Cluster Analysis

- Cluster Validation
- Determining Number of Clusters

Cluster Validation

- The procedure of evaluating the results of a clustering algorithm is known under the term cluster validity.
- How do we evaluate the "goodness" of the resulting clusters?
- Then why do we want to evaluate them?
  - To avoid finding patterns in noise
  - To compare clusterings, or clustering algorithms

Cluster Validation

- Determining the clustering tendency of a set of data.
- Comparing the results of a cluster analysis to externally known results.
- Evaluating how well the results of a cluster analysis fit the data without reference to external information.
- Comparing the results of two different sets of cluster analyses to determine which is better.
- Determining the 'correct' number of clusters.

3

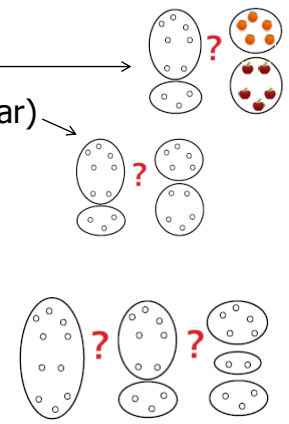
Cluster Validation

- Measures that are applied to judge various aspects of cluster validity, are classified into the following two types:
  - External Index: Used to measure the extent to which cluster labels match externally supplied class labels.
    - E.g., entropy, precision, recall
  - Internal Index: Used to measure the goodness of a clustering structure without reference to external information.
    - E.g., Sum of Squared Error (SSE)

4

Cluster Validation

- External Index
  - Validate against ground truth
  - Compare two clusters: (how similar)
- Internal Index
  - Validate without external info
  - With different number of clusters
  - Solve the number of clusters



5

External Index

- Assume that the data is labeled with some class labels.
  - This is called the "ground truth".
- It is wanted the clusters to be homogeneous with respect to classes.
- External measures are based on a matrix that summarize the number of correct predictions and wrong predictions.

6

External Index

- n = number of points
- m_i = points in cluster i
- c_j = points in class j
- n_{ij}= points in cluster i coming from class j
- $p_{ij} = n_{ij}/m_i$= probability of element from cluster i to be assigned in class j

	Class 1	Class 2	Class 3	
Cluster 1	n_{11}	n_{12}	n_{13}	m_1
Cluster 2	n_{21}	n_{22}	n_{23}	m_2
Cluster 3	n_{31}	n_{32}	n_{33}	m_3
	c_1	c_2	c_3	n

	Class 1	Class 2	Class 3	
Cluster 1	p_{11}	p_{12}	p_{13}	m_1
Cluster 2	p_{21}	p_{22}	p_{23}	m_2
Cluster 3	p_{31}	p_{32}	p_{33}	m_3
	c_1	c_2	c_3	n

7

External Index

- Entropy:
  - Of a cluster i: $e_i = -\sum_{j=1}^{L} p_{ij} \log p_{ij}$
- Precision:
  - Of a cluster i: $Prec(i,j) = p_{ij}$
    - The fraction of a cluster i that consists of objects of a class j

	Class 1	Class 2	Class 3	
Cluster 1	p_{11}	p_{12}	p_{13}	m_1
Cluster 2	p_{21}	p_{22}	p_{23}	m_2
Cluster 3	p_{31}	p_{32}	p_{33}	m_3
	c_1	c_2	c_3	n

8

External Index

- Recall:
  - Of a cluster i: $Rec(i,j) = \frac{n_{ij}}{c_j}$
    - The extent to which a cluster i contains all objects of a class j.
- F-measure:
  - $F(i,j) = \frac{2*Prec(i,j)*Rec(i,j)}{Prec(i,j)+Rec(i,j)}$

	Class 1	Class 2	Class 3	
Cluster 1	n_{11}	n_{12}	n_{13}	m_1
Cluster 2	n_{21}	n_{22}	n_{23}	m_2
Cluster 3	n_{31}	n_{32}	n_{33}	m_3
	c_1	c_2	c_3	n

9

External Index

	Class 1	Class 2	Class 3	
Cluster 1	2	3	85	90
Cluster 2	90	12	8	110
Cluster 3	8	85	7	100
	100	100	100	300

Precision: (0.94, 0.81, 0.85)
– overall 0.86
Recall: (0.85, 0.9, 0.85)
- overall 0.87

	Class 1	Class 2	Class 3	
Cluster 1	20	35	35	90
Cluster 2	30	42	38	110
Cluster 3	38	35	27	100
	100	100	100	300

Precision: (0.38, 0.38, 0.38)
– overall 0.38
Recall: (0.35, 0.42, 0.38)
– overall 0.39

(Assign to cluster i the class k_i such that $k_i = \arg\max_j n_{ij}$)

10

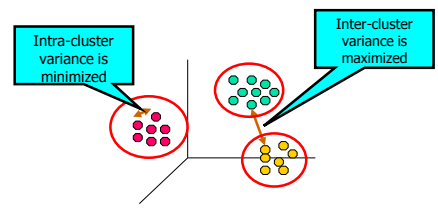
Internal Index

- Used to measure the goodness of a clustering structure without reference to external information.
- Internal index:
  - Variances of within cluster and between clusters
  - Silhouette Coefficient
  - F-Ratio
  - Davies-Bouldin Index (DBI)

11

Internal Index

- Internal validation measures are often based on the following two criteria:
  - **Cluster Cohesion:** Measures how closely related are objects in a cluster.
  - **Cluster Separation:** Measure how distinct or well-separated a cluster is from other clusters.



12

Internal Index

- Cohesion is measured by the within cluster sum of squares (SSE)

$$WSS = \sum_{i} \sum_{x \in C_i} (x - c_i)^2$$

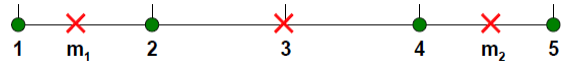
- Separation is measured by the between cluster sum of squares

$$BSS = \sum_{i} m_i (c - c_i)^2$$

13

Internal Index

- Example:
  - BSS + WSS = constant



K=2 clusters:

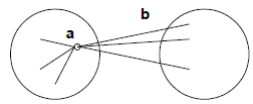
$WSS = (1-1.5)^2 + (2-1.5)^2 + (4-4.5)^2 + (5-4.5)^2 = 1$

$BSS = 2 \times (3-1.5)^2 + 2 \times (4.5-3)^2 = 9$

14

Silhouette Coefficient

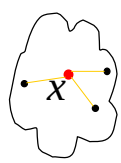
- Silhouette Coefficient combine ideas of both cohesion and separation.
- For an individual point, i
  - Calculate ***a*** = average distance of *i* to the points in its cluster
  - Calculate ***b*** = min (average distance of *i* to points in another cluster)



$$-1 \le s = \frac{b-a}{\max(a,b)} \le 1$$

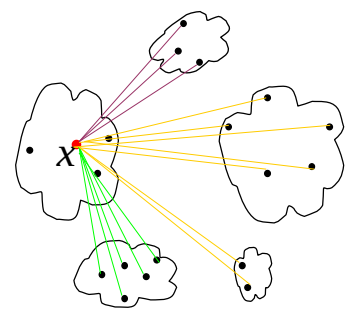
15

Silhouette Coefficient



cohesion

$a(x)$: average distance in the cluster



separation

b(x): average distances to others clusters, find minimal

16

Davies-Bouldin Index

- The Davies-Bouldin index can be calculated by the following formula:

$$DB = \frac{1}{n}\sum_{i=1}^{n}\max_{j\neq i}\left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right)$$

- c_x is the centroid of cluster x
- σx is the average distance of all elements in cluster x to centroid c_x

17

Dunn Index

- The Dunn index can be calculated by the following formula:

$$D = \frac{\min_{1\leq i<j\leq n} d(i,j)}{\max_{1\leq k\leq n} d'(k)}$$

- *d*(*i*,*j*) represents the distance between clusters *i* and *j*.
- *d* '(*k*) measures the intra-cluster distance of cluster *k*

18

F - Ratio

- Measures ratio of between-groups variance against the within-groups variance.

$$F = \frac{k \cdot \sum_{i=1}^{N} \| x_i - c_{p(i)} \|^2}{\sum_{j=1}^{k} n_j \| c_j - \bar{x} \|^2} = \frac{k \cdot SSW}{\underbrace{\sigma(X)}_{\text{BSS}} - SSW}$$

19

Determining Number of Clusters

- By rule of thumb
- Elbow method
- Choosing k using the silhoutte
- Cross validation

20

By Rule of Thumb

- It is a simple method.
- This method can by apply to any type of data set.

$$k = \sqrt{(n/2)}$$

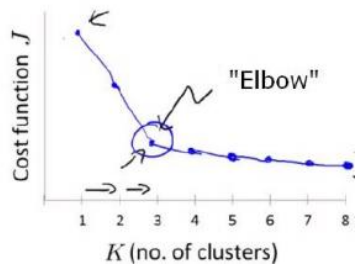
21

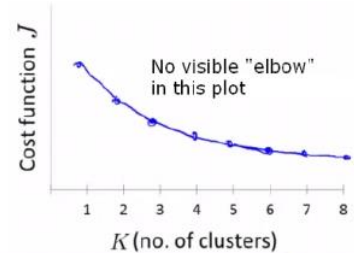
Elbow Method

- The oldest method for determining the true number of clusters in a data set is inelegantly called the elbow method.
- The idea is that Start with K=2, and keep increasing it in each step by 1.
- Calculate your clusters and the cost that comes with the training.
- At some value for K the cost drops dramatically, and after that it reaches a plateau when you increase it further.

22

Elbow Method





23

Choosing k Using The Silhoutte

- The largest average silhouette width, over different K, indicates the best number of clusters.
- The silhouette of a data instance is a measure of how closely it is matched to data within its cluster and how loosely it is matched to data of the neighbouring cluster.

$$S_i = \frac{(b_i - a_i)}{\max(a_i, b_i)} \qquad \overline{S} = \frac{1}{n}\sum_{i=1}^{n} S_i$$

24

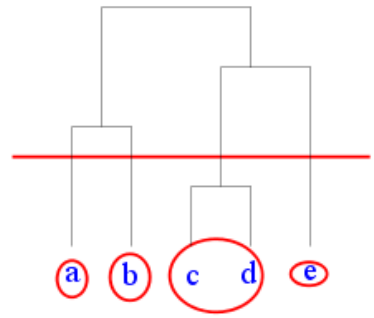
Cross Validation

- This method splits the data in two or more (K) parts.
- One part is used for clustering and the other parts are used for validation.
- The value of the objective function calculated for each part.
- These K values are calculated and averaged for each alternative number of clusters.
- The cluster number is selected that leads to only a small reduction in the objective function.

25

Hierarchical: Dendogram

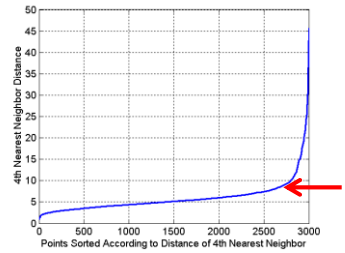
- Cutting a dendrogram at a certain level gives a set of clusters.



26

Denstiy-based: Determining ε

- The average distance to each minpts-nearest neighbors is calculated. These minpts-distances are then drawn in ascending order.



- The ε value can also be determined by the user.

27