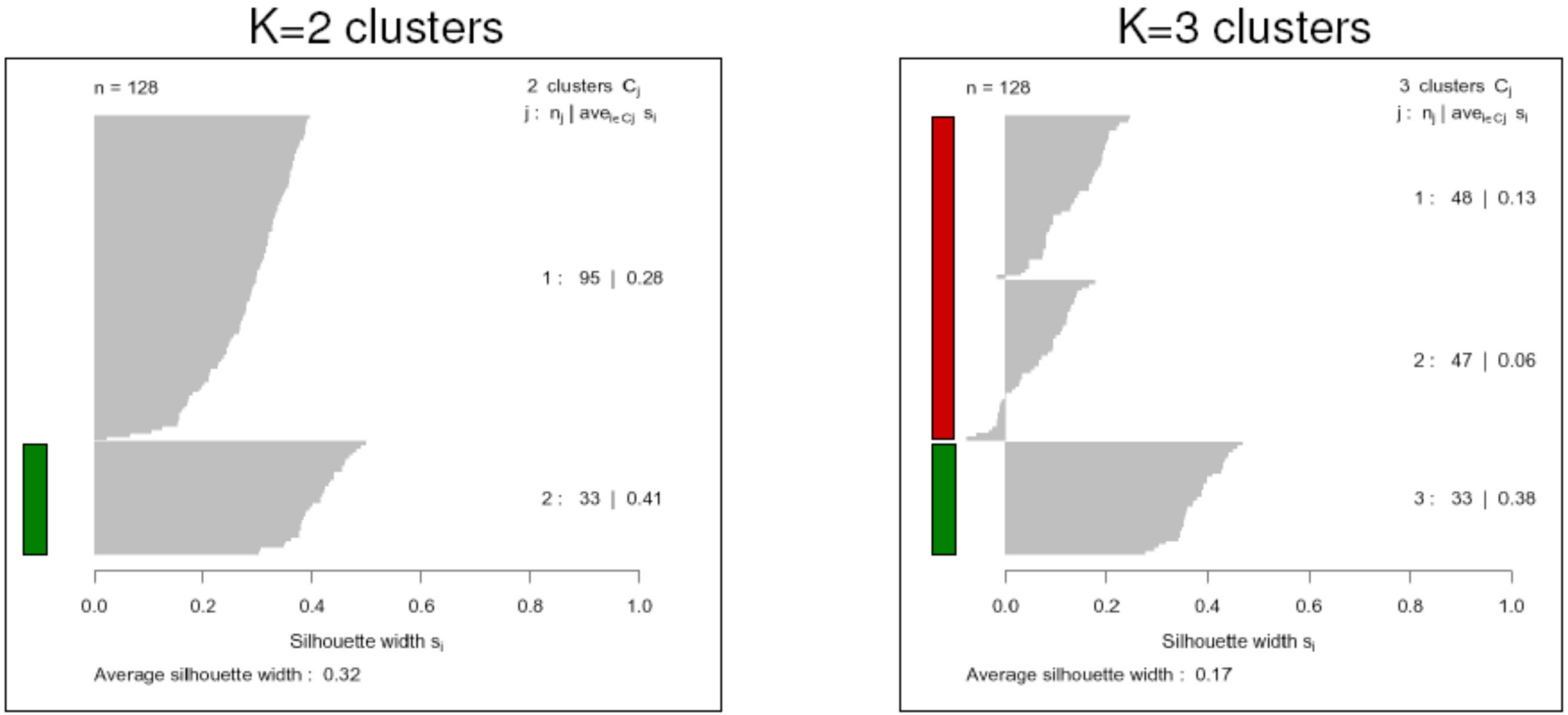
  if $b(i) \ll a(i)$ then $s(i) = -1$



- Total Silhouette: average silhouette value of all objects

Finding the correct number of clusters

- **Silhouete** can be used to identify the correct number of clusters
- How?



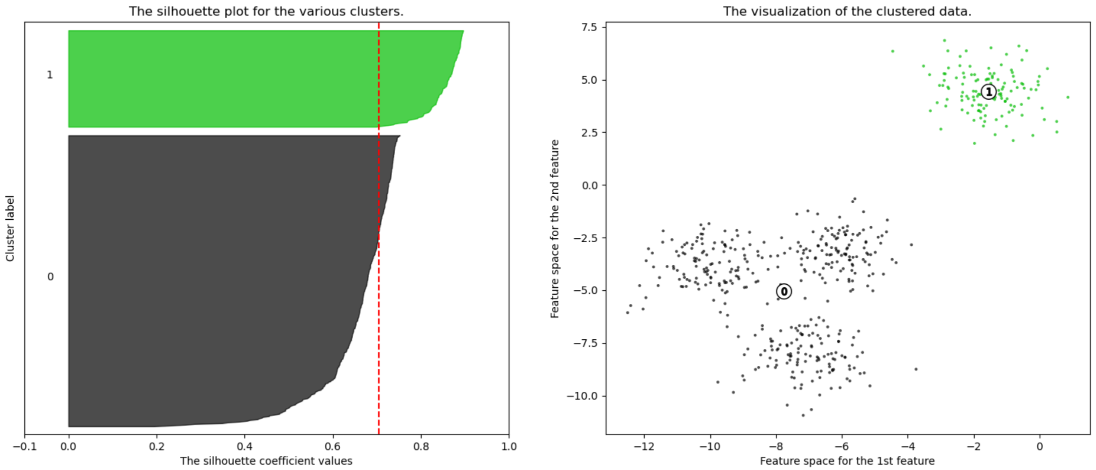
Green: Well separated cluster

Red: No clear cluster structure

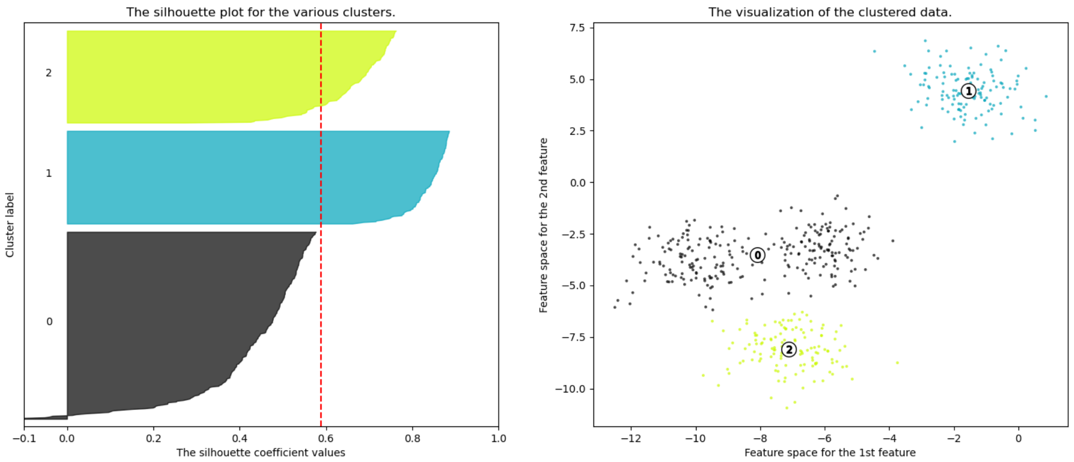
Stockholms universitet

Best number of clusters

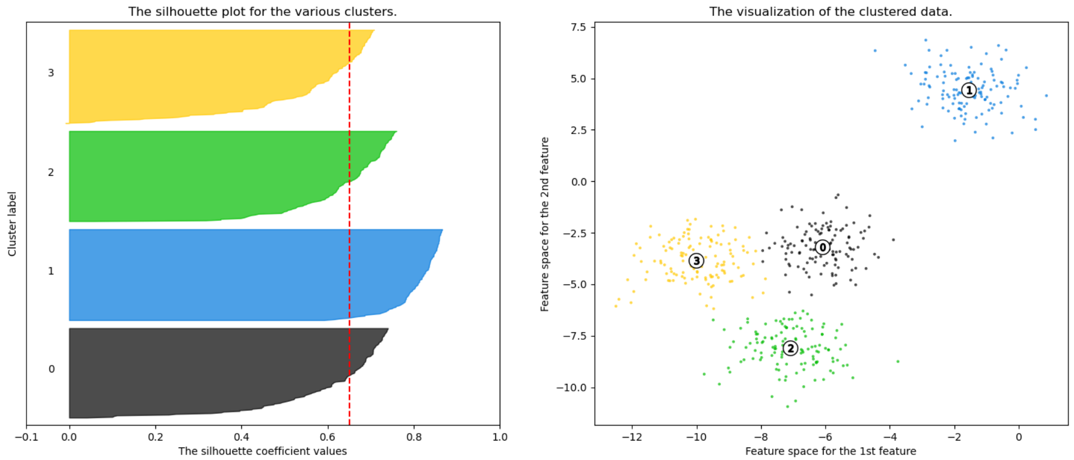
Silhouette analysis for KMeans clustering on sample data with n_clusters = 2



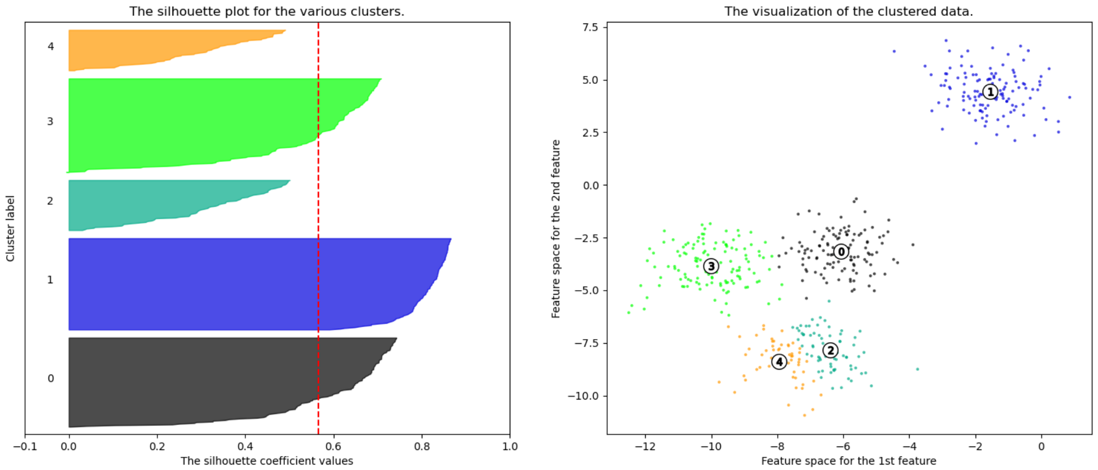
Silhouette analysis for KMeans clustering on sample data with n_clusters = 3



Silhouette analysis for KMeans clustering on sample data with n_clusters = 4



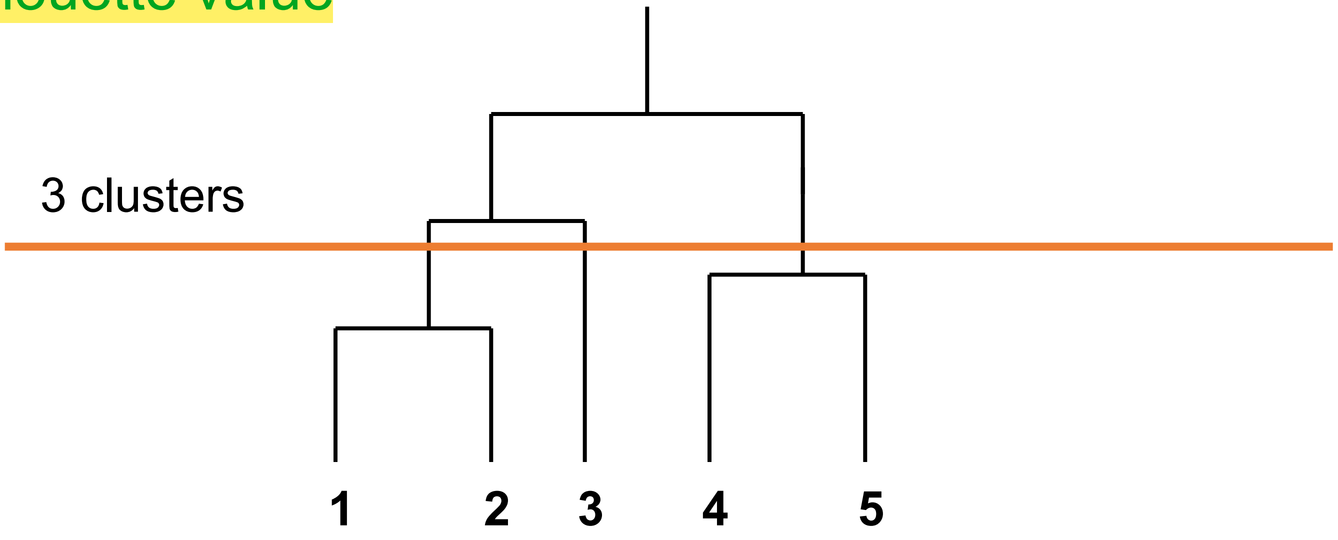
Silhouette analysis for KMeans clustering on sample data with n_clusters = 5



Finding the correct number of clusters

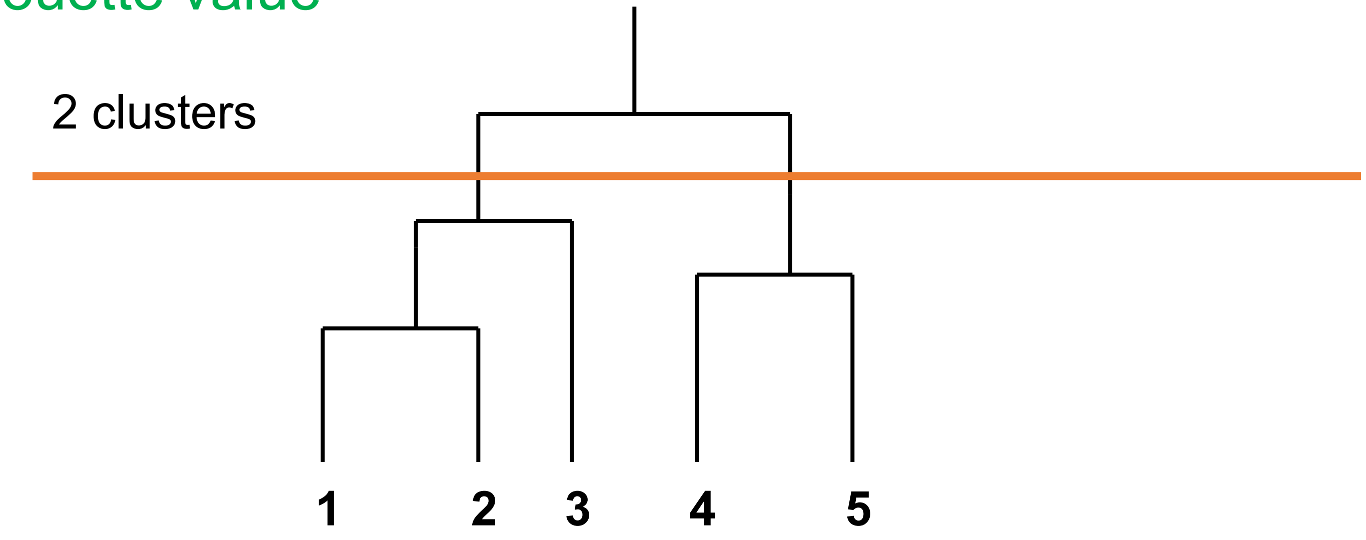
- K-means:
  - Run K-means with different number of values for K
  - Identify the clustering that produces the highest average Silhouette

- Hierarchical clustering:
  - You can choose a view of the dendrogram that provides the highest average Silhouette value



Finding the correct number of clusters

- K-means:
  - Run K-means with different number of values for K
  - Identify the clustering that produces the highest average Silhouette
- Hierarchical clustering:
  - You can choose a view of the dendrogram that provides the highest average Silhouette value



Stockholms universitet

Clustering aggregation

- Many different clusterings for the same dataset!
  - Different objective functions
  - Different algorithms
  - Different number of clusters

- Which clustering is the best?
  - **Aggregation:** we do not need to decide, but rather find a reconciliation between different outputs

The clustering-aggregation problem

- Input
  - $\boldsymbol{n}$ objects $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n\}$
  - $\boldsymbol{m}$ clusterings of the objects $\{\boldsymbol{C}_1, \dots, \boldsymbol{C}_m\}$
  - **clustering:** a collection of disjoint sets that cover $\boldsymbol{X}$

- Output
  - a **single clustering** $\boldsymbol{C}$, that is as close as possible to all input clusterings

- How do we measure ***closeness of clusterings***?
  - disagreement distance

Disagreement distance

- For two partitions C and P, and objects x, y in X define

$$I_{C,P}(x,y) = \begin{cases} 1 & \text{if } C(x) = C(y) \text{ and } P(x) \neq P(y) \\ & \text{OR} \\ & \text{if } C(x) \neq C(y) \text{ and } P(x) = P(y) \\ 0 & \text{otherwise} \end{cases}$$

U	C	P
x_1	1	1
x_2	1	2
x_3	2	1
x_4	3	3
x_5	3	4

- if $I_{C,P}(x,y) = 1$ we say that x, y create a disagreement between clusterings C and P

$D(C,P) = 3$

- **The disagreement distance between C and P is:**

$$D(C,P) = \sum_{(x,y)} I_{C,P}(x,y)$$

Stockholms universitet

Clustering aggregation

- Given m clusterings $C_1, \ldots, C_m$ find C such that

$$D(C) = \sum_{i=1}^{m} D(C, C_i)$$

aggregation cost

is minimized

U	C_1	C_2	C_3	C
x_1	1	1	1	1
x_2	1	2	2	2
x_3	2	1	1	1
x_4	2	2	2	2
x_5	3	3	3	3
x_6	3	4	3	3

Why clustering aggregation?

- Clustering categorical data

U	*City*	*Profession*	*Nationality*
x_1	New York	Doctor	U.S.
x_2	New York	Teacher	Canada
x_3	Boston	Doctor	U.S.
x_4	Boston	Teacher	Canada
x_5	Los Angeles	Lawer	Mexican
x_6	Los Angeles	Actor	Mexican

- Consider each categorical attribute as a cluster
- Merge clusters to an aggregate clustering

Why clustering aggregation?

- Detect outliers
  - outliers are defined as points for which there is no consensus by the clusterings
- Improve the robustness of clustering algorithms
  - different algorithms have different weaknesses
  - combining them can produce a better result

Complexity of Clustering Aggregation

- The clustering aggregation problem is NP-hard
  - the median partition problem [Barthelemy and LeClerc 1995]
- Look for heuristics and approximate solutions
- A simple 2-approximation algorithm: **BEST**
  - select among the input clusterings the clustering C^* that minimizes $D(C^*)$

Today …

- Metrics for assessing the **quality** of a cluster
- What is **external evaluation**?
- What is **internal evaluation**
- Finding the correct **number** of clusters
- Clustering **aggregation**

TODOs

Reading:
Main course book: chapter 7

Lab 2
Sep 14

Quiz 2

Coming up next

Thursday

Lab 2 – Clustering using Python

Lecture 6 – Classification I

Friday

Thanks!

golnaz.taheri@dsv.su.se