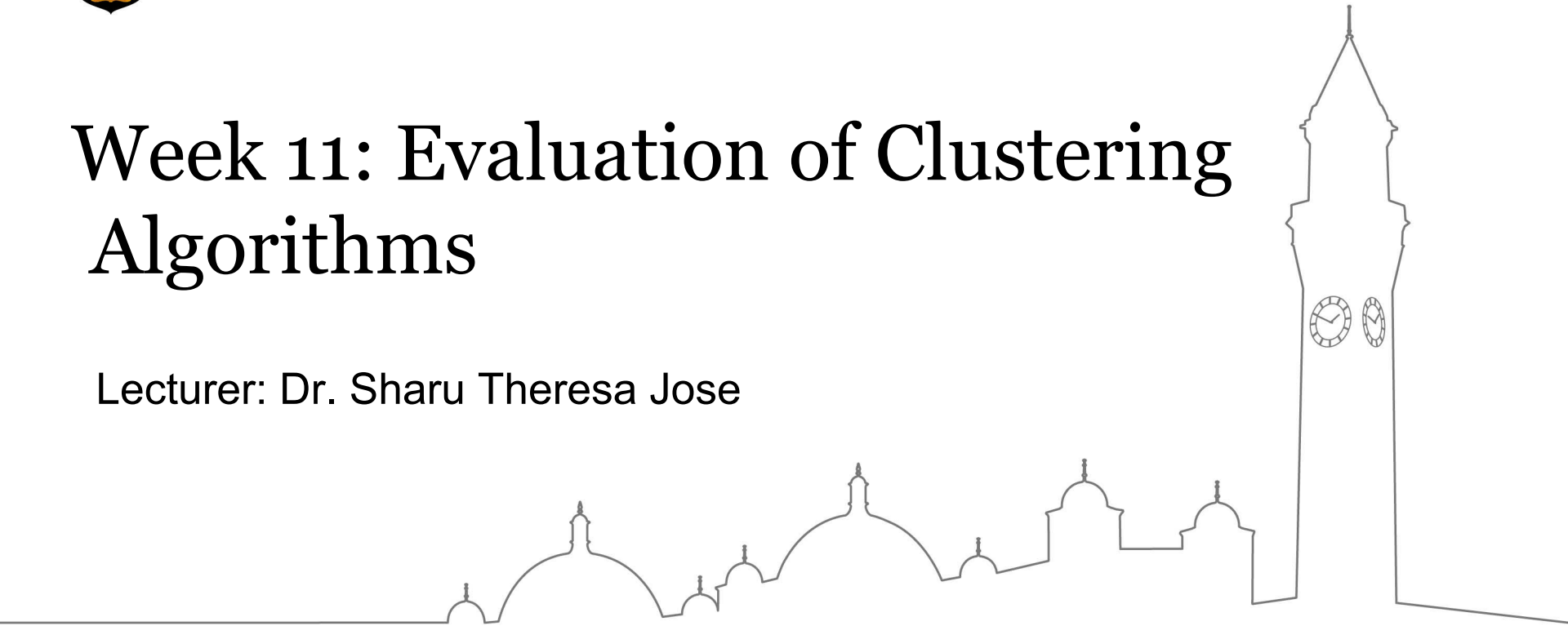
UNIVERSITY OF BIRMINGHAM

Week 11: Evaluation of Clustering Algorithms

Lecturer: Dr. Sharu Theresa Jose



Learning Outcomes

- Understand the importance of cluster analysis
- Familiarize with some commonly used cluster validation criteria

Overview of Lecture

- Introduction
  - What is cluster evaluation?
- Cluster Validation Criteria
  - Unsupervised, Supervised and Relative Validation Criteria
- Unsupervised Validation Criteria
- Supervised Validation Criteria

Introduction

- Supervised learning has well-accepted evaluation measures and procedures (e.g., accuracy, cross-validation).
- In contrast, cluster evaluation (or validation) is not trivial.
- It may even look like different clustering algorithms might need different evaluation criteria.
  - Example for a cluster validation measure:

$$dissimilarity\ or\ WCSS(\boldsymbol{C}) = \sum_{C \in \boldsymbol{C}} variability(C)\,,\ \text{where}$$

$$variability\ or\ inertia(C) = \sum_{e \in C} distance(e, centroid(C))\,.$$

  - Suitable for K-means, but not for hierarchical clustering.
- Nevertheless, cluster validation is important: every clustering algorithm will find clusters in a dataset, even if data has no natural clustering structure.

Cluster Validation Can Help Answer….

- Is there a clustering tendency in the observed data, i.e., determine whether non-random structure (or natural grouping) exists in the data?
- Can we evaluate how well the results of a clustering algorithm fit the data (or natural grouping) **without** external information?
- Can we evaluate how well the results of a clustering algorithm fit the data **with** external information?
- Can we compare two sets of clusters to determine which is better?
- Can we determine the correct number of clusters?

What is cluster validation?

- Goal: evaluate in a **quantitative** and **objective** manner the cluster structure found by an algorithm according to a validation criterion

- Validation criterion : Index used to measure the adequacy of the found cluster structures.

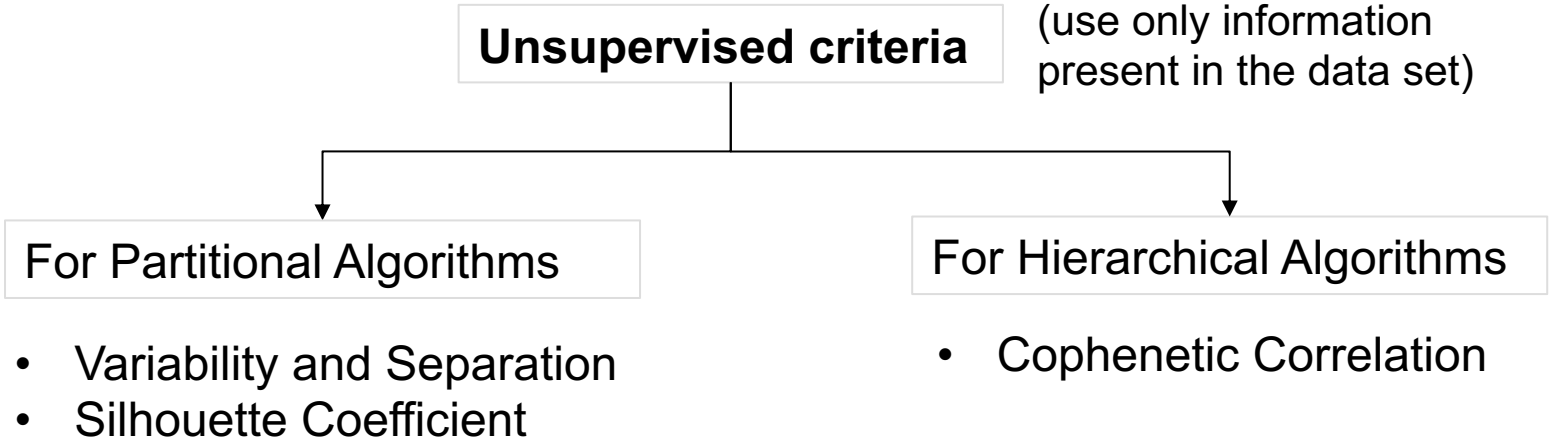
- Adequacy refers to the sense in which the found cluster structure provides true information about the data or reflect the intrinsic character of the data.

Types of Cluster Validity Criteria

- Unsupervised (Internal Indices)
  - Measures **goodness** of a clustering structure **without** reference to external information Example: $dissimilarity(\boldsymbol{C})$.
  - Can be further divided into two classes: intra-cluster and inter-cluster similarity indices.
  - Can also be used to estimate the optimal number of clusters (e.g., elbow method).
- Supervised (External Indices)
  - Measures the extent to which a clustering algorithm **matches** some external structure.
  - Example: Entropy (how well cluster labels match with externally supplied class labels).
- Relative
  - Compares two different sets of clusters or algorithms.
  - Can be a supervised or unsupervised criteria used for the purpose of comparison.
  - Example: Two K-means clusterings can be compared using either dissimilarity (WCSS) or entropy.

UNIVERSITY OF BIRMINGHAM

Unsupervised Validity Criteria

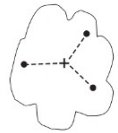


Variability and Separation Criteria

- Based on the notions of inter-cluster and intra-cluster dissimilarity.
- Centroid-based variability and separation criteria: With appropriate choice of distance function $d(\cdot,\cdot)$, we have

Intra-cluster dissimilarity of cluster C:

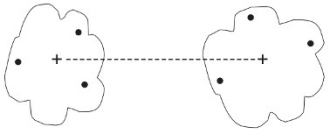
$$inertia\ or\ variability(C) = \sum_{e \in C} d(e, centroid(C))$$



$Variability(C)$

Inter-cluster dissimilarity between clusters C_1, C_2:

$$separation_c(C_1, C_2) = d(centroid(C_1), centroid(C_2))$$



$Separation_c(C_1, C_2)$

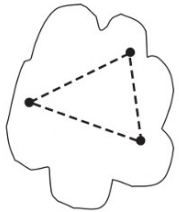
Cluster dissimilarity of C_1 with respect to the whole data:

$$separation_c(C_1) = d(centroid(C_1), centroid(data))$$

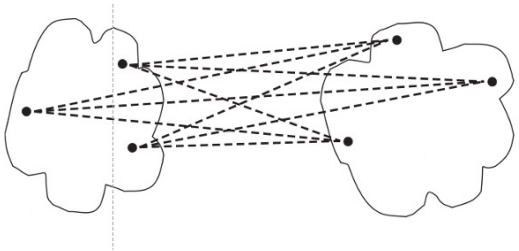
(Note: $separation_c(C_1)$ is used to compute BCSS)

- Distance-based variability and separation criteria:

$$variability_d(C) = \sum_{e_1, e_2 \in C} d(e_1, e_2)$$ (intra-cluster dissimilarity of cluster C)



$$separation_d(C_1, C_2) = \sum_{e_1, \in C_1, e_2 \in C_2} d(e_1, e_2)$$ (inter-cluster dissimilarity between clusters C_1, C_2)



- An overall cluster validity for a set of K clusters can be defined as a weighted sum of individual clusters,

$$overall\ validity = \sum_{i=1}^{K} w_i\ validity(C_i),$$

where $validity(C_i)$ can be variability or separation criteria or a combination.

Overall validity	Cluster weight (w_i)	$validity(C_i)$
$dissimilarity(\boldsymbol{C})$ or WCSS($\boldsymbol{C}$)	1	$variability(C_i)$
$BCSS(\boldsymbol{C})$	$number\ of\ examples\ in\ C_i$	$separation_c(C_i)$

- Overall validity can be used to estimate the number of clusters.
  - Example: $dissimilarity(\boldsymbol{C})$ estimates the number of clusters via elbow method in K-means

Relationship between dissimilarity and overall separation for squared Euclidean distance

- Distance function: squared Euclidean distance
- It can be verified that

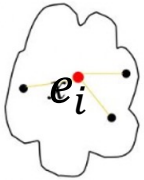
$$WCSS(\boldsymbol{C}) + BCSS(\boldsymbol{C}) = \sum_{e} d\big(e, centroid(data)\big) = a\ constant$$

- Minimizing WCSS (via cluster assignments) equals maximizing BCSS.

Silhouette Coefficient (SC)

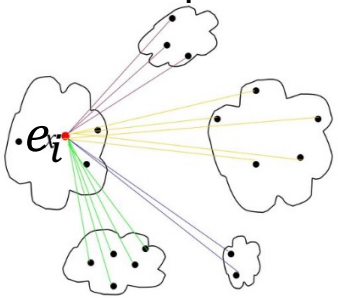
- Combines ideas of variability and separation for individual example, for clusters as well as for clusterings.
- Computing SC for an individual example: Let ith example e_i belongs to cluster C.
  - Calculate a_i = average distance of ith example to all other examples in its cluster, i.e.,

$$a_i = \frac{\sum_{e \in C, e \neq e_i} d(e_i, e)}{no.of\ examples\ in\ C - 1}.$$



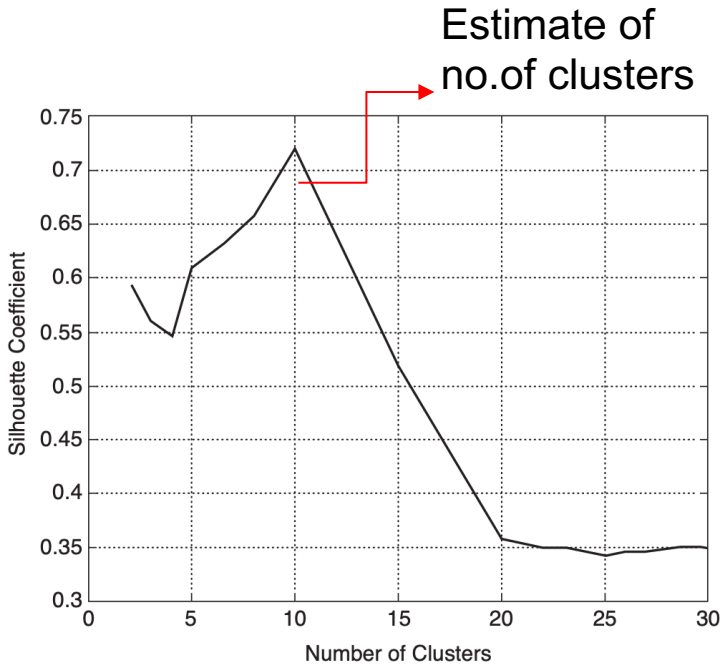
  - Calculate b_i = min (average distance of ith example to examples in another cluster), i.e.,

$$b_i = \min_{k=1,..K, C_k \neq C} \frac{\sum_{e \in C_k} d(e_i, e)}{no.of\ examples\ in\ C_k}.$$



- SC for ith example is $s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$.

- SC can vary between -1 and 1.
  - SC=-1: $(a_i > b_i = 0) \Rightarrow$ data is better fit to a neighboring cluster
  - SC=0: $(a_i = b_i) \Rightarrow$ data is on the border between two clusters
  - SC=1: $(0 = a_i < b_i) \Rightarrow$ data is well-matched to cluster
- SC of a cluster (or clustering) = average of SCs of examples in the cluster (or clustering)
- Average SC of a clustering structure can be used to estimate the optimal number of clusters in the data set.
  - Plot the average SC of clustering as a function of number of clusters.
  - Peak in the plot gives an estimate of the number of clusters.



Unsupervised Validity Criteria for Hierarchical Clustering

- Cophenetic distance between two examples = Distance at which an agglomerative clustering algorithm puts the examples in the same cluster for the first time.
- Results in a cophenetic distance matrix P whose (i, j)th entry, i.e., entry in the ith row and jth column of matrix P, is the cophenetic distance between ith and jth examples.
- For N examples, cophenetic distance matrix is a N x N matrix.
- Note that this matrix is different from the distance matrix D that summarizes the distance relationship between the data points.

- Let $P_{i,j}$ denote the (i,j)th entry of matrix P and let $D_{i,j}$ denote the (i,j)th entry of the distance matrix D.

- **CoPhenetic Correlation Coefficient (CPCC)** is the correlation between the entries of the distance matrix D and the cophenetic distance matrix P:

$$CPCC = \frac{\sum_{i,j=1..N,\, s.t\ i<j}(D_{i,j} - d)(P_{i,j} - p)}{\sqrt{\sum_{i,j=1..N,\, s.t\ i<j}(D_{i,j} - d)^2 \sum_{i,j=1..N,\, s.t\ i<j}(P_{i,j} - p)^2}},$$

where d and p respectively denote the average of the non-zero entries of the matrices D and P.

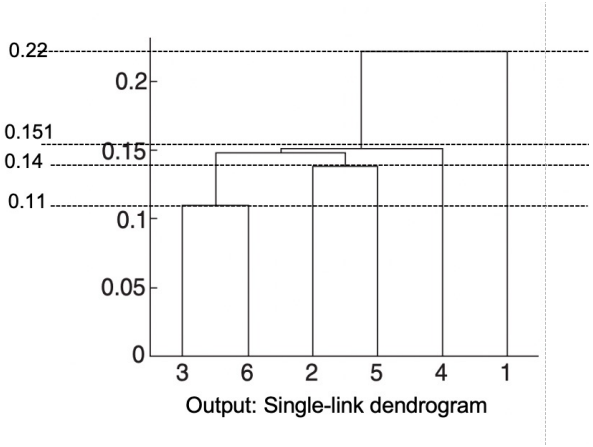
- CPCC is a measure of how well hierarchical clustering fits the data; the higher (or closer to 1) the better.

Example 1: Single-Linkage Hierarchical Clustering

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.151	0.28	0.11
p4	0.37	0.20	0.151	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Table 8.4. Euclidean distance matrix for 6 points.

Input: Distance matrix, D



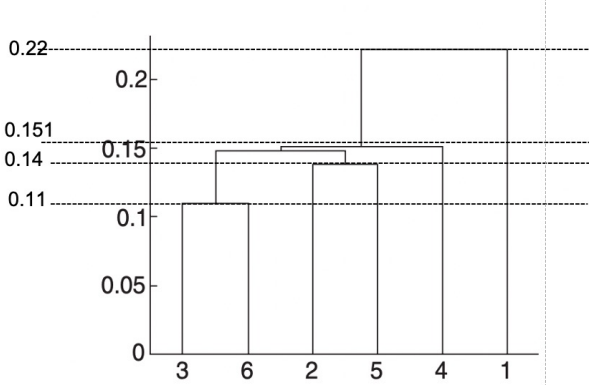
Output: Single-link dendrogram

Question: Compute the cophenetic distance matrix.

Example 1

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.151	0.28	0.11
p4	0.37	0.20	0.151	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Table 8.4. Euclidean distance matrix for 6 points.



Output: Single-link dendrogram

Solution:

$P =$

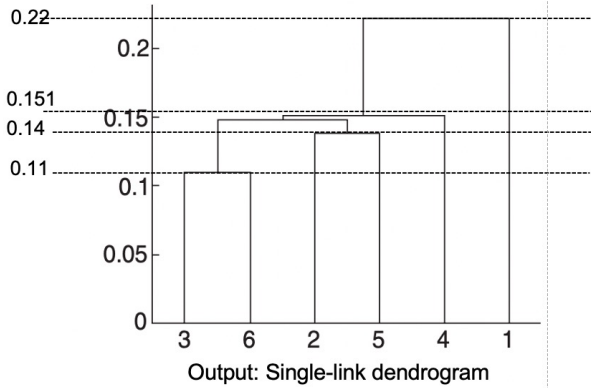
	p1	p2	p3	p4	p5	p6
p1	0	?	?	?	?	?
p2		0	?	?	?	?
p3			0	?	?	?
p4				0	?	?
p5					0	?
p6						0


UNIVERSITY OF BIRMINGHAM

Example 1

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.151	0.28	0.11
p4	0.37	0.20	0.151	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Table 8.4. Euclidean distance matrix for 6 points.



Output: Single-link dendrogram

Solution:

$P =$

	p1	p2	p3	p4	p5	p6
p1	0	0.22	0.22	0.22	0.22	0.22
p2		0	0.15	0.151	0.14	0.15
p3			0	0.151	0.15	0.11
p4				0	0.151	0.151
p5					0	0.15
p6						0

- Compute CPCC for the above problem

D =

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.151	0.28	0.11
p4	0.37	0.20	0.151	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Table 8.4. Euclidean distance matrix for 6 points.

P =

	p1	p2	p3	p4	p5	p6
p1	0	0.22	0.22	0.22	0.22	0.22
p2		0	0.15	0.151	0.14	0.15
p3			0	0.151	0.15	0.11
p4				0	0.151	0.151
p5					0	0.15
p6						0

Compute $d = \frac{0.24+0.22+0.37+0.34+0.23+0.15+0.2+0.14+0.25+0.151+0.28+0.11+0.29+0.22+0.39}{15}$ = 0.239

Compute p $= \frac{0.22*5+0.15+0.151+0.14+0.15+0.151+0.15+0.11+0.151+0.151+0.15}{15}$ =0.170

$$CPCC = \frac{\sum_{i,j=1..N,\ s.t\ i<j}(D_{i,j} - d)(P_{i,j} - p)}{\sqrt{\sum_{i,j=1..N,\ s.t\ i<j}(D_{i,j} - d)^2 \sum_{i,j=1..N,\ s.t\ i<j}(P_{i,j} - p)^2}},$$

To compute denominator,

	p1	p2	p3	p4	p5	p6
p1	0	$(0.24-d)^2$	$(0.22-d)^2$	$(0.37-d)^2$	$(0.34-d)^2$	$(0.23-d)^2$
p2		0	$(0.15-d)^2$	$(0.20-d)^2$	$(0.14-d)^2$	$(0.25-d)^2$
p3			0	$(0.151-d)^2$	$(0.28-d)^2$	$(0.11-d)^2$
p4				0	$(0.29-d)^2$	$(0.22-d)^2$
p5					0	$(0.39-d)^2$
p6						0

Updated Distance matrix, D

Sum of all the entries of the above matrix = sum_D

	p1	p2	p3	p4	p5	p6
p1	0	$(0.22-p)^2$	$(0.22-p)^2$	$(0.22-p)^2$	$(0.22-p)^2$	$(0.22-p)^2$
p2		0	$(0.15-p)^2$	$(0.151-p)^2$	$(0.14-p)^2$	$(0.15-p)^2$
p3			0	$(0.151-p)^2$	$(0.15-p)^2$	$(0.11-p)^2$
p4				0	$(0.151-p)^2$	$(0.151-p)^2$
p5					0	$(0.15-p)^2$
p6						0

Updated Cophenetic distance matrix, P

Sum of all the entries of the above matrix = sum_P

$$denominator = \sqrt{sum_P * sum_D}$$

UNIVERSITY OF BIRMINGHAM

To compute numerator,

	p1	p2	p3	p4	p5	p6
p1	0	$(0.24 - d)$ (cell D_{12})	$(0.22 - d)$ (cell D_{13})	$(0.37 - d)$ (cell D_{14})	$(0.34 - d)$ (cell D_{15})	$(0.23 - d)$ (cell D_{16})
p2		0	$(0.15 - d)$ (cell D_{23})	$(0.20 - d)$ (cell D_{24})	$(0.14 - d)$ (cell D_{25})	$(0.25 - d)$ (cell D_{26})
p3			0	$(0.151 - d)$ (cell D_{34})	$(0.28 - d)$ (cell D_{35})	$(0.11 - d)$ (cell D_{36})
p4				0	$(0.29 - d)$ (cell D_{45})	$(0.22 - d)$ (cell D_{46})
p5					0	$(0.39 - d)$ (cell D_{56})
p6						0

Updated Distance matrix, D

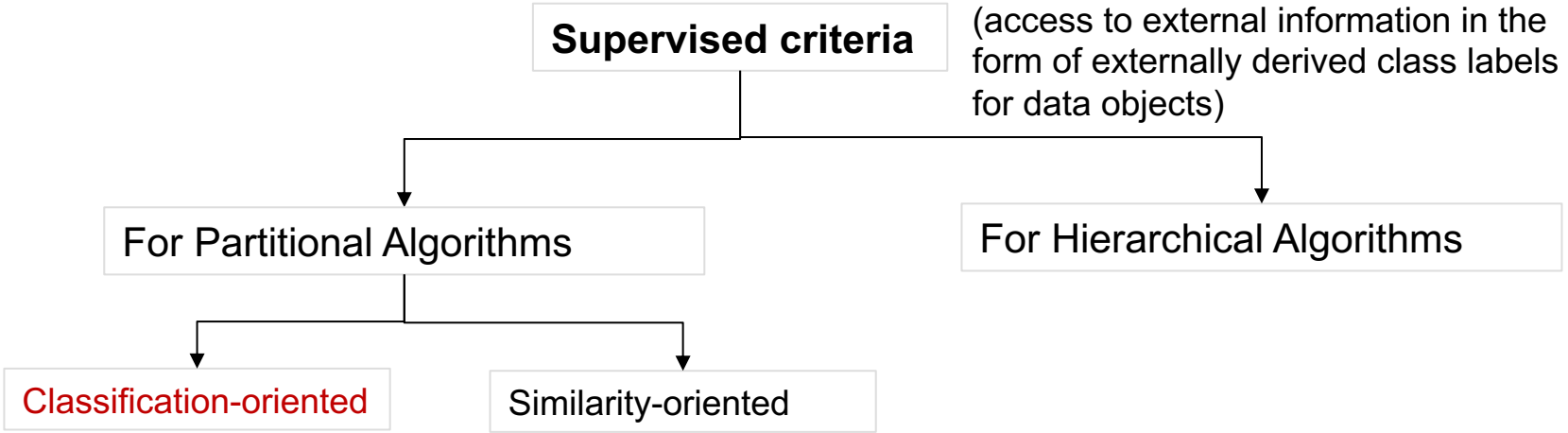
	p1	p2	p3	p4	p5	p6
p1	0	$0.22 - p$ (cell P_{12})	$0.22 - p$ (cell P_{13})	$0.22 - p$ (cell P_{14})	$0.22 - p$ (cell P_{15})	$0.22 - p$ (cell P_{16})
p2		0	$0.15 - p$ (cell P_{23})	$0.151 - p$ (cell P_{24})	$0.14 - p$ (cell P_{25})	$0.15 - p$ (cell P_{26})
p3			0	$0.151 - p$ (cell P_{34})	$0.15 - p$ (cell P_{35})	$0.11 - p$ (cell P_{36})
p4				0	$0.151 - p$ (cell P_{45})	$0.151 - p$ (cell P_{46})
p5					0	$0.15 - p$ (cell P_{56})
p6						0

Updated Cophenetic distance matrix, P

$$numerator = P_{12}D_{12} + P_{13}D_{13} + P_{14}D_{14} + P_{15}D_{15} + P_{16}D_{16} + P_{23}D_{23} + P_{24}D_{24} + P_{25}D_{25} + P_{26}D_{26} + P_{34}D_{34} + P_{35}D_{35} + P_{36}D_{36} + P_{45}D_{45} + P_{46}D_{46} + P_{56}D_{56}$$

$$CPCC = \frac{numerator}{denominator} = 0.4599$$

Supervised Validity Criteria



- ❑ uses measures from classification
- ❑ extent to which a cluster contains objects of a single class
  - Entropy
  - Purity
  - Precision, Recall, F-measure

- ❑ related to similarity measures for binary data
- ❑ extent to which two objects in the same class are in the same cluster and vice versa
  - Jaccard measure
  - Rand statistic

UNIVERSITY OF BIRMINGHAM

Classification-Oriented Validity Measures

- Uses externally derived class labels for data examples.
- Example: confusion matrix for the output of clustering algorithm on LA Times dataset

Predicted Cluster labels

True Class labels (externally supplied)	Cluster 1	Cluster 2	Cluster 3	Total
Entertainment	10	11	50	71
Finance	15	60	13	88
Foreign	20	21	9	50
Metro	3	15	2	20
National	45	2	11	58
Sports	12	28	56	96
Total	105	137	141	383

Confusion Matrix for the output of Clustering Algorithm on LA Times Dataset

Each entry: number of objects in a cluster that belongs to the corresponding class

- Let L denote the number of classes and K denote the number of clusters.
- Probability that an example of cluster i belongs to class j:

$$p_{i,j} = \frac{no.of\ examples\ of\ class\ j\ in\ cluster\ i}{no.of\ examples\ in\ cluster\ i}$$

Precision, Recall and F-measure

- Precision of cluster i with respect to class j:

$$precision(i,j) = p_{i,j}$$

  - Measures extent to which a cluster contains objects of a single class
- Recall of cluster i with respect to class j :

$$recall(i,j) = \frac{number\ of\ objects\ of\ class\ j\ in\ cluster\ i}{number\ of\ objects\ in\ class\ j}$$

  - Determines the fraction of class j contained in cluster i
- F-measure of cluster i with respect to class j :

$$F(i,j) = \frac{2 * precision(i,j) * recall(i,j)}{precision(i,j) + recall(i,j)}$$

  - Measures extent to which a cluster contains only objects of a particular class and all objects of that class
  - Combination of both precision and recall

Entropy

- Degree to which each cluster consists of examples of a single class
- Entropy of ith cluster:

$$e_i = -\sum_{j=1}^{L} p_{i,j} \log_2 p_{i,j}$$

- Total entropy of a set of clusters:

$$e = \sum_{i=1}^{K} \frac{no.of\ examples\ in\ cluster\ i}{total\ no.of\ examples} e_i$$

- Low entropy 

Purity

- Another measure of the extent to which a cluster consists of examples of a single class.
- Purity of ith cluster:

$$p_i = \max_j p_{i,j}$$

- Overall purity of a set of clusters:

$$p = \sum_{i=1}^{K} \frac{no.of\ examples\ in\ cluster\ i}{total\ no.of\ examples} p_i$$

- High purity ✓

Example:

$j = 1$ $j = 2$ $j = 3$ $j = 4$ $j = 5$ $j = 6$

Table 8.9. K-means clustering results for the *LA Times* document data set.

Cluster	Entertainment	Financial	Foreign	Metro	National	Sports	Entropy	Purity
1	3	5	40	506	96	27	1.2270	0.7474
2	4	7	280	29	39	2	1.1472	0.7756
3	1	1	1	7	4	671	0.1813	0.9796
4	10	162	3	119	73	2	1.7487	0.4390
5	331	22	5	70	13	23	1.3976	0.7134
6	5	358	12	212	48	13	1.5523	0.5525
Total	354	555	341	943	273	738	1.1450	0.7203

Best cluster

$$p_{1,1} = \frac{3}{3+5+40+506+96+27} = \frac{3}{677} = 0.0044 \qquad p_{1,2} = \frac{5}{677} = 0.0073 \qquad p_{1,3} = \frac{40}{677} = 0.0590$$

$$p_{1,4} = \frac{506}{677} = 0.7474 \qquad p_{1,5} = \frac{96}{677} = 0.1418 \qquad p_{1,6} = \frac{27}{677} = 0.0398 \qquad purity\ of\ cluster\ 1 = 0.7474$$

$$entropy\ of\ cluster\ 1$$
$$= -0.0044 \log_2 0.0044 - 0.0073 \log_2 0.0073 - 0.0590 \log_2 0.0590 - 0.7474 \log_2 0.7474 - 0.1418 \log_2 0.1418 - 0.0398 \log_2 0.0398 = 1.2270$$

$precision(1,4) = 0.7474$, recall(1,4) = 506/943 = 0.5365, F(1,4)=0.8019/1.2839 = 0.624

UNIVERSITY OF BIRMINGHAM

Similarity-Oriented Measures

- Measures the extent to which two examples in the same class belong to the same cluster and vice versa.
- Comparison of two N xN matrices:
  - Ideal cluster similarity matrix: has 1 in the (i, j)th entry if two examples $i \; and \; j$ are in the same cluster, and 0 otherwise.
  - Ideal class similarity matrix: has 1 in the (i, j)th entry if two examples $i \; and \; j$ are in the same class, and 0 otherwise.
- Can use two measures:
  - Measure 1: Compute the correlation between the above two matrices.
  - Measure 2: Binary similarity-based measures

Binary Similarity Based Measures:

Compute the following quantities:

f_{00} = number of pairs of objects having a different class and a different cluster
f_{01} = number of pairs of objects having a different class and the same cluster
f_{10} = number of pairs of objects having the same class and a different cluster
f_{11} = number of pairs of objects having the same class and the same cluster

	Same Cluster	Different Cluster
Same Class	f_{11}	f_{10}
Different Class	f_{01}	f_{00}

- Rand statistic: $$\text{Rand statistic} = \frac{f_{00} + f_{11}}{f_{00} + f_{01} + f_{10} + f_{11}}$$

- Jaccard coefficient: $$\text{Jaccard coefficient} = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

- Measures how well the obtained clustering agrees with the original class labels.
- Takes value between 0 and 1; higher the better

Example:

Table 8.10. Ideal cluster similarity matrix.

Point	p1	p2	p3	p4	p5
p1	1	1	1	0	0
p2	1	1	1	0	0
p3	1	1	1	0	0
p4	0	0	0	1	1
p5	0	0	0	1	1

Table 8.11. Ideal class similarity matrix.

Point	p1	p2	p3	p4	p5
p1	1	1	0	0	0
p2	1	1	0	0	0
p3	0	0	1	1	1
p4	0	0	1	1	1
p5	0	0	1	1	1

Compute the Rand statistic and Jaccard coefficient based on the above tables.

Solution: f_{00}=4, $f_{01} = 2, f_{10} = 2, f_{11} = 2$
Rand statistic = (2+4)/(4+2+2+2)=0.6

Jaccard coefficient = 2/(2+2+2)=0.33

Final Remarks

- Measures of clustering tendency:
  - Evaluate whether a data set has clusters without clustering
  - Example: Hopkins Statistic
- There is more to cluster evaluation and is an active area of research.
  - Assessing the significance of cluster validity measures: The validity criteria discussed above give a single number as a measure of goodness of cluster. How to interpret the significance of this number?
  - Naïve solution – define the range of cluster validity criteria and use statistics to evaluate whether the value we have obtained is unusually low or high.

Reference

- Introduction to Data Mining, by Tan, Steinbach and Kumar - Chapter 8