CS 484
Data Mining

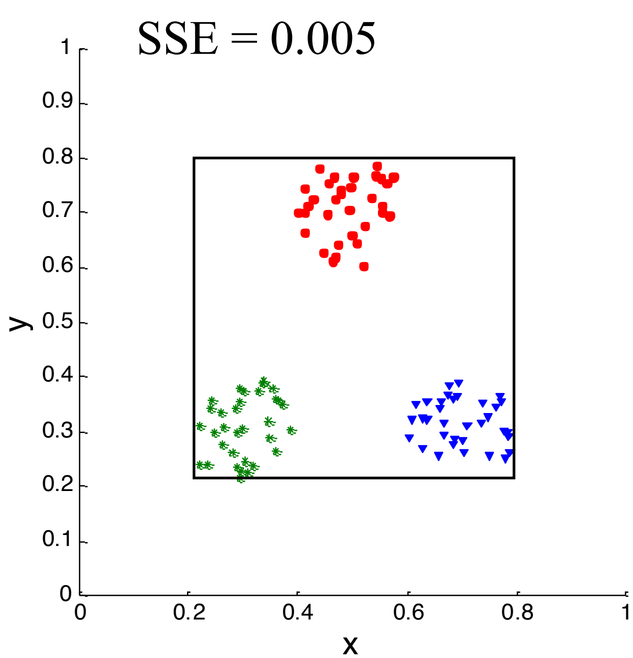
Clustering 6

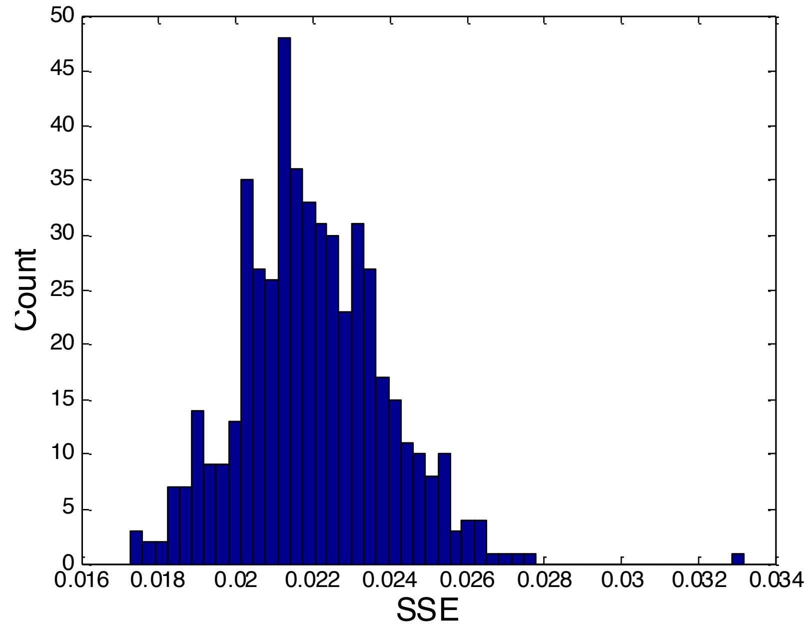
Framework for Cluster Validity

- 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 data or clusterings 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 SSE

- Example
  - Compare SSE of 0.005 against three clusters in random data
  - Histogram shows SSE of three clusters in 500 sets of random data points of size 100 distributed over the range 0.2 – 0.8 for x and y values





Internal Measures: Cohesion and Separation

- Cluster Cohesion: Measures how closely related are objects in a cluster
  - Example: SSE
- Cluster Separation: Measure how distinct or well-separated a cluster is from other clusters
- Example: Squared Error
  - Cohesion is measured by the within cluster sum of squares (SSE)

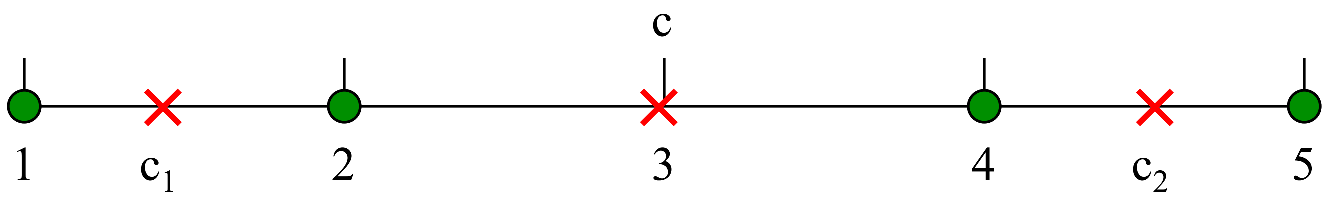
$$SSE = \sum_i \sum_{x \in C_i} (x - c_i)^2$$

  - Separation is measured by the between cluster sum of squares, or by between cluster to overall prototype sum of squares (shown)

$$SSB = \sum_i |C_i| (c - c_i)^2$$

where $|C_i|$ is the size of cluster i, c_i is the centroid of cluster i, and c is the overall centroid.

Total Sum of Squares (TSS)



K=1 cluster:

$$TSS = (1-3)^2 + (2-3)^2 + (4-3)^2 + (5-3)^2 = 10$$

$$SSE = (3-1)^2 + (3-2)^2 + (4-3)^2 + (5-3)^2 = 10$$

$$SSB = 4 \times (3-3)^2 = 0$$

K=2 clusters:

$$TSS = (1-3)^2 + (2-3)^2 + (4-3)^2 + (5-3)^2 = 10$$

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

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

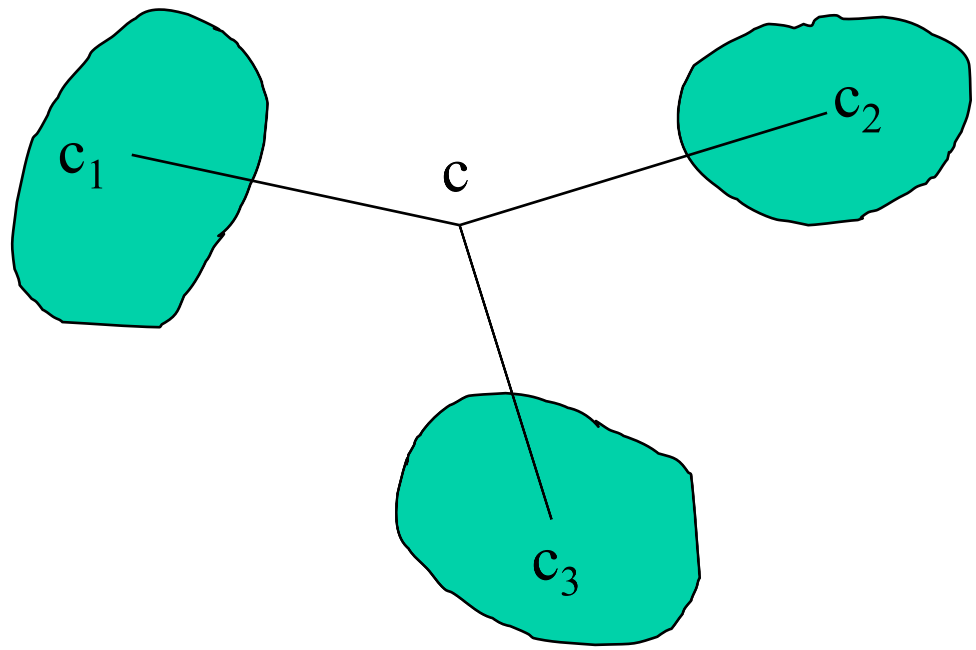
TSS = SSE + SSB

Total Sum of Squares (TSS)

$$TSS = \sum dist(x, c)^2$$

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

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



c: overall mean

c_i: centroid of each cluster C_i

$|C_i|$: number of points in cluster C_i

Total Sum of Squares (TSS)

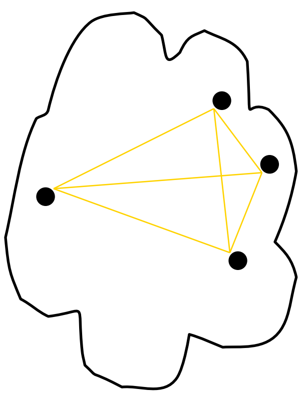
TSS = SSE + SSB

- Given a data set, TSS is fixed
- A clustering with large SSE has small SSB, while one with small SSE has large SSB

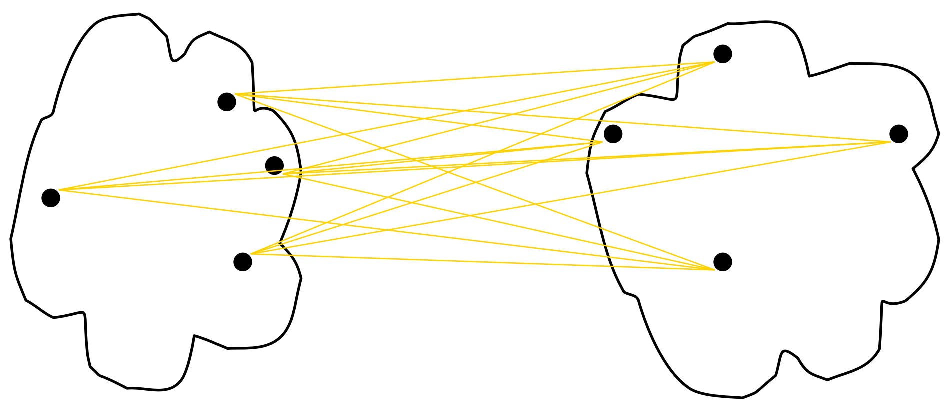
- Goal is to minimize SSE and maximize SSB

Internal Measures: Cohesion and Separation

- A proximity graph based approach can also be used for cohesion and separation.
  - Cluster cohesion is the sum of the weight of all links within a cluster.
  - Cluster separation is the sum of the weights between nodes in the cluster and nodes outside the cluster.



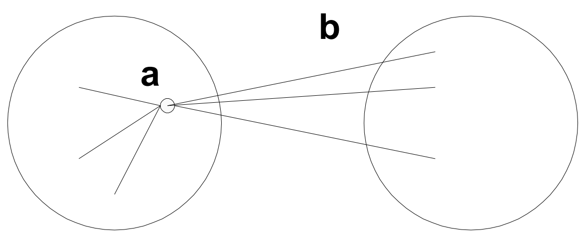
cohesion



separation

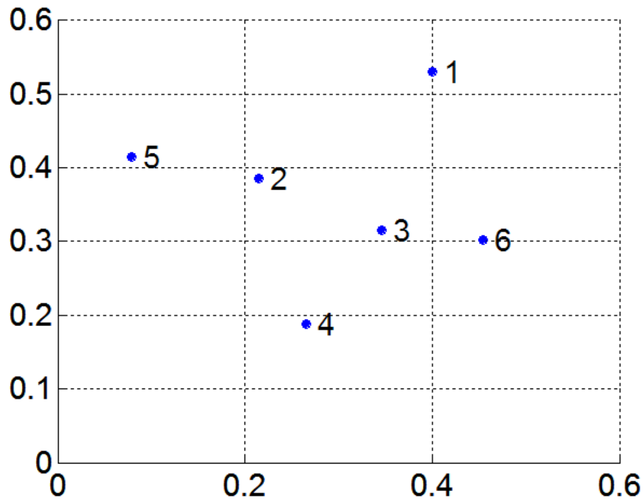
Internal Measures: Silhouette Coefficient

- Silhouette Coefficient combine ideas of both cohesion and separation, but for individual points, as well as clusters and clusterings
- 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)
  - The silhouette coefficient for a point is then given by
    s = 1 – a/b if a < b, (or s = b/a - 1 if a ≥ b, not the usual case)
  - Typically between 0 and 1 (but can be negative if **a ≥ b**).
  - The closer to 1 the better.



- Can calculate the Average Silhouette width for a cluster or a clustering

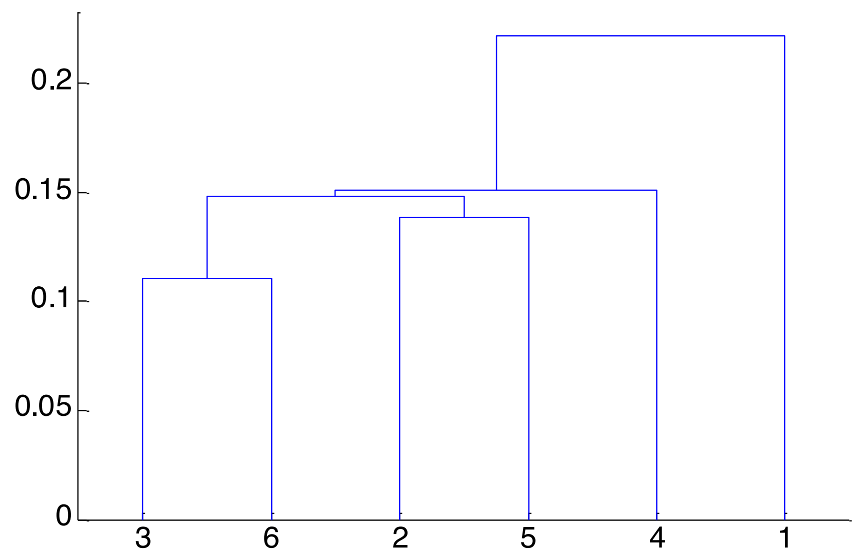
Unsupervised Evaluation of Hierarchical Clustering



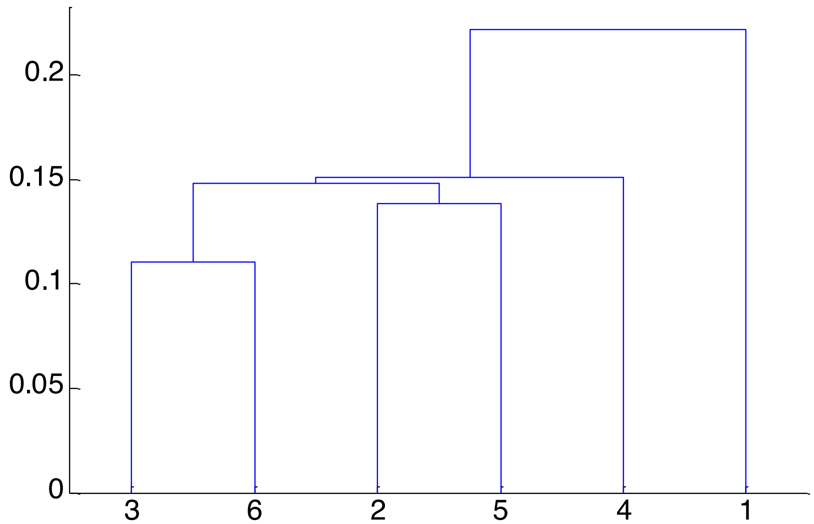
Distance Matrix:

	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

Single Link



Unsupervised Evaluation of Hierarchical Clustering



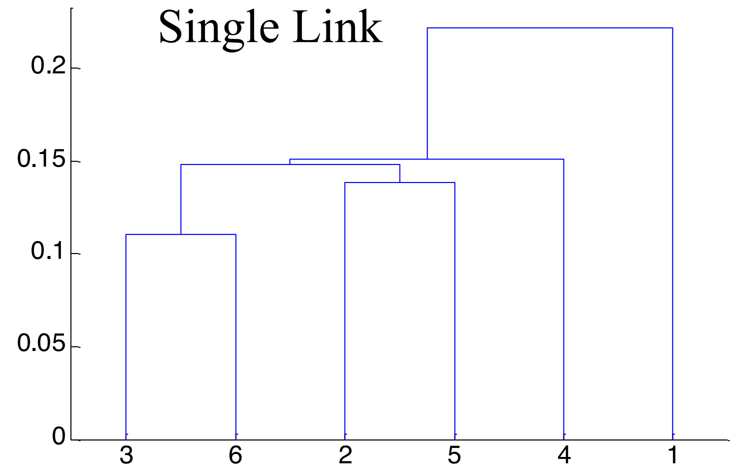
Single Link

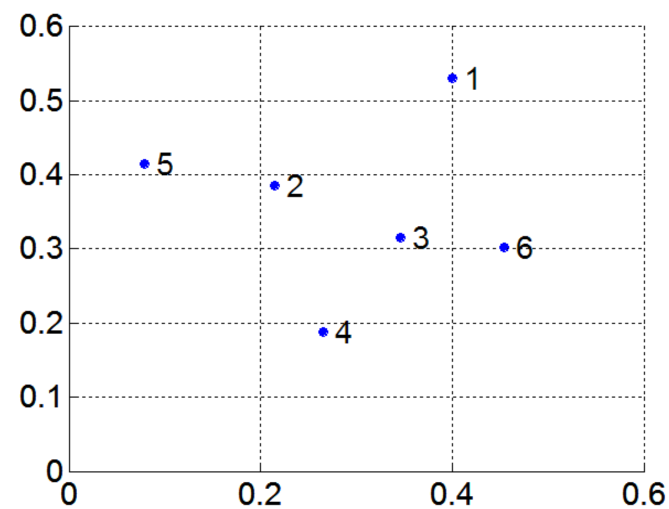
Point	P1	P2	P3	P4	P5	P6
P1	0	0.222	0.222	0.222	0.222	0.222
P2	0.222	0	0.148	0.151	0.139	0.148
P3	0.222	0.148	0	0.151	0.148	0.110
P4	0.222	0.151	0.151	0	0.151	0.151
P5	0.222	0.139	0.148	0.151	0	0.148
P6	0.222	0.148	0.110	0.151	0.148	0

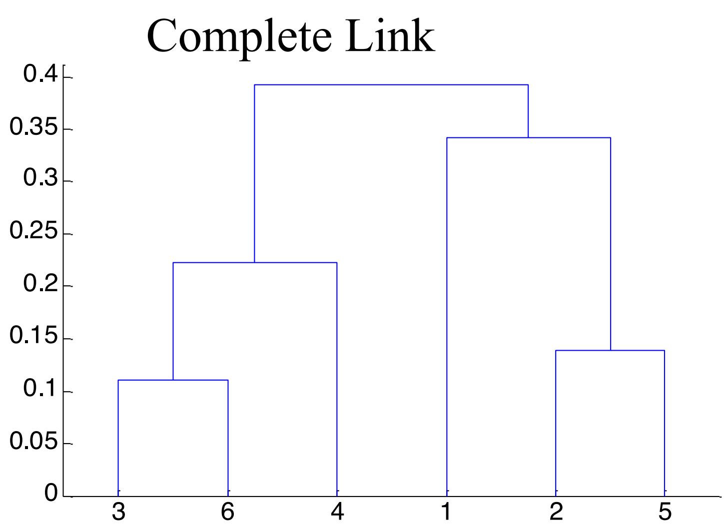
Cophenetic Distance Matrix for Single Link

- Cophenetic distance
  - the proximity at which the clustering technique puts the objects in the same cluster for the first time.
  - E.g. if two clusters are merged with distance = 0.1, then all points in one cluster have a cophenetic distance of 0.1 wrt the points in the other cluster.
- CPCC (CoPhenetic Correlation Coefficient)
  - Correlation between original distance matrix and cophenetic distance matrix

Unsupervised Evaluation of Hierarchical Clustering







Technique	CPCC
Single Link	0.44
Complete Link	0.63
Group Average	0.66
Ward's	0.64

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 $purity = \sum_{i=1}^{K} \frac{m_i}{m} purity_j$.

Supervised Cluster Validation: Precision and Recall

Cluster i

m_{i1}: class 1

m_{i2}: class 2

Overall Data

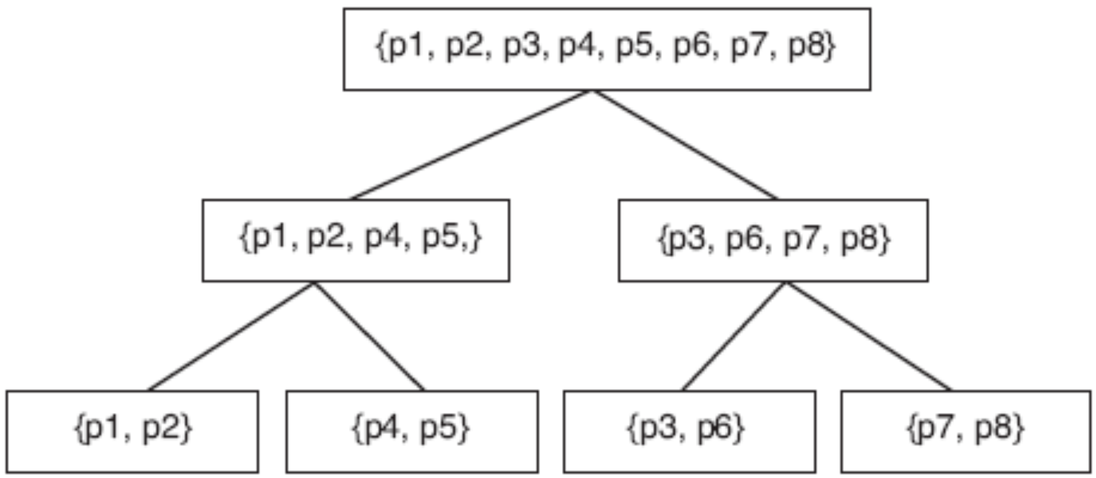
m_1: class 1

m_2: class 2

- Precision for cluster i w.r.t. class j = $\dfrac{m_{ij}}{\sum_k m_{ik}}$

- Recall for cluster i w.r.t. class j = $\dfrac{m_{ij}}{\sum_k m_{kj}} = \dfrac{m_{ij}}{m_j}$

Supervised Cluster Validation: Hierarchical Clustering



Hierarchical F-measure:

$$F = \sum_j \frac{m_j}{m} \max_i F(i, j)$$

where the maximum is taken over all clusters i at all levels, m_j is the number of objects in class j, and m is the total number of objects.

Supervised Cluster Validation: Binary Similarity

- Consider all pairs of distinct objects
  - f_{00} = # of pairs of objects having a different class and a different cluster
  - f_{01} = # of pairs of objects having a different class and the same cluster
  - f_{10} = # of pairs of objects having the same class and a different cluster
  - f_{11} = # of pairs of objects having the same class and the same cluster

Supervised Cluster Validation: Binary Similarity

- Rand Statistic (Simple matching coefficient):

$$\frac{f_{00} + f_{11}}{f_{00} + f_{01} + f_{10} + f_{11}}$$

- Jaccard Coefficient:

$$\frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

	Same Cluster	**Different Cluster**
Same Class	f11	f10
Different Class	f01	f00

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