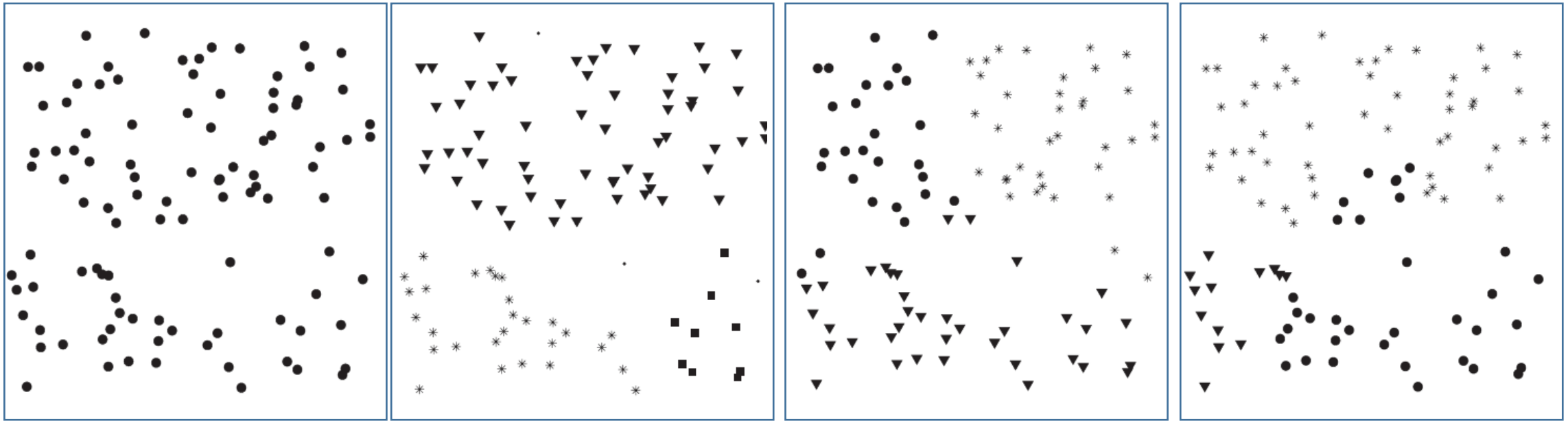

DOGBERT CONSULTS
CUSTOMER DATA IS AN ASSET THAT YOU CAN SELL.
IT'S TOTALLY ETHICAL BECAUSE OUR CUSTOMERS WOULD DO THE SAME THING TO US IF THEY COULD.
SOUNDS FAIR.
IN PHASE ONE, WE'LL DEHUMANIZE THE ENEMY BY CALLING THEM "DATA."
Dilbert.com DilbertCartoonist@gmail.com
10-13-10 ©2010 Scott Adams, Inc./Dist. by UFS, Inc.

Cluster Validation

Clustering Problem

- Almost every clustering algorithm will find clusters in a data set, even if the data has no natural cluster structure



(a) Original points. (b) Three clusters found by DBSCAN. (c) Three clusters found by K-means. (d) Three clusters found by complete link.

Figure 8.26. Clustering of 100 uniformly distributed points.

Clustering Tendency

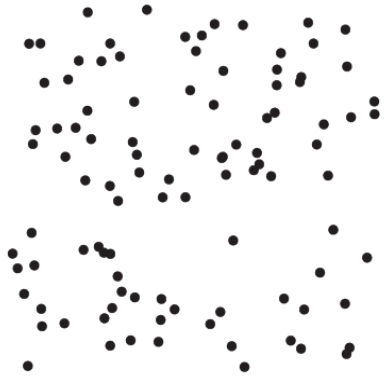
- Evaluate whether a data set has clusters, without clustering
- Statistical tests for spatial randomness (mostly for data in low-dimensional Euclidean space)

Hopkins Statistic

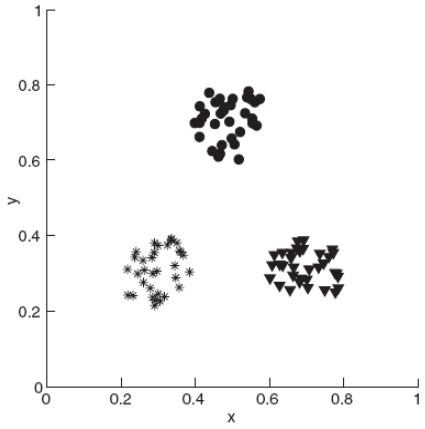
- Generate m points that are randomly distributed across the data space
- Also sample m actual data points from the data set
- For both sets of points, find the distance to the nearest neighbor in the original data set

$$H = \frac{\sum_{i=1}^{m} u_i}{\sum_{i=1}^{m} u_i + \sum_{i=1}^{m} w_i}$$

u_i = the nearest neighbor distances of the artificially generated points
w_i = the nearest neighbor distances of the sample points from the original data



m=20, 100 different trials
Average H=0.56, σ=0.03



m=20, 100 different trials
Average H=0.95, σ=0.006

Similarity Matrix

	P1	P2	P3	P4	P5	P6	P7	P8	P9
P1	0	4	5	1	2	3	4	5	2
P2	4	0	1	4	2	7	8	8	5
P3	5	1	0	3	1	5	6	7	5
P4	1	4	3	0	3	5	6	7	4
P5	2	2	1	3	0	5	6	7	4
P6	3	7	5	5	5	0	1	2	3
P7	4	8	6	6	6	1	0	1	3
P8	5	8	7	7	7	2	1	0	4
P9	2	5	5	4	4	3	3	4	0

Fill matrix with a measure of similarity, like Euclidian distance

Similarity Matrix

	P1	P2	P3	P4	P5	P6	P7	P8	P9
P1		4	5	1	3	3	4	5	2
P2	4		1	4	2	7	8	8	5
P3	5	1		3	1	5	6	7	5
P4	1	4	3		3	5	6	7	1
P5	3	2	1	3		5	6	7	4
P6	3	7	5	5	5		1	2	3
P7	4	8	6	6	6	1		1	3
P8	5	8	7	7	7	2	1		4
P9	2	5	5	1	4	3	3	4	

Color matrix according to values

Similarity Matrix

	P2	P3	P5	P1	P4	P9	P6	P7	P8
P2		1	2	4	4	5	7	8	8
P3	1		1	5	3	5	5	6	7
P5	2	1		3	3	4	5	6	7
P1	4	5	3		1	2	3	4	5
P4	4	3	3	1		1	5	6	7
P9	5	5	4	2	1		3	3	4
P6	7	5	5	3	5	3		1	2
P7	8	6	6	4	6	3	1		1
P8	8	7	7	5	7	4	2	1	

Cluster the data and
sort points according to cluster assignment

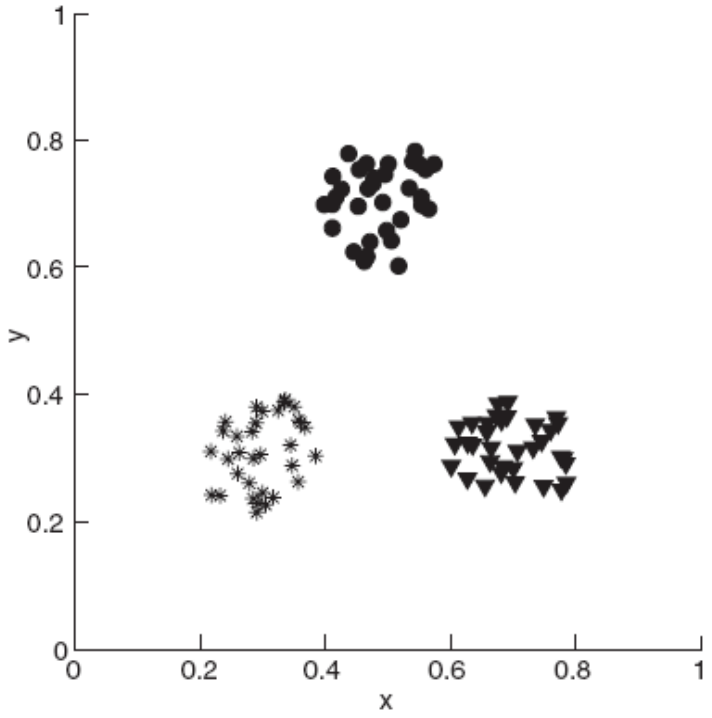
C1 = {P2,P3,P5}
C2 = {P1,P4,P9}
C3 = {P6,P7,P8}

Cohesive and well-separated clusters will have a strong block-diagonal pattern

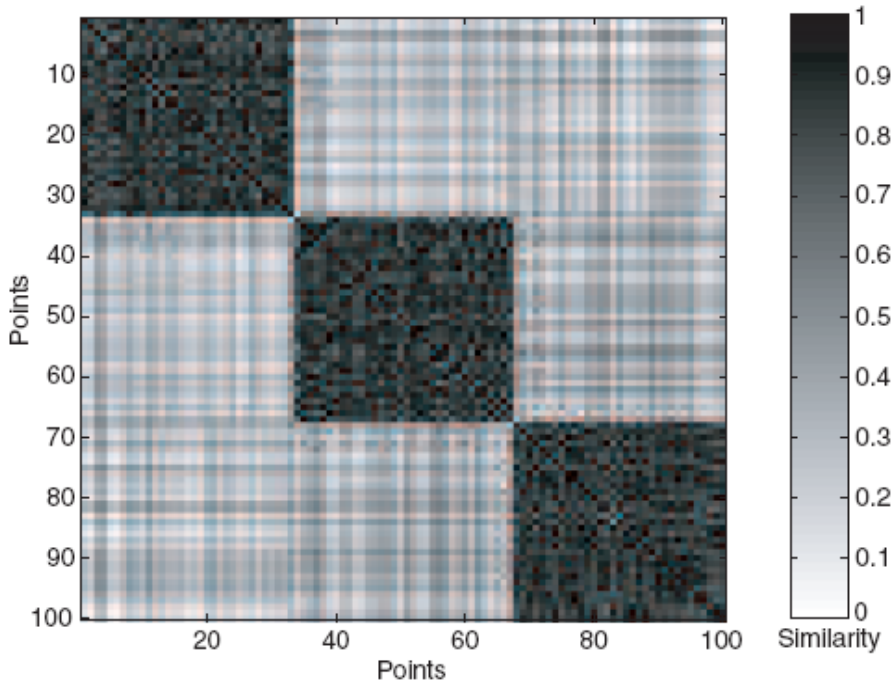
Similarity Matrix

- Transform distance to similarity

$s = 1 - (d - min\_d) / (max\_d - min\_d)$

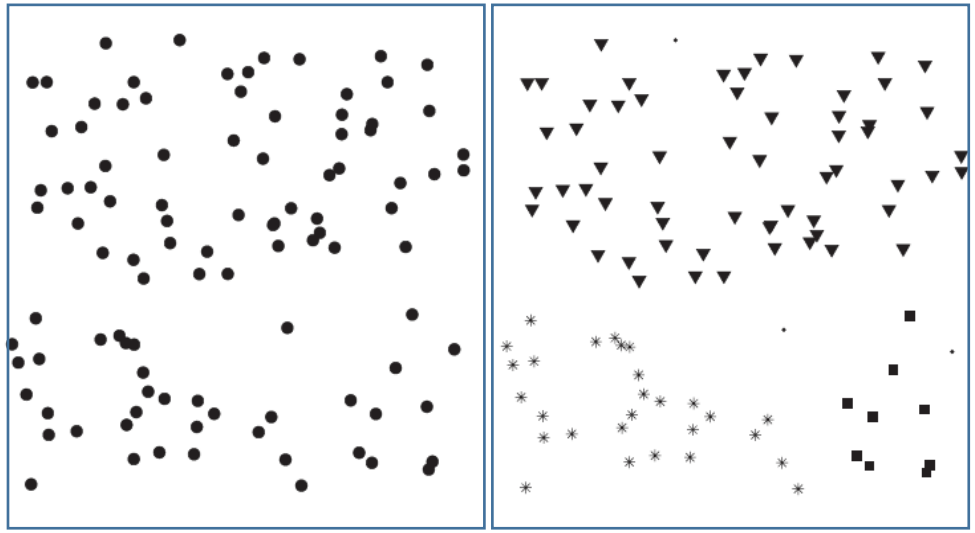


(a) Well-separated clusters.



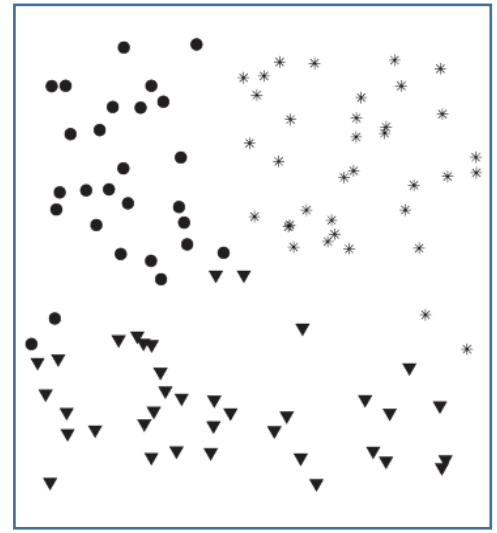
(b) Similarity matrix sorted by K-means cluster labels.

Figure 8.30. Similarity matrix for well-separated clusters.

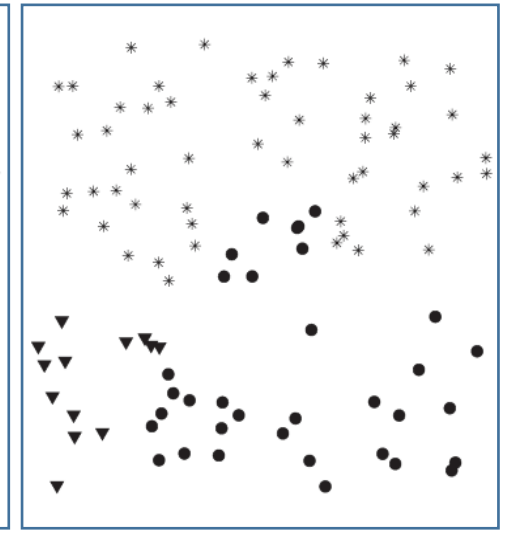


(a) Original points.

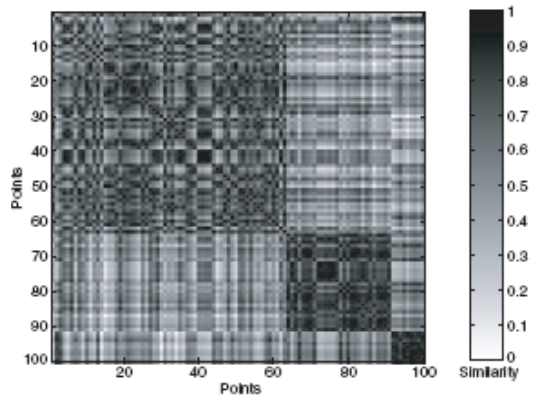
(b) Three clusters found by DBSCAN.



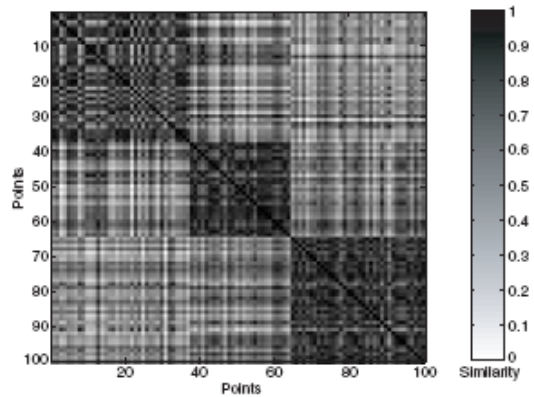
(c) Three clusters found by K-means.



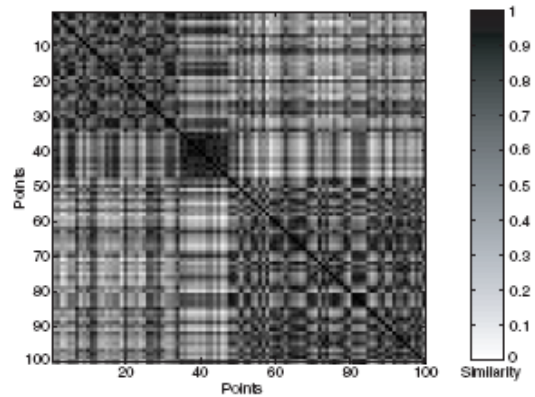
(d) Three clusters found by complete link.



(a) Similarity matrix sorted by DBSCAN cluster labels.



(b) Similarity matrix sorted by K-means cluster labels.



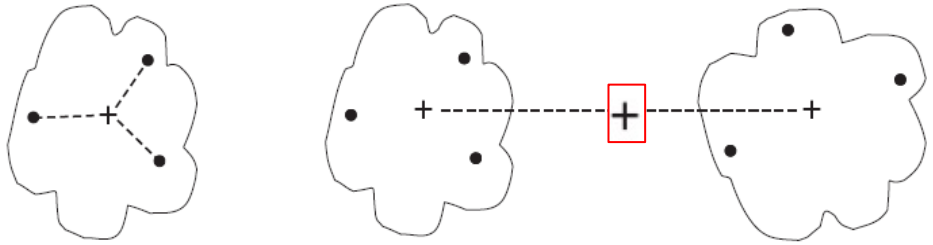
(c) Similarity matrix sorted by complete link cluster labels.

Figure 8.31. Similarity matrices for clusters from random data.

Internal Measures

- Unsupervised Validation
- Measures the goodness of a clustering using only information present in the dataset

Cohesion and Separation



(a) Cohesion. (b) Separation.

- **Cluster Cohesion:** How closely related are objects in a cluster
  - Example: Within cluster sum of squares (WSS=SSE)

$$\boldsymbol{WSS = SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist(c_i, x)^2}$$

- **Cluster Separation:** How distinct is a cluster from other clusters
  - Example: Between cluster sum of squares (BSS/SSB)

$$\boldsymbol{BSS = SSB = \sum_{i=1}^{K} |C_i| \, dist(c_i, c)^2}$$

Where K is the number of clusters
C_i is the i^{th} cluster
$|C_i|$ is the size of the i^{th} cluster
c_i is the centroid of cluster C_i
c is the overall centroid of all the data
x is a data point

Total Sum of Squares

- **Total Sum of Squares:**
  - TSS = WSS + BSS

  - TSS is a constant: The sum of squares of the distance of each point to the overall centroid of all the data

$$\boldsymbol{TSS} = \sum_{\boldsymbol{x}} \boldsymbol{dist}(\boldsymbol{c}, \boldsymbol{x})^{\boldsymbol{2}}$$

Silhouette Coefficient

- Combines both cohesion and separation
- For each point:
  - Calculate its average distance to all other objects in its cluster *(call this a)*
  - Calculate its average distance to all other objects in a different cluster. Do this for each different cluster. Find the minimum value of these *(call this b)*
  - The silhouette coefficient for the data point is

$$s = (b-a) / max(a,b)$$

- The value of the silhouette coefficient will be between -1 to 1

- Can compute the average silhouette coefficient of an individual cluster by averaging across all data points in the cluster
- Can compute the average silhouette coefficient of the total clustering by averaging across all data points in all clusters

Silhouette Coefficient

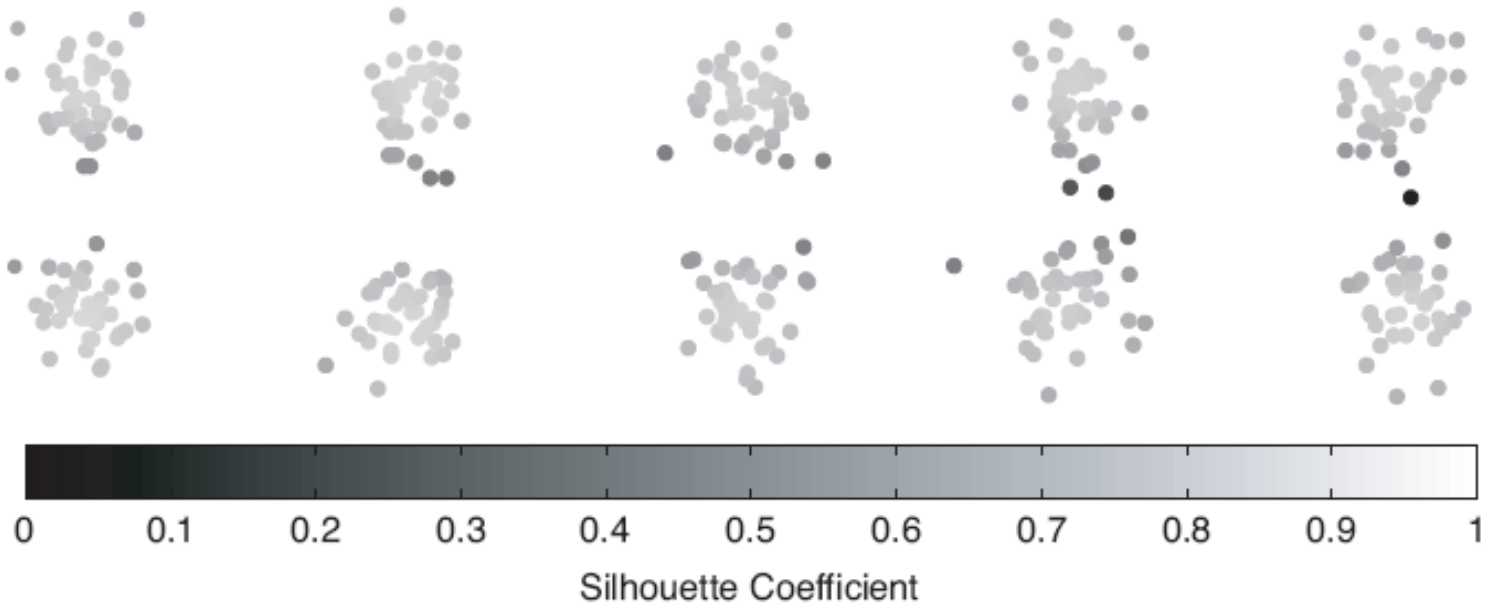


Figure 8.29. Silhouette coefficients for points in ten clusters.

Determining the Correct Number of Clusters

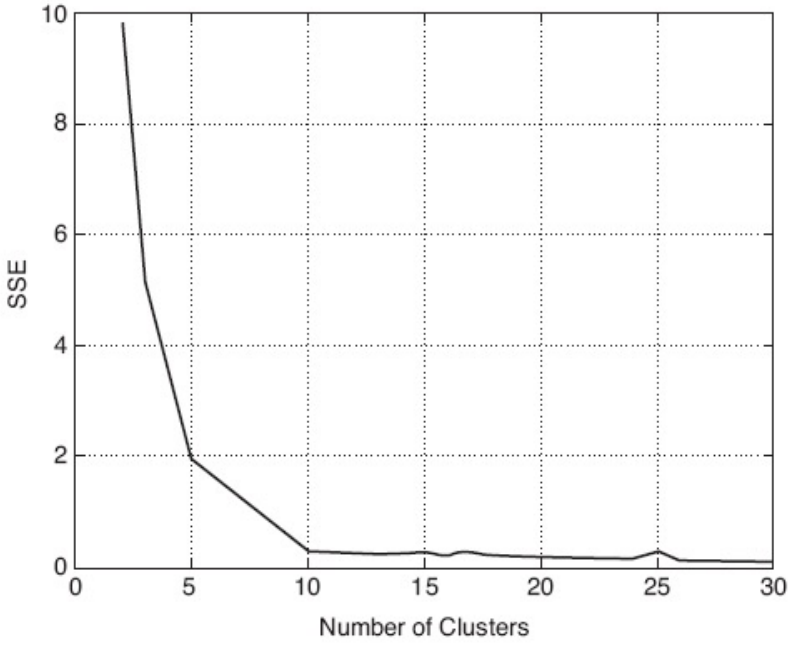


Figure 8.32. SSE versus number of clusters for the data of Figure 8.29.

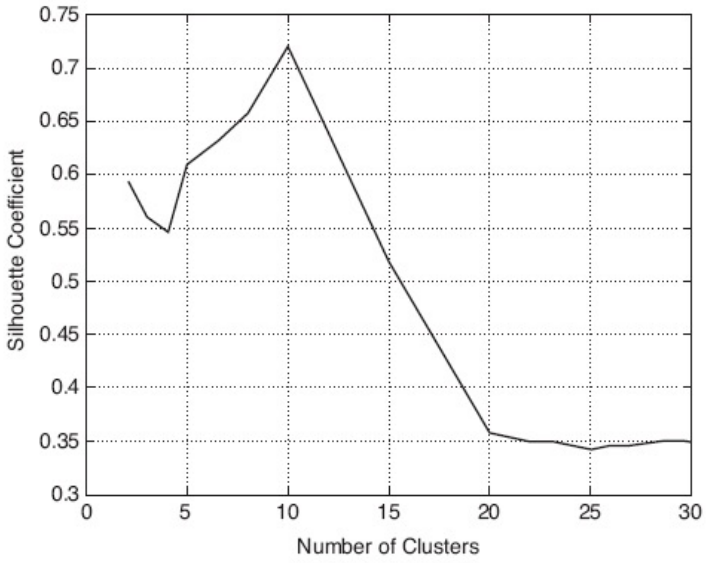
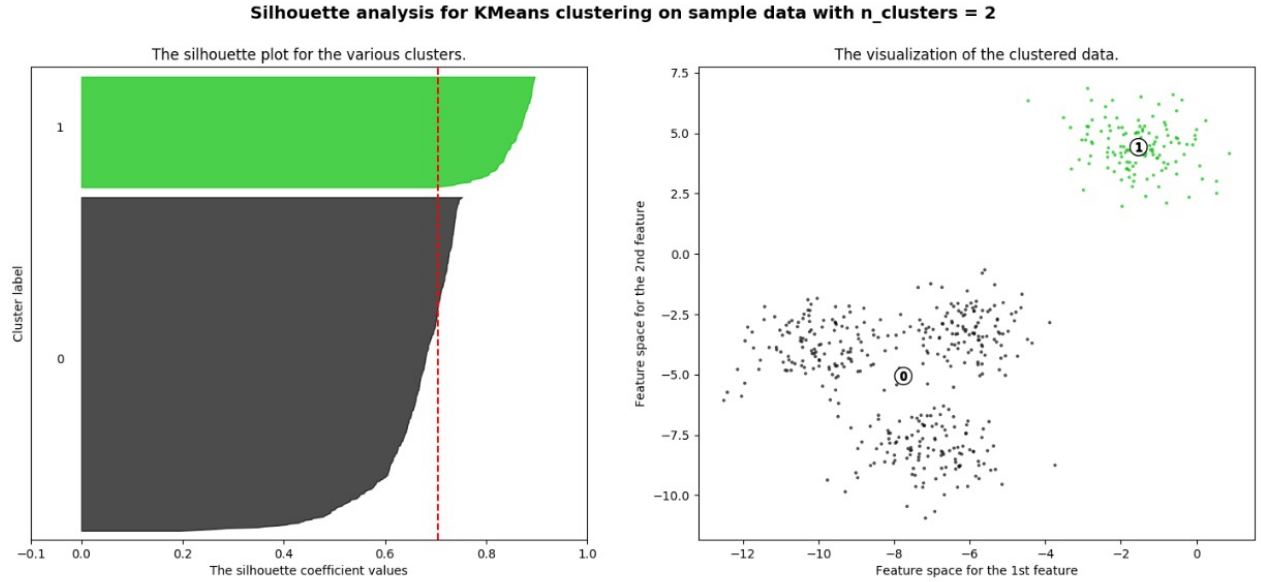
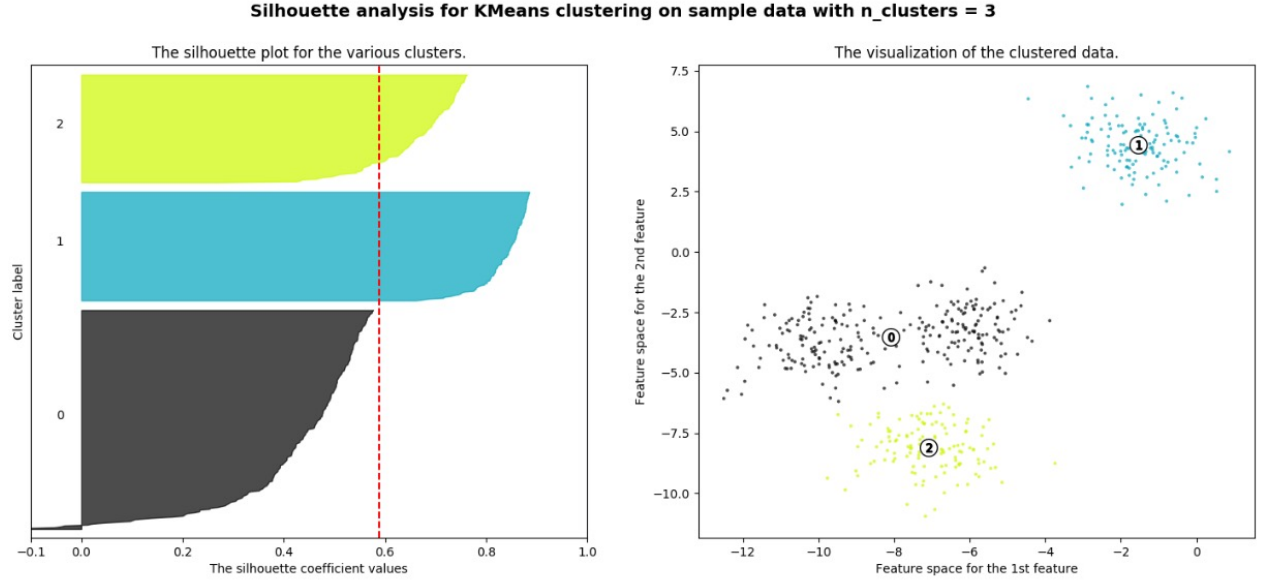
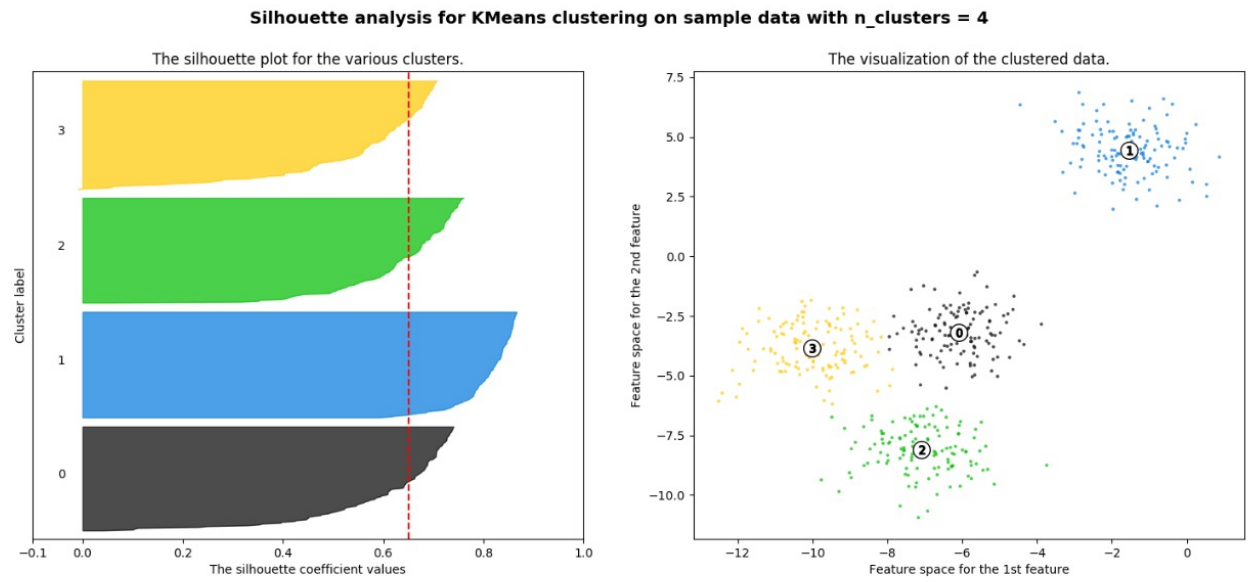


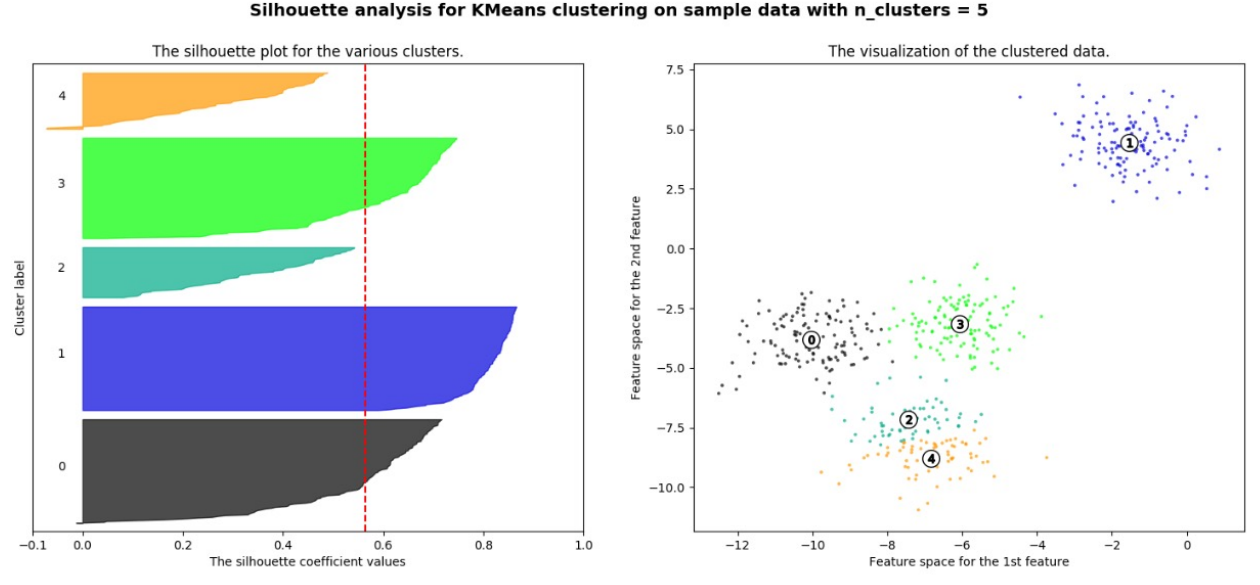
Figure 8.33. Average silhouette coefficient versus number of clusters for the data of Figure 8.29.

Silhouette plots



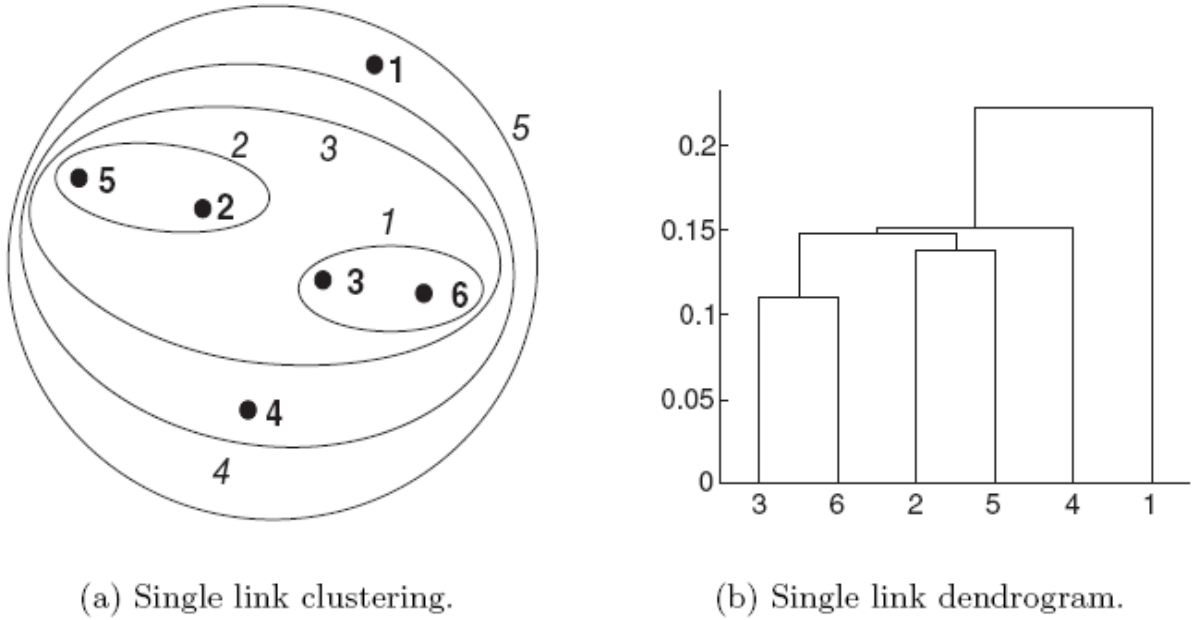






Evaluation of Hierarchical Clustering

- **Cophenetic distance** between two points is the proximity at which the agglomerative clustering put them in the same cluster



(a) Single link clustering. (b) Single link dendrogram.

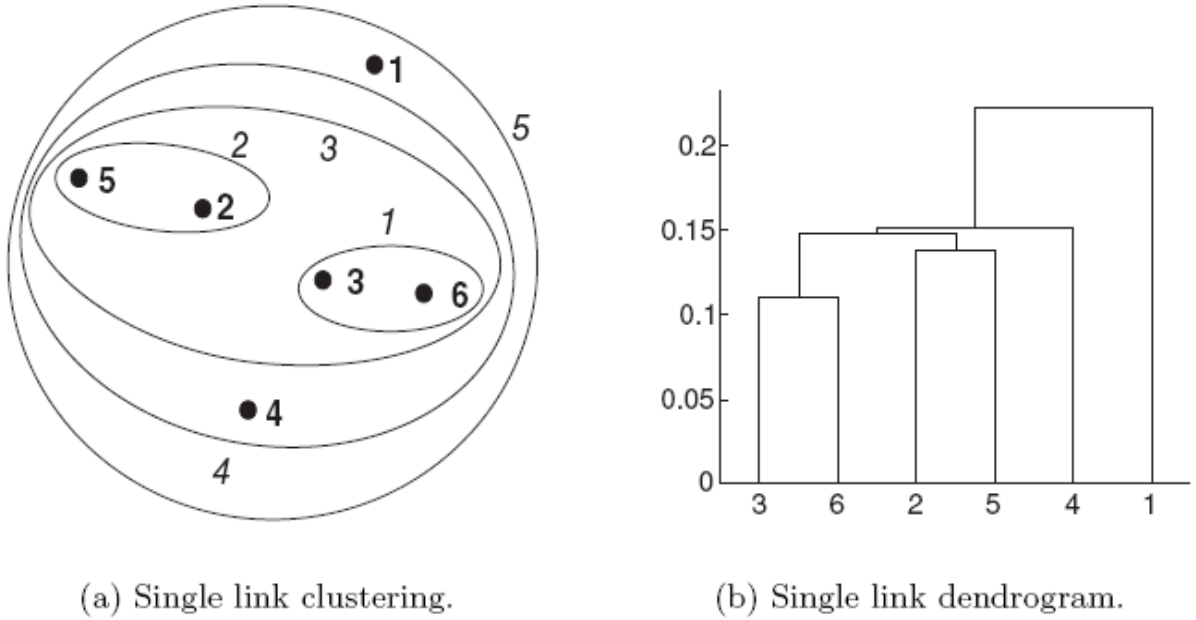
Figure 8.16. Single link clustering of the six points shown in Figure 8.15.

	p1	p2	p3	p4	p5	p6
p1	0					
p2		0				
p3			0			0.110
p4				0		
p5					0	
p6			0.110			0

Cophenetic distance matrix for single link

Evaluation of Hierarchical Clustering

- **Cophenetic distance** between two points is the proximity at which the agglomerative clustering put them in the same cluster



(a) Single link clustering. (b) Single link dendrogram.

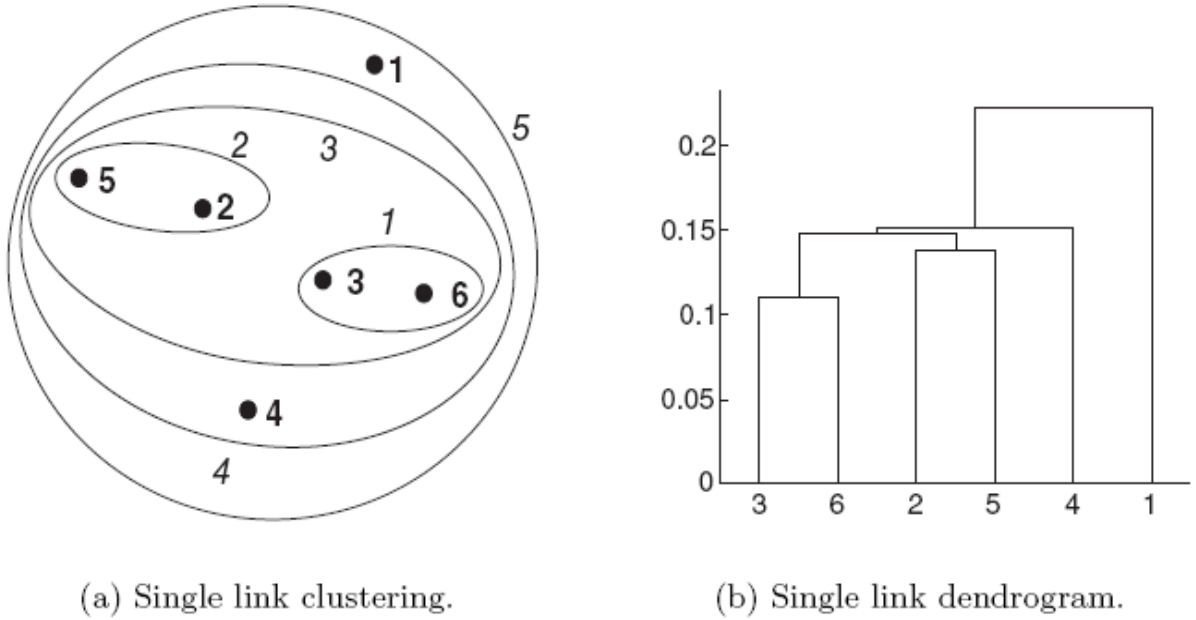
Figure 8.16. Single link clustering of the six points shown in Figure 8.15.

	p1	p2	p3	p4	p5	p6
p1	0					
p2		0			0.139	
p3			0			0.110
p4				0		
p5		0.139			0	
p6			0.110			0

Cophenetic distance matrix for single link

Evaluation of Hierarchical Clustering

- **Cophenetic distance** between two points is the proximity at which the agglomerative clustering put them in the same cluster



(a) Single link clustering. (b) Single link dendrogram.

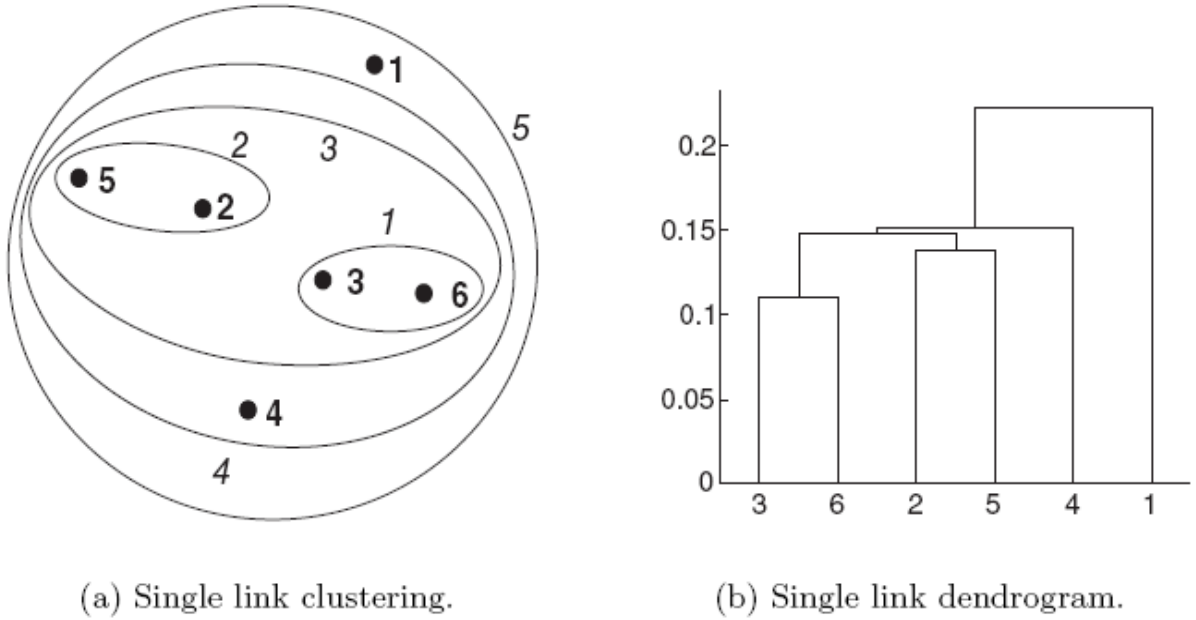
Figure 8.16. Single link clustering of the six points shown in Figure 8.15.

	p1	p2	p3	p4	p5	p6
p1	0					
p2		0	0.148		0.139	0.148
p3		0.148	0		0.148	0.110
p4				0		
p5		0.139	0.148		0	0.148
p6		0.148	0.110		0.148	0

Cophenetic distance matrix for single link

Evaluation of Hierarchical Clustering

- **Cophenetic distance** between two points is the proximity at which the agglomerative clustering put them in the same cluster



(a) Single link clustering. (b) Single link dendrogram.

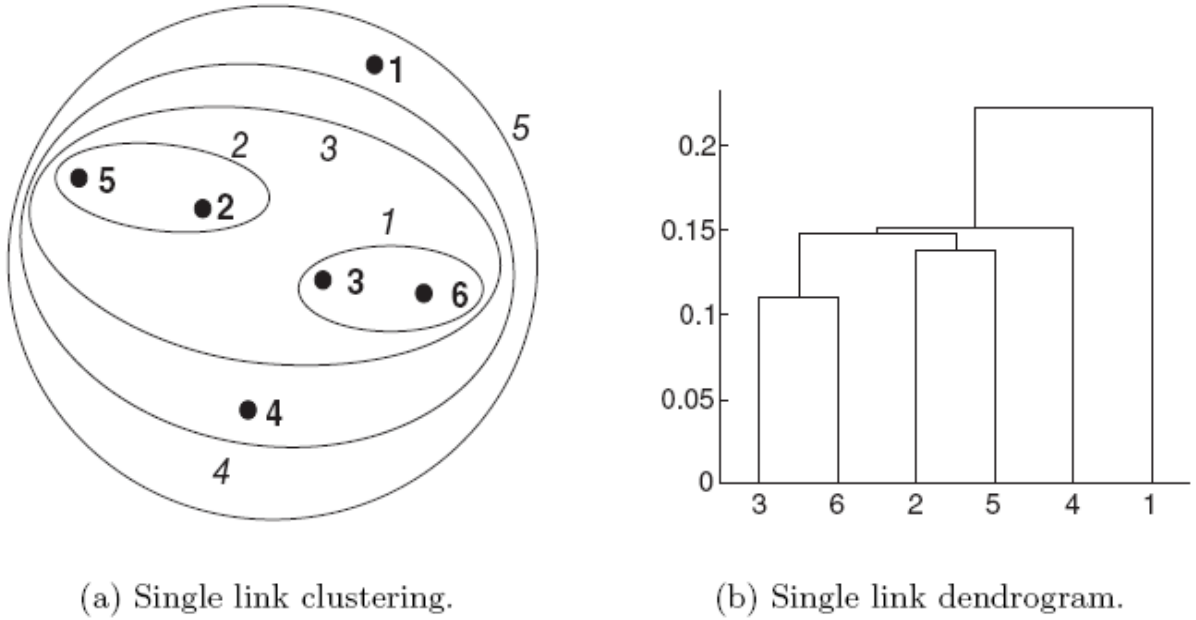
Figure 8.16. Single link clustering of the six points shown in Figure 8.15.

	p1	p2	p3	p4	p5	p6
p1	0					
p2		0	0.148	0.151	0.139	0.148
p3		0.148	0	0.151	0.148	0.110
p4		0.151	0.151	0	0.151	0.151
p5		0.139	0.148	0.151	0	0.148
p6		0.148	0.110	0.151	0.148	0

Cophenetic distance matrix for single link

Evaluation of Hierarchical Clustering

- **Cophenetic distance** between two points is the proximity at which the agglomerative clustering put them in the same cluster



(a) Single link clustering. (b) Single link dendrogram.

Figure 8.16. Single link clustering of the six points shown in Figure 8.15.

	p1	p2	p3	p4	p5	p6
p1	0	0.222	0.222	0.222	0.222	0.222
p2	0.222	0	0.148	0.151	0.139	0.148
p3	0.222	0.148	0	0.151	0.148	0.110
p4	0.222	0.151	0.151	0	0.151	0.151
p5	0.222	0.139	0.148	0.151	0	0.148
p6	0.222	0.148	0.110	0.151	0.148	0

Cophenetic distance matrix for single link

Cophenetic Correlation Coefficient (CPCC)

- Correlation between the cophenetic distance matrix and the proximity matrix of the original data points

	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	0.15	0.28	0.11
p4	0.37	0.20	0.15	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

Original proximity matrix

	p1	p2	p3	p4	p5	p6
p1	0	0.222	0.222	0.222	0.222	0.222
p2	0.222	0	0.148	0.151	0.139	0.148
p3	0.222	0.148	0	0.151	0.148	0.110
p4	0.222	0.151	0.151	0	0.151	0.151
p5	0.222	0.139	0.148	0.151	0	0.148
p6	0.222	0.148	0.110	0.151	0.148	0

Cophenetic distance matrix for single link

dist	Cdist
0.24	0.222
0.22	0.222
0.37	0.222
0.34	0.222
0.23	0.222
0.15	0.148
0.20	0.151
0.14	0.139
0.25	0.148
0.15	0.151
0.28	0.148
0.11	0.11
0.29	0.151
0.22	0.151
0.39	0.148

Pearson Correlation coefficient = 0.45

Type of Clustering	CPCC
Single Link	0.45
Complete Link	0.63
Group Average	0.66
Ward's	0.64

Comparison of agglomerative hierarchical clustering techniques on the same data set

External Measures

- Supervised Cluster Validation
- If we have external information about the class labels, we can measure the cluster labels against the class labels, using standard classification performance measures
  - Impurity/Entropy: The degree to which each cluster consists of objects of a single class
  - Precision: The fraction of a cluster that consists of objects of a specified class
  - Recall: The extent to which a cluster contains all objects of a specified class
  - F-measure: Combines precision and recall to measure the extent to which a cluster contains *only* objects of a particular class and *all* objects of that class

Example

Actual

Predicted

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

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

Confusion Matrix

Cluster 1, Metro class:

Precision (PPV)= 506/677 = 0.7474
Recall (TPR) = 506/943 = 0.537

Comparing Cluster and Class Matrices

Two Clusters
C_1={p1,p2,p3} C_2={p4,p5}

	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

Ideal cluster similarity matrix

Two Classes
L1={p1,p2} L2={p3,p4,p5}

	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

Ideal class similarity matrix

	Same Cluster	Different Cluster
Same Class		
Different Class		

Confusion Matrix

Correlation of the two matrices, Γ *Statistic* = 0.359

Comparing Cluster and Class Matricies

Two Clusters
C_1={p1,p2,p3} C_2={p4,p5}

	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

Ideal cluster similarity matrix

Two Classes
L1={p1,p2} L2={p3,p4,p5}

	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

Ideal class similarity matrix

	Same Cluster	Different Cluster
Same Class	(TP) 2	(FN) 2
Different Class	(FP) 2	(TN) 4

Confusion Matrix / Contingency Table

$$Rand\ Index = \frac{TP+TN}{TP+FP+TN+FN} = \frac{6}{10} = 0.6$$

$$Jaccard\ Coefficient = \frac{TP}{TP+FP+FN} = \frac{2}{6} = 0.333$$

Cluster Validation Summary

- **Clustering Tendency:** determine whether a non-random structure actually exists in the data
- **Internal Validation:** Evaluate how well the results of the clustering fit the data, *without* reference to external information (unsupervised)
- **External Validation:** Evaluate how well the results of the clustering fit the data, *with* externally provided class labels (supervised)
- **Relative Cluster Validation**:
  - Comparing two sets of clusters to determine which is better
    - Compare different clustering algorithms
    - Determine the correct number of clusters (K)
  - Compare two clusters to each other
  - Evaluate individual points – are they clustered well?