Cluster Quality

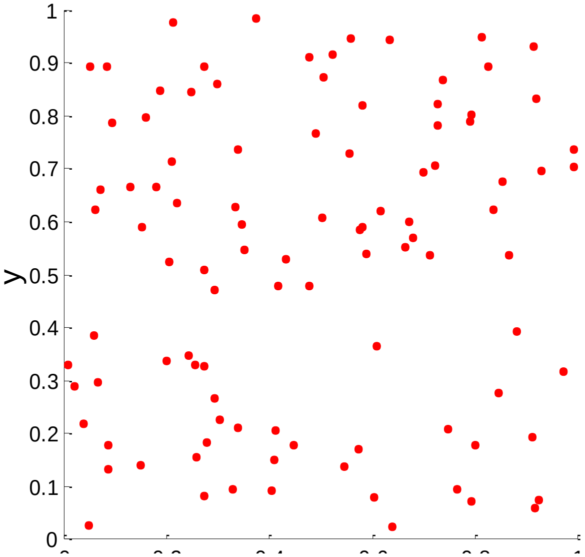
1. Defining Cluster Validity

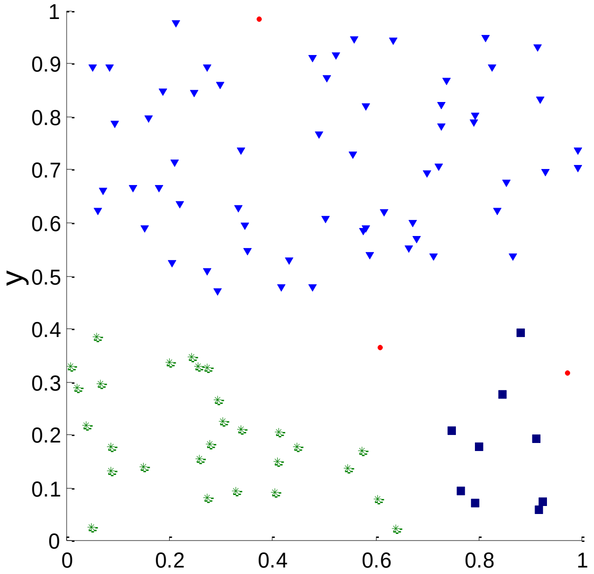
Cluster Validity

- For supervised classification we have a variety of measures to evaluate how good our model is
  - Accuracy, precision, recall
- For cluster analysis, the analogous question is how to evaluate the “goodness” of the resulting clusters?
- But “clusters are in the eye of the beholder”!
- Then why do we want to evaluate them?
  - To avoid finding patterns in noise
  - To compare clustering algorithms
  - To compare two sets of clusters
  - To compare two clusters

Clusters found in Random Data

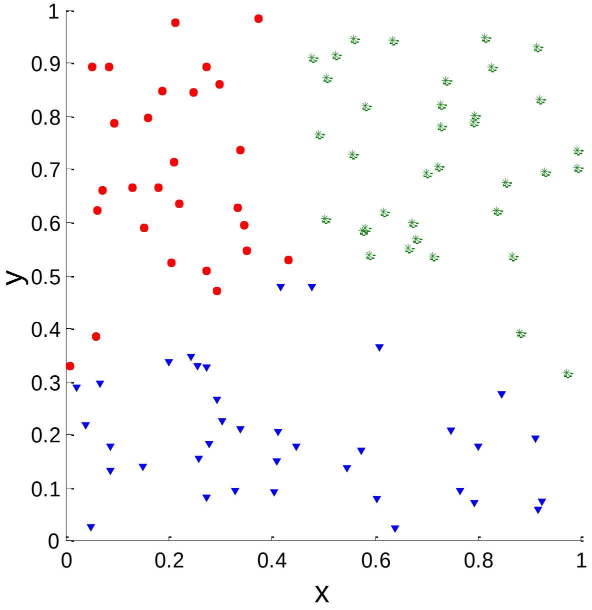
Random Points

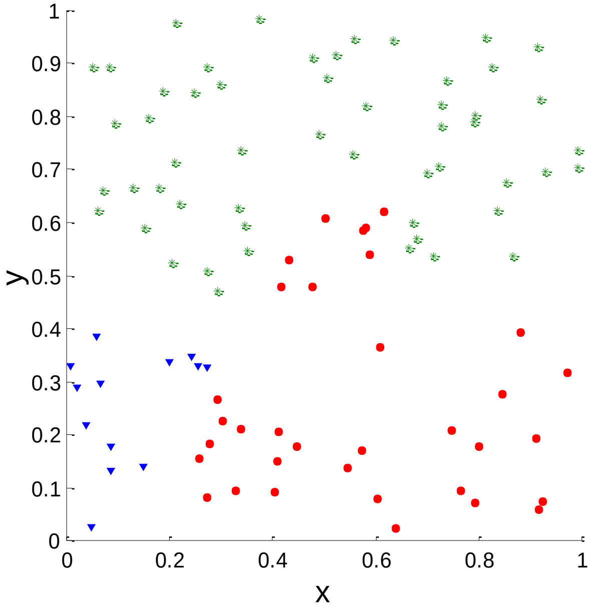




DBSCAN

K-means





Complete Link

Different Aspects of Cluster Validation

1. Determining the clustering tendency of a set of data, i.e., distinguishing whether non-random structure actually exists in the data.
2. Comparing the results of a cluster analysis to externally known results, e.g., to externally given class labels.
3. Evaluating how well the results of a cluster analysis fit the data *without* reference to external information.

   - Use only the data
4. Comparing the results of two different sets of cluster analyses to determine which is better.
5. Determining the 'correct' number of clusters.

   For 2, 3, and 4, we can further distinguish whether we want to evaluate the entire clustering or just individual clusters.

Measuring Clustering Tendency

- Clustering algorithm will find clusters in any data. Is our data random?
  - If random means fitting a given model (i.e. known spatial distribution) need to estimate parameters and evaluate statistical significance
  - If random data means distributed uniformly at random (special, but ubiquitous case of the above) then can use Hopkins statistics.
  - For $D \subset \mathbb{R}^n$ be the set of data points, do the following:

1. Take a sample S, of $|\mathrm{S}| = p$ points in D;
2. Generate a set B $\mathbb{R}^n$ uniformly at random over the range of D, such that $|D| >> |B|$, but $|B| > p$
3. Take a sample S' of points in B that has same size as S
4. For each point x_i in S compute its nearest neighbor distance w_i in D
5. For each point y_i compute its nearest neighbor distance u_i in B
6. Compute tendency $H = \frac{\sum_{i=1}^{p} w_i}{\sum_{i=1}^{p} u_i + \sum_{i=1}^{p} w_i}$

Measures of Cluster Validity

- Numerical measures that are applied to judge various aspects of cluster validity, are classified into the following three types.
  - External Index: Used to measure the extent to which cluster labels match externally supplied class labels.
    - Entropy
  - Internal Index: Used to measure the goodness of a clustering structure *without* respect to external information.
    - Sum of Squared Error (SSE)
  - Relative Index: Used to compare two different clusterings or clusters.
    - Often an external or internal index is used for this function, e.g., SSE or entropy
- Sometimes these are referred to as criteria instead of indices
  - However, sometimes criterion is the general strategy and index is the numerical measure that implements the criterion.

1. Defining Cluster Validity

2. Validity via Matrix Correlation

Cluster Validity Via Correlation

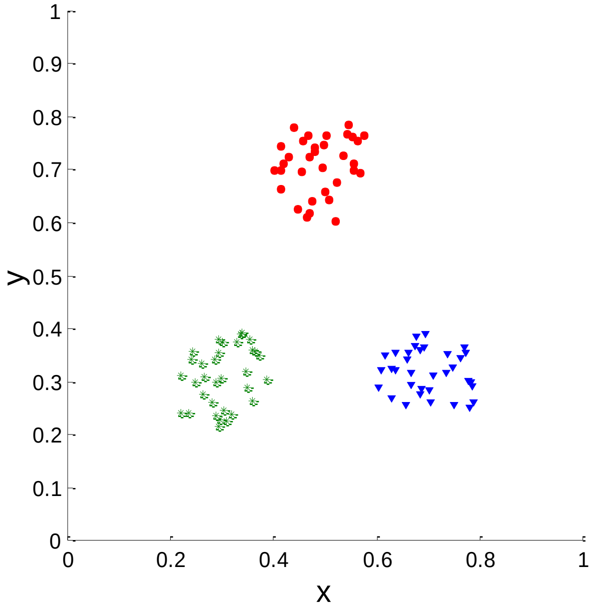
- Two matrices
  - Proximity Matrix
  - “Incidence” Matrix
    - One row and one column for each data point
    - An entry is 1 if the associated pair of points belong to the same cluster (each cluster is a complete graph)
    - An entry is 0 if the associated pair of points belongs to different clusters (each cluster is a separate connected component)
- Measuring correlation between proximity and incidence matrix:
  - If the matrix structure of proximity matrix is unimportant (e.g. spatial data), then treat matrices as vectors (flatten using as.vector in **R**) and compute correlation between vectors
  - If matrix structure is important (e.g. observations of multi-dimensional process in time) then compute canonical correlations

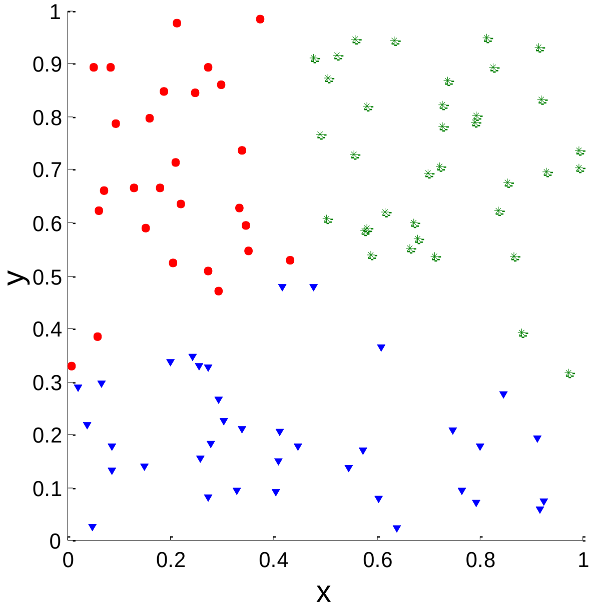
lCluster Validity Via Correlation – cont.

- Compute the correlation between the two matrices
  - Since the matrices are symmetric, only the correlation between
    n(n-1) / 2 entries needs to be calculated.
- High correlation indicates that points that belong to the same cluster are close to each other.
- Not a good measure for some density or contiguity based clusters.

Cluster Validity Via Correlation - continued

- Correlation of incidence and proximity matrices for the K-means clusterings of the following two data sets.





Corr = -0.9235

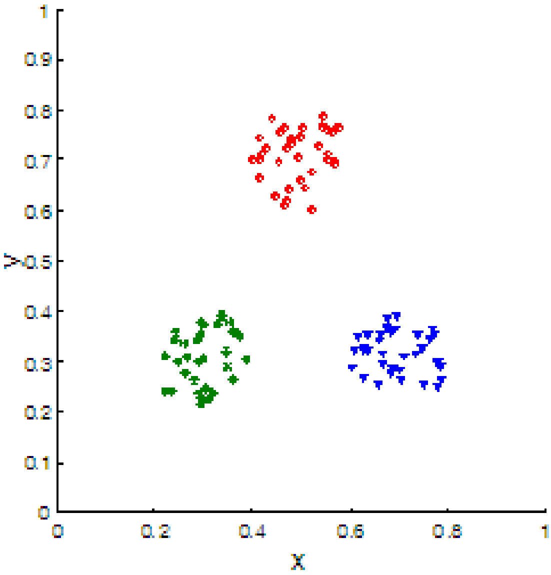
Corr = -0.5810

Framework for Cluster Validity

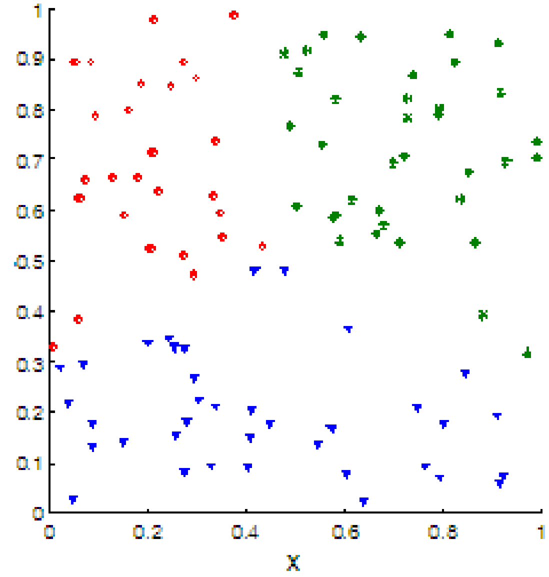
- Need a framework to interpret any measure.
  - For example, if our measure of evaluation has the value, 10, is that good, fair, or poor?
- Statistics provide a framework for cluster validity
  - The more "atypical" a clustering result is, the more likely it represents valid structure in the data
  - Can compare the values of an index that result from random clustering to those of a clustering result.
    - If the value of the index is unlikely, then the cluster results are valid
  - These approaches are more complicated and harder to understand.
- For comparing the results of two different sets of cluster analyses, a framework is less necessary.
  - However, there is the question of whether the difference between two index values is significant

Statistical Framework for Correlation

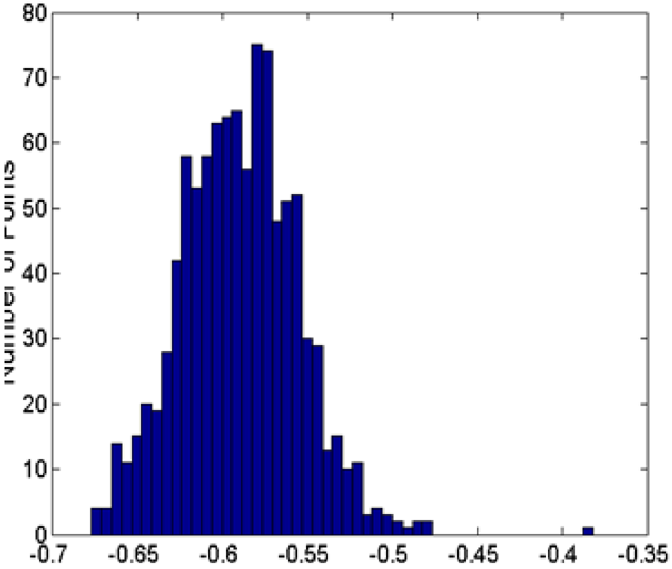
- Correlation of incidence and proximity matrices for the K-means clustering of the following two data sets.
- Histogram: correlation of three clusters vs. 500 sets of random data points size 100 distributed over the range 0 – 1,0 for (x,y) values



Corr = -0.9235



Corr = -0.5810



Correlation of random points

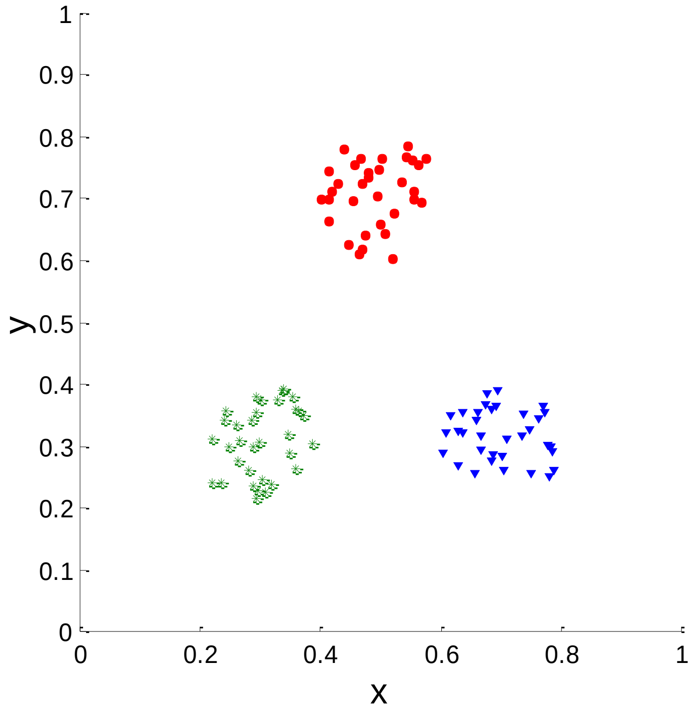
1. Defining Cluster Validity

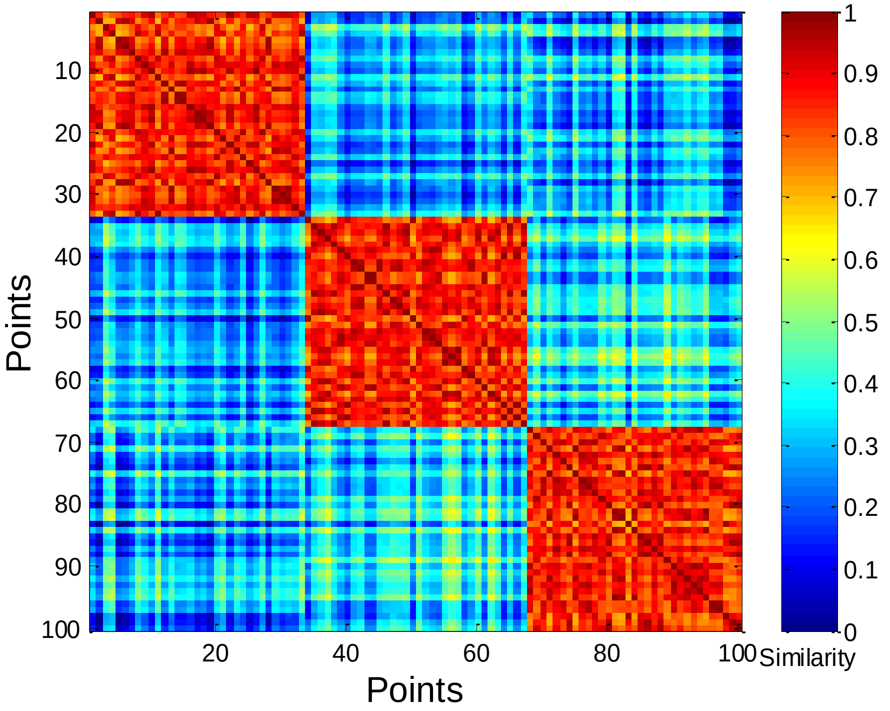
2. Validity via Matrix Correlation

3. Visualization of Validity

Similarity Matrix for Validation/Visualization

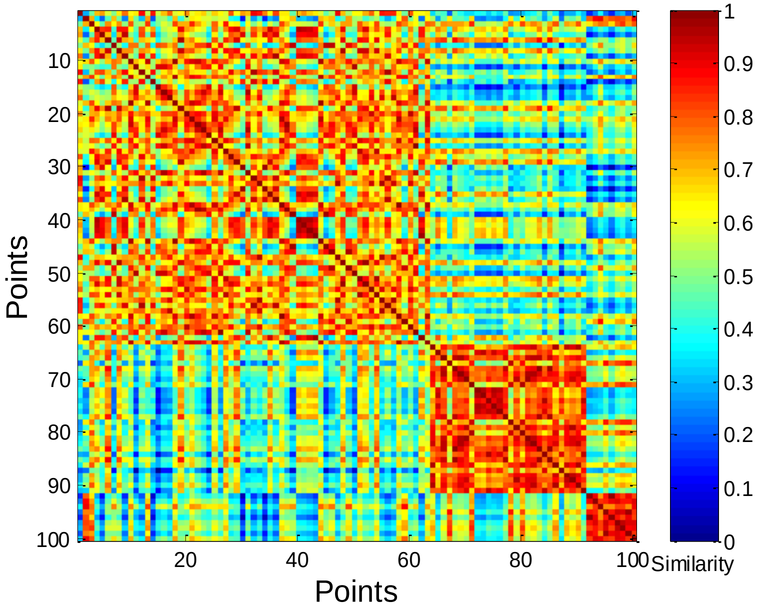
- Order the similarity matrix with respect to (manual) cluster labels and inspect visually.

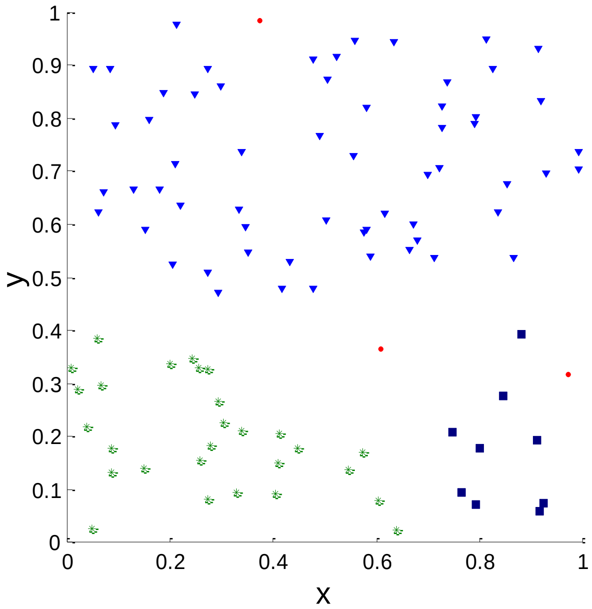




Similarity Matrix for Validation/Visualization

- Clusters in random data are not so crisp
  - Cluster identification using DBSCAN

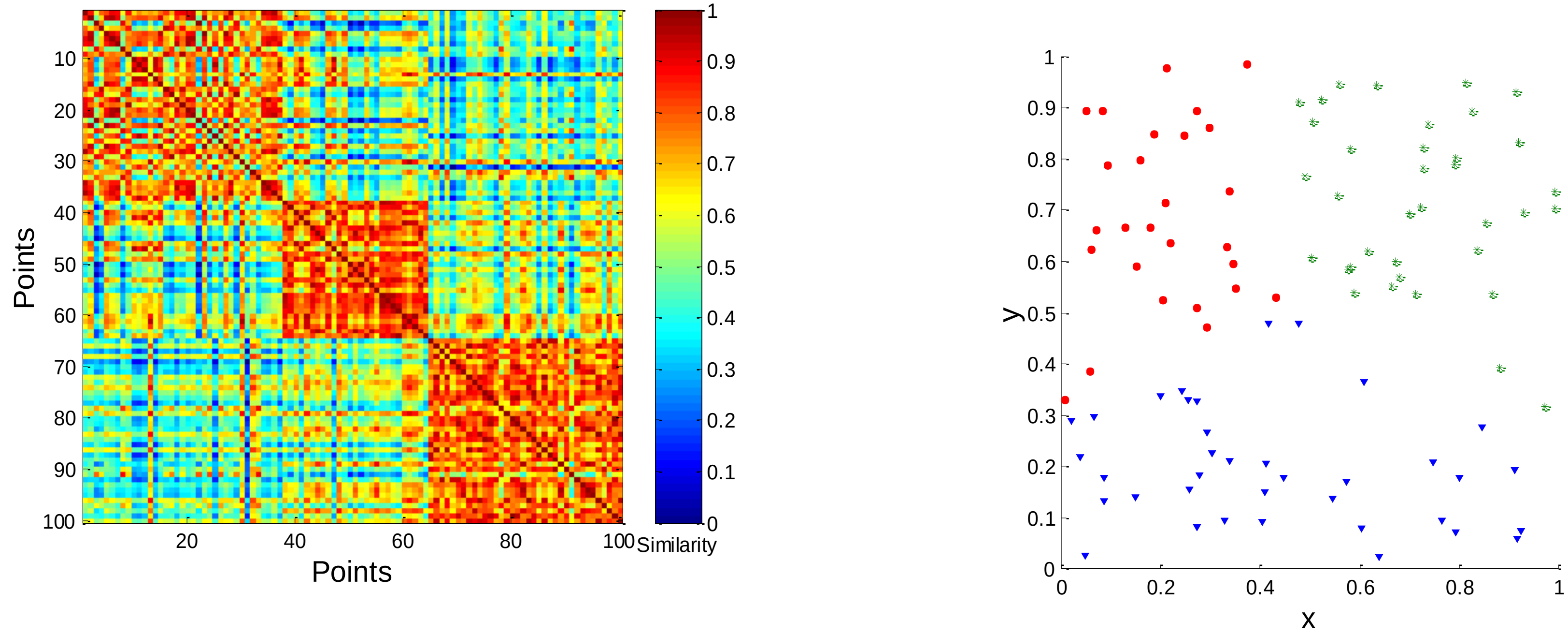




DBSCAN

Similarity Matrix for Validation/Visualization

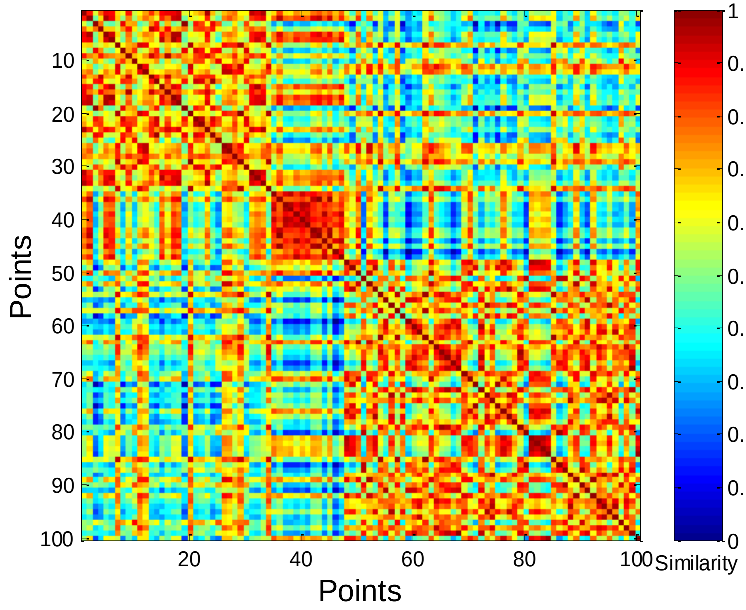
- Clusters in random data are not so crisp
- Cluster identification using 3-means

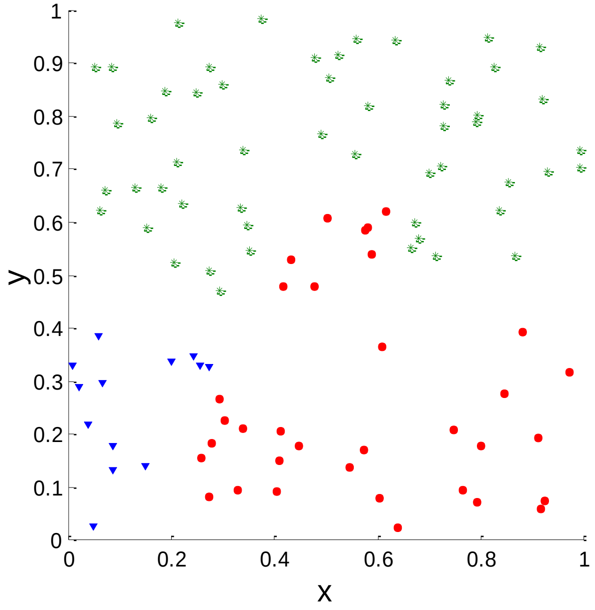


K-means

Similarity Matrix for Validation/Visualization

- Clusters in random data are not so crisp
  - Cluster identification using Complete Graph

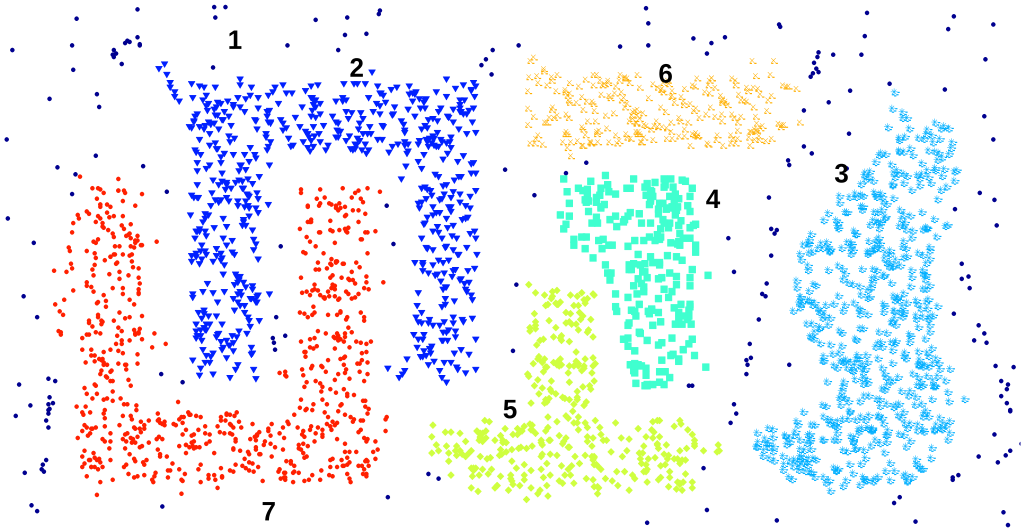


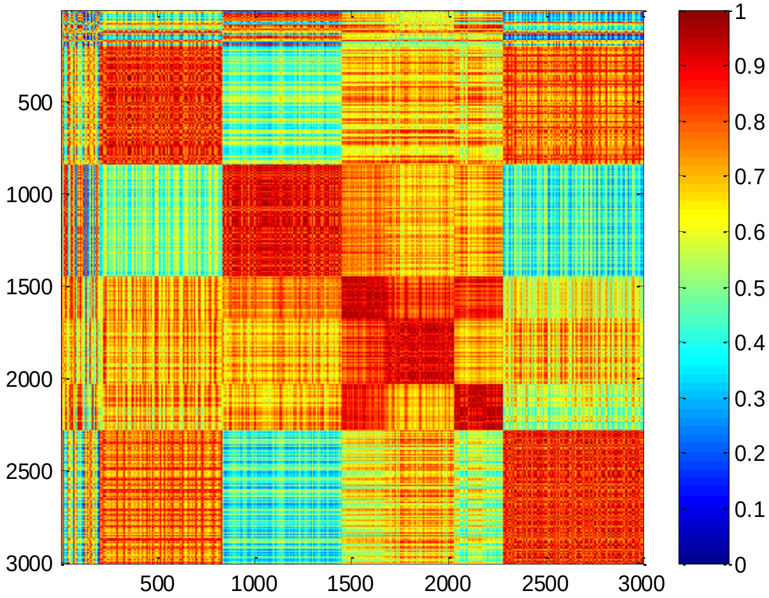


Complete Graph

Similarity Matrix for Validation/Visualization

- Clusters in meaningful data test
  - Cluster identification using DBSCAN





DBSCAN

1. Defining Cluster Validity

2. Validity via Matrix Correlation

3. Visualization of Validity

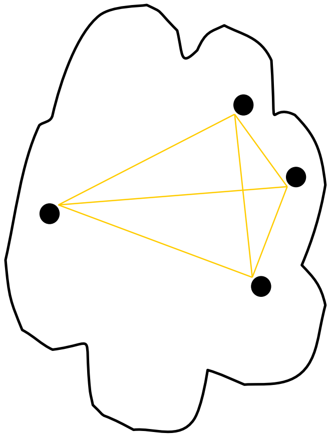
4. Internal Measures

Internal Measures: Cohesion and Separation

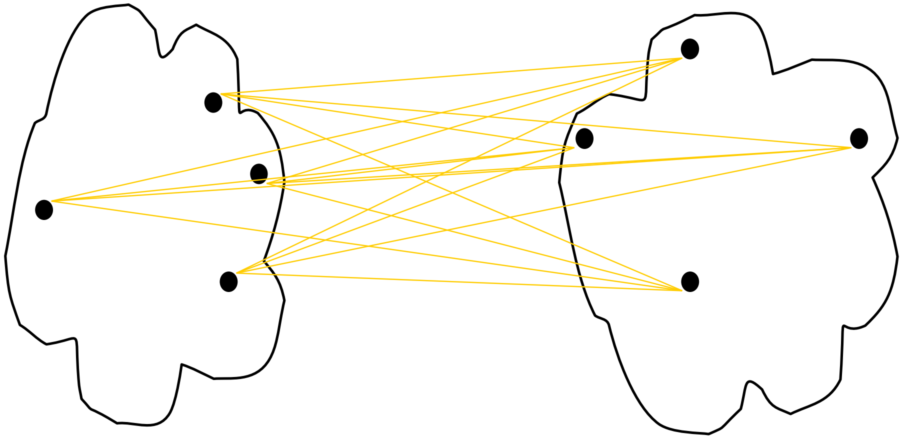
- Cluster Cohesion: Measures how closely related are objects in a cluster
- Cluster Separation: Measures how distinct or well-separated a cluster is from other clusters
- Two classes of measures – prototype based and graph based

Internal Measures: Graph Cohesion/Separation

- Let D be data, $G_D$proximity graph
- Cluster Ci = subgraph induced by data points in C_i
- Graph-based cohesion:
  - Cohesion of a cluster C_i
    $Coh_G(C_i) = w_i \sum_{x, y \in Ci} prox(x, y)$
    or total weighted edge capacity (distance) of a subgraph C_i
  - Total cohesion $Coh_G = \sum_i Coh_G(C_i)$
- Graph based separation:
  - $Sep_G(C_i, Cj) =$
    $w_{ij} \sum_{x \in C_i, y \in C_j} prox(x, y)$
    or the weighted capacity (distance) of the cut separating C_{i} and C_j
  - Total separation $Sep_G =$
    $\sum_i \sum_j Sep_G(C_i, C_j\ )$



cohesion



separation

Internal Measures: Prototype Cohesion/Separation

- Prototypes = centroids or medoids. Cohesion (with prototypes): $Coh_p(Ci) = \sum_{y\ \in Ci} w_i\, prox(c_i, y)$ where c_i –centroid/medoid of cluster C_i and w_i – weight assigned to a cluster

  Example: Within-cluster Squared Error: Cohesion is measured by the within cluster sum of squares (SS) where

  $prox(x, y) = \|x - y\|^2$, i.e.

$$Coh_{WSE}(C_i) = \sum_{x \in C_i} \|x - c_i\|^2$$

  Then total cohesion is $Coh_{WSE} = \sum_i Coh\_WSE(C_i)$

- Separation (with prototypes) $Sep_p(C_i) = \sum_i^k w_i prox(c_i\ , c)$ where c is overall centroid

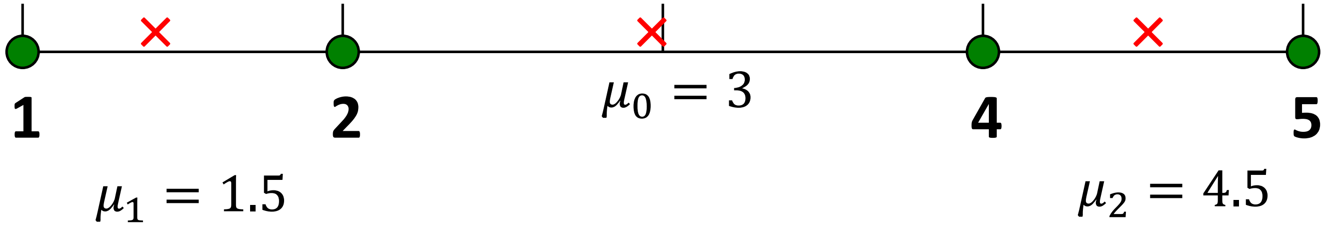
  Example: Total Separation by Between-cluster Sum of Squares. Weight in cluster separation is its size: $Sep_{BSS}(C_i) = |C_i|\|c_i - c\|$, where $|Ci|$ is the size of cluster i.

  So total separation is $Sep_{BSS} = \sum_i Sep_{BSS}(C_i)$

Internal Measures: Cohesion and Separation

Example: $prox = SS$

- Total sum of squares TSS=BSS + WSE =constant



- K=1 - one cluster:
  - $Coh_{WSE} = (1-3)^2 + (2-3)^2 + (4-2)^2 + (5-3)^2 = 10$
  - $Sep_{BSS} = 4 \times (3-3) = 0$
  - $TSS = 10 + 0$
- K=2 –two clusters:
  - $Coh_{WSE} = (1-1.5)^2 + (2-1.5)^2 + (4-4.5)^2 + (5-4.5)^2 = 1$
  - $Sep_{BSS} = 2 \times (3-1.5)^2 + 2 \times (4.5-3)^2 = 9$
  - $TSS = 1 + 9 = 10$

SS ⟹ Graph Cohesion = Prototype Cohesion

- Let $prox$ function be SS and weight of each pair of points in a cluster be $\frac{1}{2|C_i|}$

$$\textbf{Coh}\ (C_i) = \frac{1}{2|C_i|}\sum_{x,y\in C_i}(x-y)^2 = \frac{1}{2|C_i|}\sum_{x,y\in C_i}((x-c_i)-(y-c_i))^2$$

$$= \frac{1}{2|C_i|}\left(\sum_{x,y\in C_i}(x-c_i)^2 - 2\sum_{x,y\in C_i}(x-c_i)(y-c_i) + \sum_{x,y\in C_i}(y-c_i)^2\right)$$

$$= \frac{1}{2|C_i|}\left(|C_i|\sum_{x\in C_i}(x-c_i)^2 + |C_i|\sum_{y\in C_i}(y-c_i)^2\right) = \sum_{x\in C_i}(x-c_i)^2$$

$$= SSE(C_i)$$

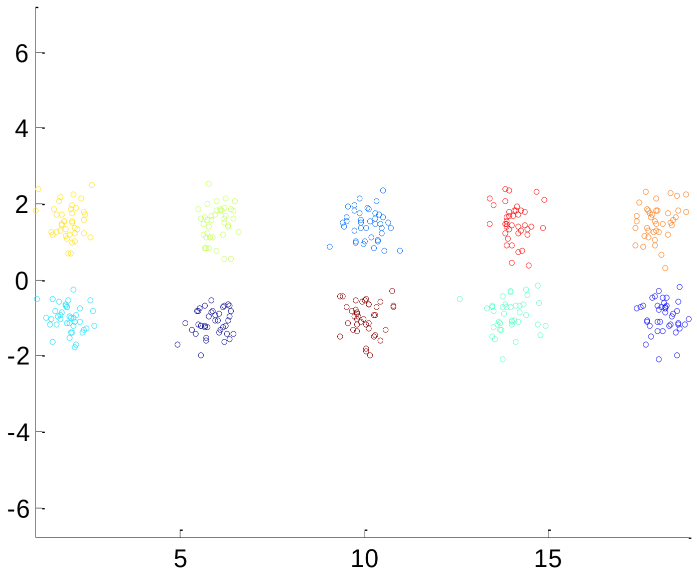
where

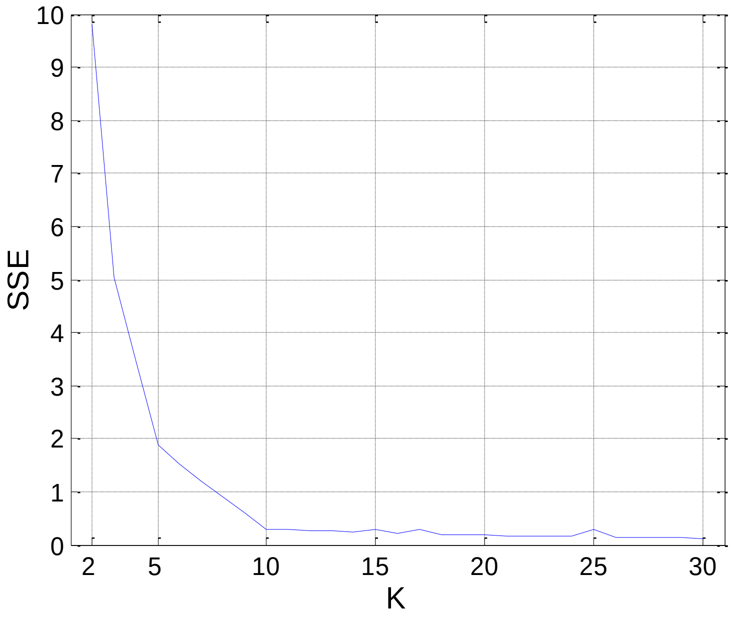
$$\sum_{x,y\in C_i}(x-c_i)(y-c_i) = \sum_{x\in C_i}x\sum_{y\in C_i}y - \sum_{x\in C_i}x\sum_{y\in C_i}c_i - \sum_{x\in C_i}c_i\sum_{y\in C_i}y + \sum_{x\in C_i}c_i\sum_{y\in C_i}c_i$$

$$=|C_i|c_i|C_i|c_i - 2|C_i|c_i\sum_{y\in C_i}c_i + \sum_{x\in C_i}c_i\sum_{y\in C_i}c_i = 0$$

Internal Measures: SSE

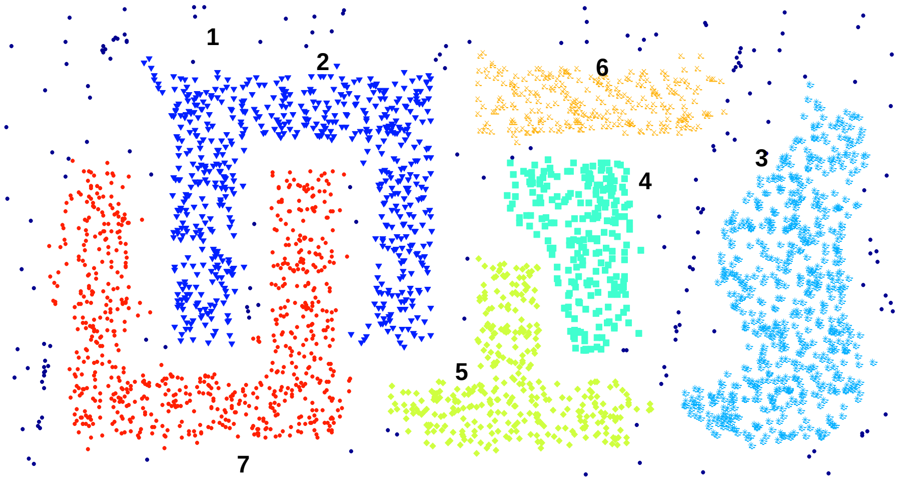
- Clusters in more complicated figures aren’t well separated
- Internal Index: Used to measure the goodness of a clustering structure without respect to external information
  - SSE
- SSE is good for comparing two clusterings or two clusters (average SSE).
- Can also be used to estimate the number of clusters

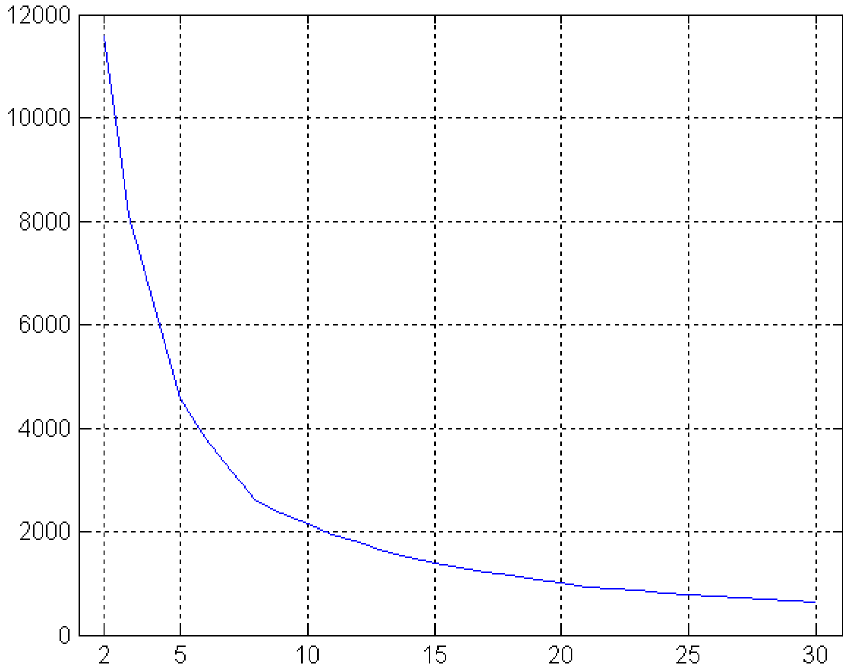




Internal Measures: SSE

- SSE curve for a more complicated data set

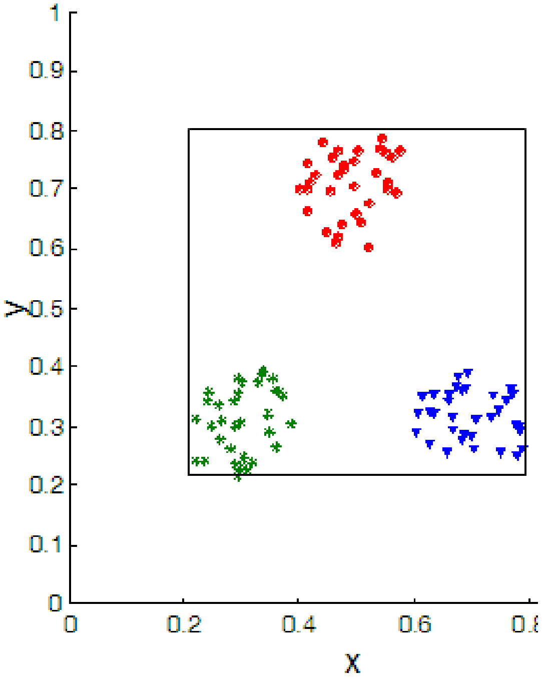




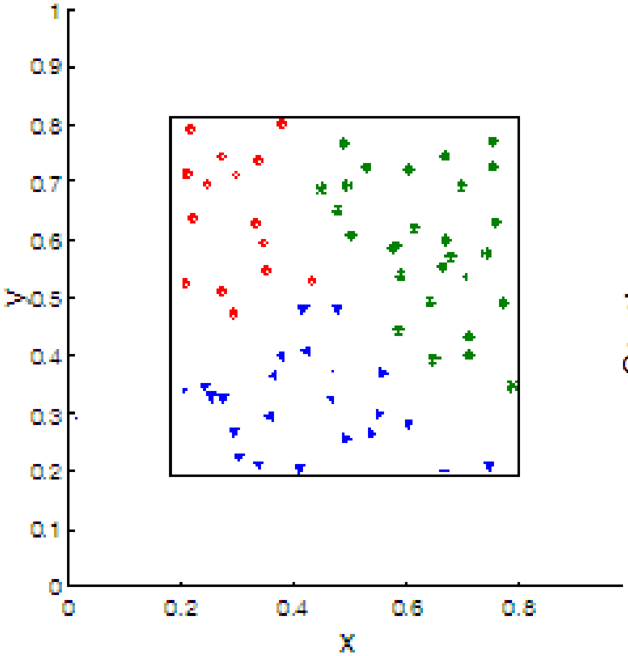
SSE of clusters found using K-means

Statistical Framework for SSE

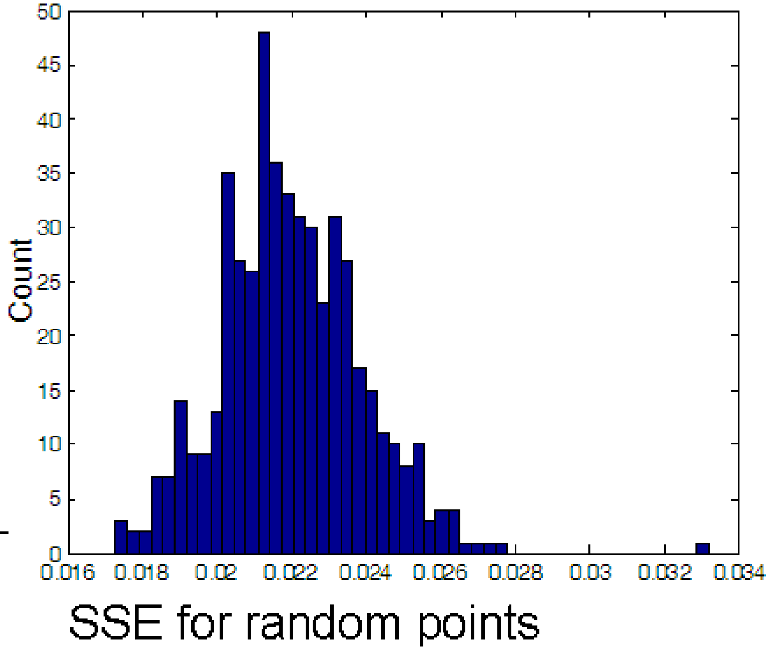
- Example
  - Compare SSE of 0.005 of three clusters against random data
  - Histogram: SSE of three clusters vs. 500 sets of random data points size 100 distributed over the range 0.2 – 0.8 for (x,y) values



SSE=0,005



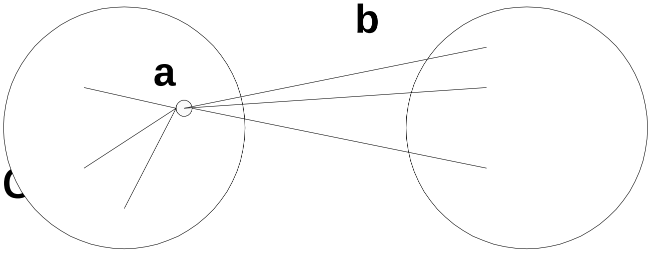
SSE=0.024



Silhouette Coefficient

- Silhouette Coefficient combine ideas of both cohesion and separation, but for individual points, as well as clusters and clusterings. For individual point x:
  - *Cohesion a*(*x*): average distance of *x* to all other vectors in the same cluster.
  - *Separation b*(*x*): average distance of *x* to the vectors in other clusters. Find the minimum among the clusters.
  - *silhouette*

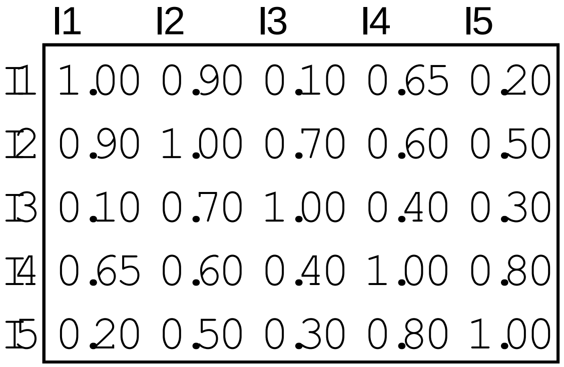
$$s(x) = \frac{b(x) - a(x)}{\max\{a(x), b(x)\}}$$



  - $s(x) = [-1, +1]$: -1=bad, 0=indifferent, 1=go
  - Silhouette coefficient (SC):

Type equation here.

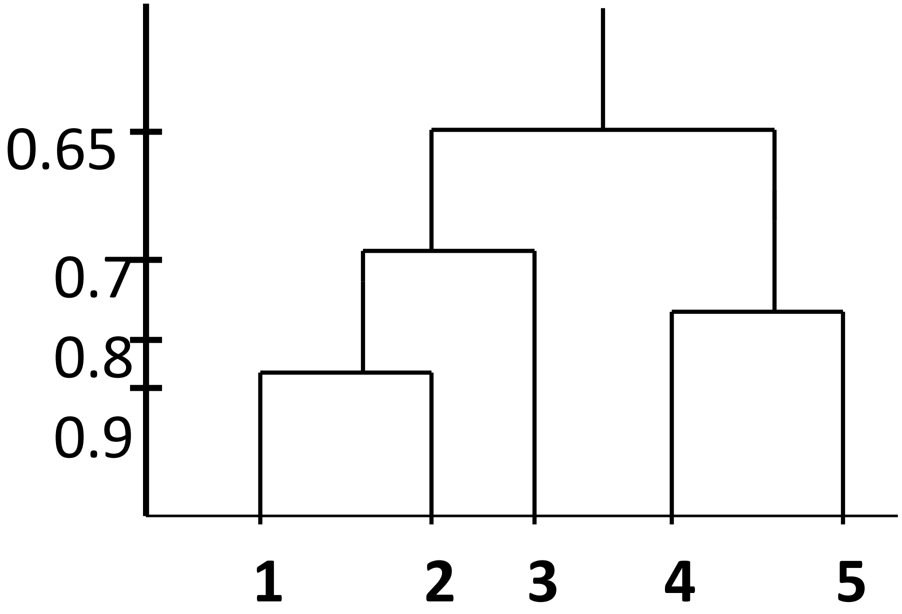
Cophenetic Matrix – Example: Single Link



	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00

	p1	p2	p3	p4	p5
p1		0.9	0.7	0.65	0.65
p2			0.7	0.65	0.65
p3				0.65	0.65
p4					0.8
p5					

Cophenetic Matrix



Cophenetic correlation coefficient – computed between cophenetic matrix and original similarity matrix

1. Defining Cluster Validity

2. Cluster Validity via Matrix Correlation

3. Visualization of Validity

4. Internal Measures

5. External Measures

External Measures of Cluster Validity: Entropy and Purity

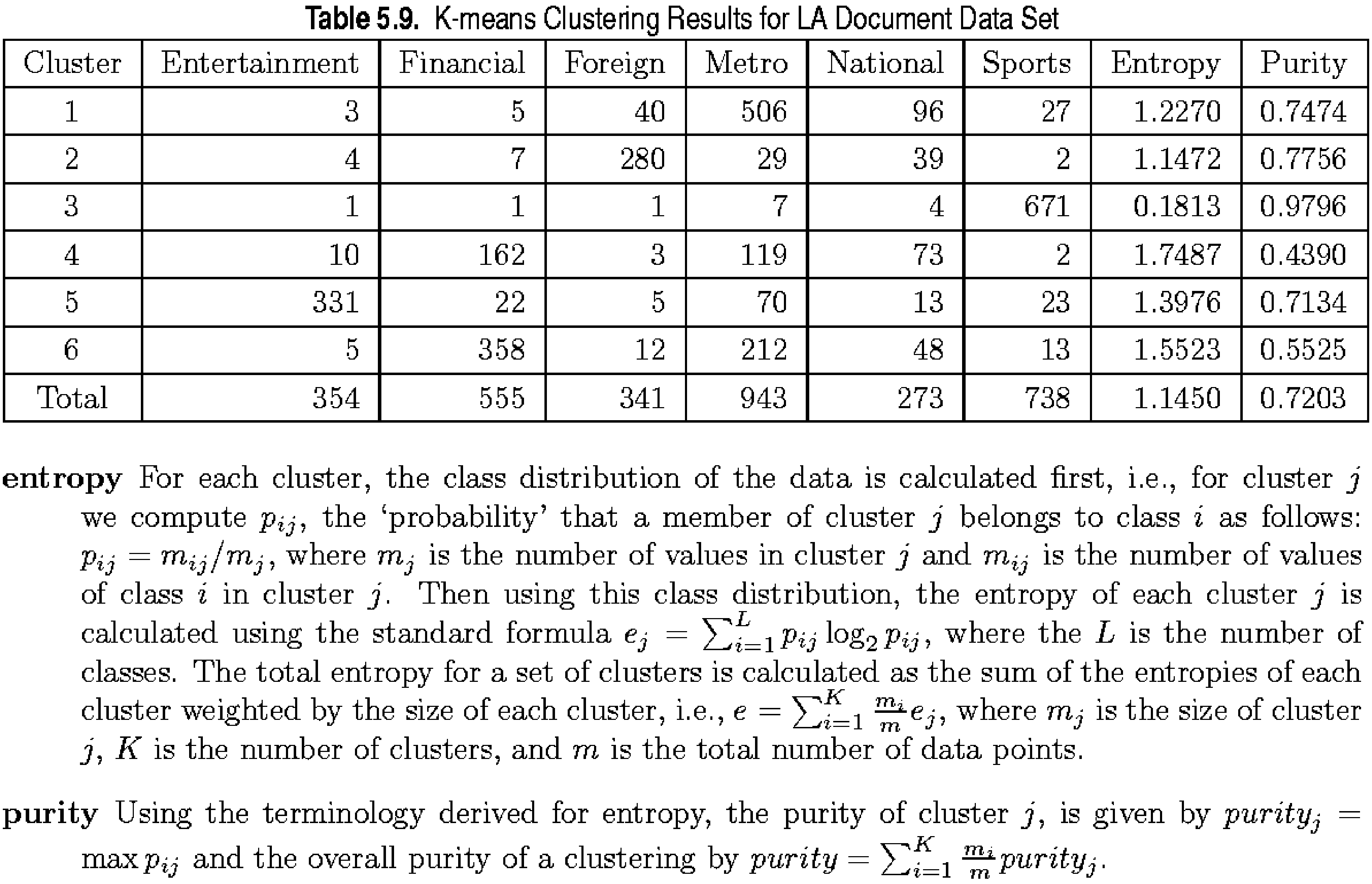


Table 5.9. K-means Clustering Results for LA Document Data Set

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

entropy For each cluster, the class distribution of the data is calculated first, i.e., for cluster j we compute p_{ij}, the 'probability' that a member of cluster j belongs to class i as follows: $p_{ij} = m_{ij}/m_j$, where m_j is the number of values in cluster j and m_{ij} is the number of values of class i in cluster j. Then using this class distribution, the entropy of each cluster j is calculated using the standard formula $e_j = \sum_{i=1}^{L} p_{ij} \log_2 p_{ij}$, where the L is the number of classes. The total entropy for a set of clusters is calculated as the sum of the entropies of each cluster weighted by the size of each cluster, i.e., $e = \sum_{i=1}^{K} \frac{m_i}{m} e_j$, where m_j is the size of cluster j, K is the number of clusters, and m is the total number of data points.

purity Using the terminology derived for entropy, the purity of cluster j, is given by $purity_j = \max p_{ij}$ and the overall purity of a clustering by $purity = \sum_{i=1}^{K} \frac{m_i}{m} purity_j$.

More External Measures of Cluster Validity

- Similarity-Oriented
  - From cluster labels compute ideal cluster matrix (i.e. block matrix, $C(x, y) = 1$ if x, y belongs to same cluster and $C(x, y) = 0$ otherwise)
  - From class labels compute class matrix (same computation)
  - Rand statistic= $(f_{00} + f_{11})/(f_{00} + f_{01} + f_{10} + f_{11})$
  - Jaccard coefficient $= f_{11}/(f_{00} + f_{01} + f_{10} + f_{11})$

Cluster Validity for Hierarchical Clustering

- The idea: for each class there must be at least one cluster that is good w.r.t. a chosen measure. So we take a cluster that is best w.r.t. this measure and then combine these using weighted average of all per-class measures
  - if chosen measure is purity then

    $\sum_{j=1}^{k}\left(\frac{m_j}{m}\right)\max_{i} p_{ij}$ where p_{ij} frequency of class i in cluster j
  - If chosen measure is F-measure then

    $\sum_{j=1}^{k}\left(\frac{m_j}{m}\right)\max_{i} F_{ij}$ where
    - For each class j maximum is taken over all clusters;
    - $F_{ij} = \frac{2p_{ij}r_{ij}}{p_{ij}+r_{ij}}$ measures the extent to which cluster i contains only class j
    - For each class j and cluster i *recall* r_{ij} is fraction of class j contained in cluster i, i.e. $r_{ij} = \frac{m_{ij}}{m_i}$;

Final Comment on Cluster Validity

“The validation of clustering structures is the most difficult and frustrating part of cluster analysis.

Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage.”

Algorithms for Clustering Data, Jain and Dubes

Reading

- 8.5