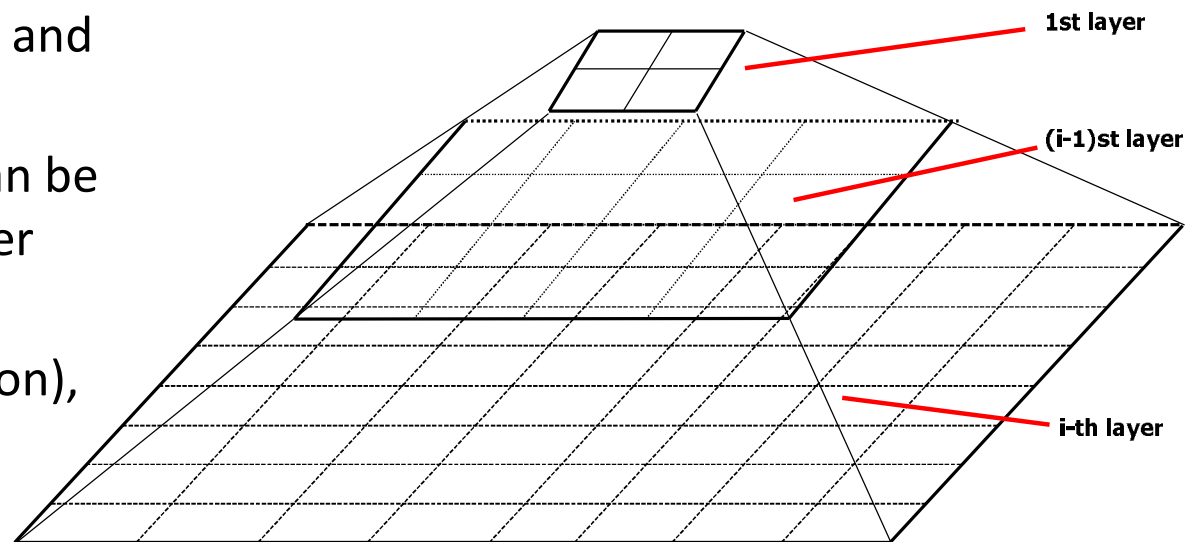


Data Mining Fundamentals

Chapter 10. Cluster Analysis: Basic Concepts and Methods

STING: A Statistical Information Grid Approach

- ❑ STING (Statistical Information Grid) (Wang, Yang and Muntz, VLDB'97)
- ❑ The spatial area is divided into rectangular cells at different levels of resolution, and these cells form a tree structure
- ❑ A cell at a high level contains a number of smaller cells of the next lower level
- ❑ Statistical information of each cell is calculated and stored beforehand and is used to answer queries
- ❑ Parameters of higher level cells can be easily calculated from that of lower level cell, including
 - ❑ *count, mean, s*(standard deviation), *min, max*
 - ❑ type of distribution—*normal, uniform, etc.*



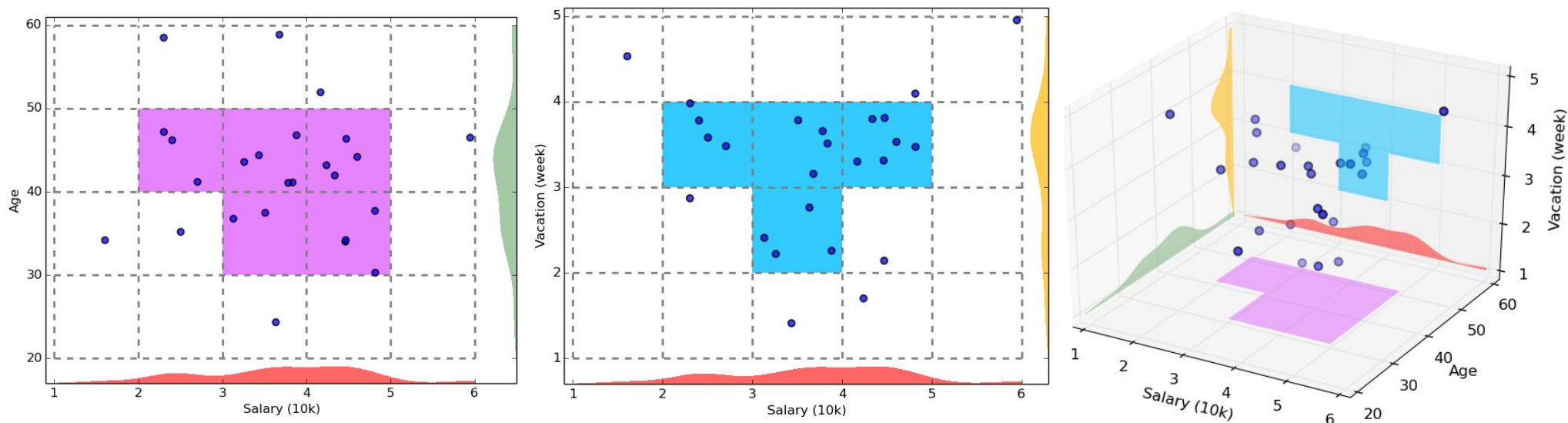
Query Processing in STING and Its Analysis

- ❑ To process a region query
 - ❑ Start at the root and proceed to the next lower level, using the STING index
 - ❑ Calculate the likelihood that a cell is relevant to the query at some confidence level using the statistical information of the cell
 - ❑ Only children of likely relevant cells are recursively explored
 - ❑ Repeat this process until the bottom layer is reached
- ❑ Advantages
 - ❑ Query-independent, easy to parallelize, incremental update
 - ❑ Efficiency: Complexity is $O(K)$
 - ❑ K : # of grid cells at the lowest level, and $K \ll N$ (i.e., # of data points)
- ❑ Disadvantages
 - ❑ Its probabilistic nature may imply a loss of accuracy in query processing

CLIQUE: Grid-Based Subspace Clustering

- ❑ CLIQUE (Clustering In QUEst) (Agrawal, Gehrke, Gunopulos, Raghavan: SIGMOD'98)
- ❑ CLIQUE is a **density-based** and **grid-based** **subspace clustering** algorithm
 - ❑ **Grid-based**: It discretizes the data space through a grid and estimates the density by counting the number of points in a grid cell
 - ❑ **Density-based**: A cluster is a maximal set of connected dense units in a subspace
 - ❑ A unit is dense if the fraction of total data points contained in the unit exceeds the input model parameter
 - ❑ **Subspace clustering**: A subspace cluster is a set of neighboring dense cells in an arbitrary subspace. It also discovers some minimal descriptions of the clusters
- ❑ It automatically identifies subspaces of a high dimensional data space that allow better clustering than original space using the Apriori principle

CLIQUE: SubSpace Clustering with Apriori Pruning



- ❑ Start at 1-D space and discretize numerical intervals in each axis into grid
- ❑ Find dense regions (clusters) in each subspace and generate their minimal descriptions
- ❑ Use the dense regions to find promising candidates in 2-D space based on the Apriori principle
- ❑ Repeat the above in level-wise manner in higher dimensional subspaces

Major Steps of the CLIQUE Algorithm

- ❑ Identify subspaces that contain clusters
 - ❑ Partition the data space and find the number of points that lie inside each cell of the partition
 - ❑ Identify the subspaces that contain clusters using the Apriori principle
- ❑ Identify clusters
 - ❑ Determine dense units in all subspaces of interests
 - ❑ Determine connected dense units in all subspaces of interests
- ❑ Generate minimal descriptions for the clusters
 - ❑ Determine maximal regions that cover a cluster of connected dense units for each cluster
 - ❑ Determine minimal cover for each cluster

Additional Comments on *CLIQUE*


□ Strengths

- *Automatically* finds subspaces of the highest dimensionality as long as high density clusters exist in those subspaces
- *Insensitive* to the order of records in input and does not presume some canonical data distribution
- Scales *linearly* with the size of input and has good scalability as the number of dimensions in the data increases

□ Weaknesses

- As in all grid-based clustering approaches, the quality of the results crucially depends on the appropriate choice of the number and width of the partitions and grid cells

Chapter 10. Cluster Analysis: Basic Concepts and Methods

- ❑ Cluster Analysis: An Introduction
- ❑ Partitioning Methods
- ❑ Hierarchical Methods
- ❑ Density- and Grid-Based Methods
- ❑ Evaluation of Clustering 
- ❑ Summary

Clustering Validation

- ❑ Clustering Validation: Basic Concepts
- ❑ Clustering Evaluation: Measuring Clustering Quality
- ❑ External Measures for Clustering Validation
 - ❑ I: Matching-Based Measures
 - ❑ II: Entropy-Based Measures
 - ❑ III: Pairwise Measures
- ❑ Internal Measures for Clustering Validation
- ❑ Relative Measures
- ❑ Cluster Stability
- ❑ Clustering Tendency

Clustering Validation and Assessment

- Major issues on clustering validation and assessment
 - **Clustering evaluation**
 - Evaluating the goodness of the clustering
 - **Clustering stability**
 - To understand the sensitivity of the clustering result to various algorithm parameters, e.g., # of clusters
 - **Clustering tendency**
 - Assess the suitability of clustering, i.e., whether the data has any inherent grouping structure

Measuring Clustering Quality

- ❑ **Clustering Evaluation:** Evaluating the goodness of clustering results
 - ❑ No commonly recognized best suitable measure in practice
- ❑ **Three categorization of measures:** External, internal, and relative
 - ❑ **External:** Supervised, employ criteria not inherent to the dataset
 - ❑ Compare a clustering against prior or expert-specified knowledge (i.e., the ground truth) using certain clustering quality measure
 - ❑ **Internal:** Unsupervised, criteria derived from data itself
 - ❑ Evaluate the goodness of a clustering by considering how well the clusters are separated and how compact the clusters are, e.g., silhouette coefficient
 - ❑ **Relative:** Directly compare different clusterings, usually those obtained via different parameter settings for the same algorithm

Measuring Clustering Quality: External Methods

- Given the **ground truth** T , $Q(C, T)$ is the **quality measure** for a clustering C
- $Q(C, T)$ is good if it satisfies the following **four** essential criteria
 - **Cluster homogeneity**
 - The purer, the better
 - **Cluster completeness**
 - Assign objects belonging to the same category in the ground truth to the same cluster
 - **Rag bag better than alien**
 - Putting a heterogeneous object into a pure cluster should be penalized more than putting it into a *rag bag* (i.e., “miscellaneous” or “other” category)
 - **Small cluster preservation**
 - Splitting a small category into pieces is more harmful than splitting a large category into pieces

Commonly Used External Measures

❑ Matching-based measures

- ❑ Purity, maximum matching, F-measure

❑ Entropy-Based Measures

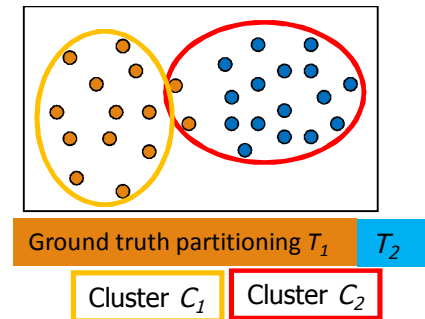
- ❑ Conditional entropy
- ❑ Normalized mutual information (NMI)
- ❑ Variation of information

❑ Pairwise measures

- ❑ Four possibilities: True positive (TP), FN, FP, TN
- ❑ Jaccard coefficient, Rand statistic, Fowlkes-Mallow measure

❑ Correlation measures

- ❑ Discretized Huber static, normalized discretized Huber static



Matching-Based Measures (I): Purity vs. Maximum Matching

- **Purity:** Quantifies the extent that cluster C_i contains points only from one (ground truth) partition:

$$purity_i = \frac{1}{n_i} \max_{j=1}^k \{n_{ij}\}$$

- Total purity of clustering C :

$$purity = \sum_{i=1}^r \frac{n_i}{n} purity_i = \frac{1}{n} \sum_{i=1}^r \max_{j=1}^k \{n_{ij}\}$$

- Perfect clustering if $purity = 1$ and $r = k$ (the number of clusters obtained is the same as that in the ground truth)

- Ex. 1 (green or orange): $purity_1 = 30/50$; $purity_2 = 20/25$; $purity_3 = 25/25$; $purity = (30 + 20 + 25)/100 = 0.75$

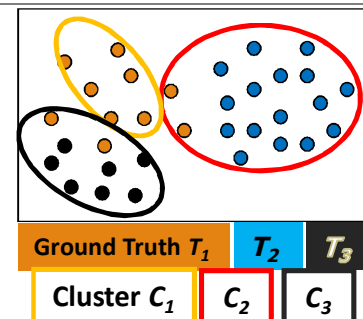
- Two clusters may share the same majority partition

- **Maximum matching:** Only one cluster can match one partition

- Match: Pairwise matching, weight $w(e_{ij}) = n_{ij}$ $w(M) = \sum_{e \in M} w(e)$

- Maximum weight matching: $match = \arg \max_M \left\{ \frac{w(M)}{n} \right\}$

- Ex2. (green) $match = purity = 0.75$; (orange) $match = 0.65 > 0.6$



$C \setminus T$	T_1	T_2	T_3	Sum
C_1	0	20	30	50
C_2	0	20	5	25
C_3	25	0	0	25
m_j	25	40	35	100

$C \setminus T$	T_1	T_2	T_3	Sum
C_1	0	30	20	50
C_2	0	20	5	25
C_3	25	0	0	25
m_j	25	50	25	100

Matching-Based Measures (II): F-Measure

- **Precision:** The fraction of points in C_i from the majority partition T_{j_i} (i.e., the same as purity), where j_i is the partition that contains the maximum # of points from C_i

$$prec_i = \frac{1}{n_i} \max_{j=1}^k \{n_{ij}\} = \frac{n_{ij_i}}{n_i}$$

- Ex. For the green table

- $prec_1 = 30/50$; $prec_2 = 20/25$; $prec_3 = 25/25$

- **Recall:** The fraction of point in partition T_{j_i} shared in common with cluster C_i , where $m_{j_i} = |T_{j_i}|$

$$recall_i = \frac{n_{ij_i}}{|T_{j_i}|} = \frac{n_{ij_i}}{m_{j_i}}$$

- Ex. For the green table

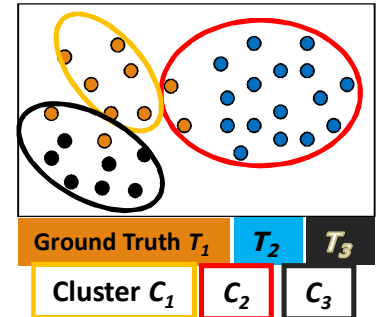
- $recall_1 = 30/35$; $recall_2 = 20/40$; $recall_3 = 25/25$

- **F-measure** for C_i : The harmonic means of $prec_i$ and $recall_i$: $F_i = \frac{2n_{ij_i}}{n_i + m_{j_i}}$

- F-measure for clustering C : average of all clusters: $F = \frac{1}{r} \sum_{i=1}^r F_i$

- Ex. For the green table

- $F_1 = 60/85$; $F_2 = 40/65$; $F_3 = 1$; $F = 0.774$



C \ T	T ₁	T ₂	T ₃	Sum
C ₁	0	20	30	50
C ₂	0	20	5	25
C ₃	25	0	0	25
m _j	25	40	35	100

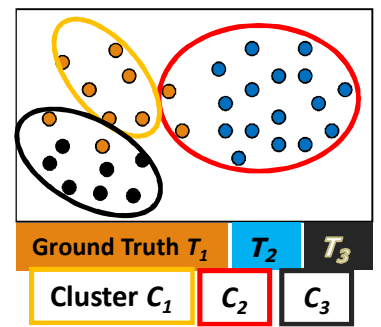
Entropy-Based Measures (I): Conditional Entropy

□ Entropy of clustering \mathcal{C} : $H(\mathcal{C}) = - \sum_{i=1}^r p_{C_i} \log p_{C_i}$ $p_{C_i} = \frac{n_i}{n}$ (i.e., the probability of cluster C_i)

□ Entropy of partitioning \mathcal{T} : $H(\mathcal{T}) = - \sum_{j=1}^k p_{T_j} \log p_{T_j}$

□ Entropy of \mathcal{T} with respect to cluster C_j : $H(\mathcal{T}|C_i) = - \sum_{j=1}^k \left(\frac{n_{ij}}{n_i}\right) \log\left(\frac{n_{ij}}{n_i}\right)$

□ Conditional entropy of \mathcal{T} with respect to clustering \mathcal{C} : $H(\mathcal{T}|\mathcal{C}) = - \sum_{i=1}^r \left(\frac{n_i}{n}\right) H(\mathcal{T}|C_i) = - \sum_{i=1}^r \sum_{j=1}^k p_{ij} \log\left(\frac{p_{ij}}{p_{C_i}}\right)$



- The more a cluster's members are split into different partitions, the higher the conditional entropy
- For a perfect clustering, the conditional entropy value is 0, where the worst possible conditional entropy value is $\log k$

$$\begin{aligned}
 H(\mathcal{T}|\mathcal{C}) &= - \sum_{i=1}^r \sum_{j=1}^k p_{ij} (\log p_{ij} - \log p_{C_i}) = - \sum_{i=1}^r \sum_{j=1}^k p_{ij} \log p_{ij} + \sum_{i=1}^r (\log p_{C_i} \sum_{j=1}^k p_{ij}) \\
 &= - \sum_{i=1}^r \sum_{j=1}^k p_{ij} \log p_{ij} + \sum_{i=1}^r (p_{C_i} \log p_{C_i}) = H(\mathcal{C}, \mathcal{T}) - H(\mathcal{C})
 \end{aligned}$$

Pairwise Measures: Four Possibilities for Truth Assignment

Four possibilities based on the agreement between cluster label and partition label

TP: true positive—Two points \mathbf{x}_i and \mathbf{x}_j belong to the same partition T , and they also in the same cluster C

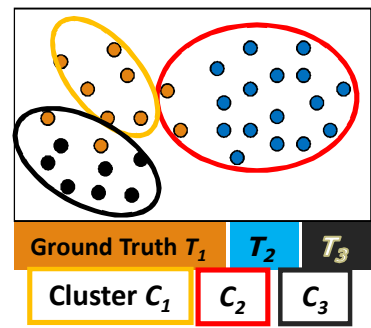
$$TP = |\{(\mathbf{x}_i, \mathbf{x}_j) : y_i = y_j \text{ and } \hat{y}_i = \hat{y}_j\}|$$

where y_i : the true partition label, and \hat{y}_i : the cluster label for point \mathbf{x}_i

FN: false negative: $FN = |\{(\mathbf{x}_i, \mathbf{x}_j) : y_i = y_j \text{ and } \hat{y}_i \neq \hat{y}_j\}|$

FP: false positive $FP = |\{(\mathbf{x}_i, \mathbf{x}_j) : y_i \neq y_j \text{ and } \hat{y}_i = \hat{y}_j\}|$

TN: true negative $TN = |\{(\mathbf{x}_i, \mathbf{x}_j) : y_i \neq y_j \text{ and } \hat{y}_i \neq \hat{y}_j\}|$



Calculate the four measures:

$$N = \binom{n}{2}$$

Total # of pairs of points

$$TP = \sum_{i=1}^r \sum_{j=1}^k \binom{n_{ij}}{2} = \frac{1}{2} \left(\sum_{i=1}^r \sum_{j=1}^k n_{ij}^2 \right) - n$$

$$FN = \sum_{j=1}^k \binom{m_j}{2} - TP$$

$$FP = \sum_{i=1}^r \binom{n_i}{2} - TP$$

$$TN = N - (TP + FN + FP) = \frac{1}{2} \left(n^2 - \sum_{i=1}^r n_i^2 - \sum_{j=1}^k m_j^2 + \sum_{i=1}^r \sum_{j=1}^k n_{ij}^2 \right)$$

Pairwise Measures: Jaccard Coefficient and Rand Statistic

- ❑ **Jaccard coefficient:** Fraction of true positive point pairs, but after ignoring the true negatives (thus asymmetric)
 - ❑ $Jaccard = TP / (TP + FN + FP)$ [i.e., denominator ignores TN]
 - ❑ Perfect clustering: $Jaccard = 1$

- ❑ **Rand Statistic:**

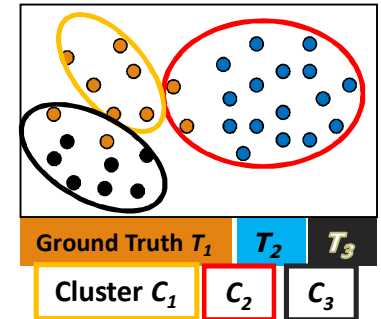
- ❑ $Rand = (TP + TN) / N$
 - ❑ Symmetric; perfect clustering: $Rand = 1$

- ❑ **Fowlkes-Mallow Measure:**

- ❑ Geometric mean of precision and recall

$$FM = \sqrt{prec \times recall} = \frac{TP}{\sqrt{(TP + FN)(TP + FP)}}$$

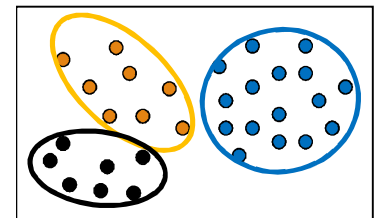
- ❑ Using the above formulas, one can calculate all the measures for the green table (leave as an exercise)



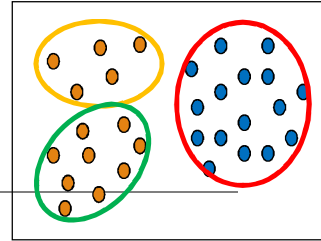
$C \setminus T$	T_1	T_2	T_3	Sum
C_1	0	20	30	50
C_2	0	20	5	25
C_3	25	0	0	25
m_j	25	40	35	100

Internal Measures (I): BetaCV Measure

- A trade-off in maximizing intra-cluster compactness and inter-cluster separation
- Given a clustering $C = \{C_1, \dots, C_k\}$ with k clusters, cluster C_i containing $n_i = |C_i|$ points
 - Let $W(S, R)$ be sum of weights on all edges with one vertex in S and the other in R
 - The sum of all the intra-cluster weights over all clusters: $W_{in} = \frac{1}{2} \sum_{i=1}^k W(C_i, C_i)$
 - The sum of all the inter-cluster weights: $W_{out} = \frac{1}{2} \sum_{i=1}^k W(C_i, \bar{C}_i) = \sum_{i=1}^{k-1} \sum_{j>i} W(C_i, C_j)$
 - The number of distinct intra-cluster edges: $N_{in} = \sum_{i=1}^k \binom{n_i}{2}$
 - The number of distinct inter-cluster edges: $N_{out} = \sum_{i=1}^{k-1} \sum_{j=i+1}^k n_i n_j$
- **Beta-CV measure:** $BetaCV = \frac{W_{in} / N_{in}}{W_{out} / N_{out}}$
 - The ratio of the mean intra-cluster distance to the mean inter-cluster distance
 - The smaller, the better the clustering



Relative Measure



- Relative measure: Directly compare different clusterings, usually those obtained via different parameter settings for the same algorithm

- Silhouette coefficient as an internal measure:** Check cluster cohesion and separation

- For each point \mathbf{x}_i , its silhouette coefficient s_i is: $s_i = \frac{\mu_{out}^{\min}(\mathbf{x}_i) - \mu_{in}(\mathbf{x}_i)}{\max\{\mu_{out}^{\min}(\mathbf{x}_i), \mu_{in}(\mathbf{x}_i)\}}$
 where $\mu_{in}(\mathbf{x}_i)$ is the mean distance from \mathbf{x}_i to points in its own cluster

$\mu_{out}^{\min}(\mathbf{x}_i)$ is the mean distance from \mathbf{x}_i to points in its closest cluster

- Silhouette coefficient (SC) is the mean values of s_i across all the points: $SC = \frac{1}{n} \sum_{i=1}^n s_i$
- SC close to +1 implies good clustering

- Points are close to their own clusters but far from other clusters

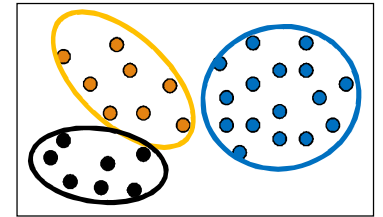
- Silhouette coefficient as a relative measure:** Estimate the # of clusters in the data

$$SC_i = \frac{1}{n_i} \sum_{x_j \in C_i} s_j$$

Pick the k value that yields the best clustering, i.e., yielding high values for SC and SC_i ($1 \leq i \leq k$)

Cluster Stability

- ❑ Clusterings obtained from several datasets sampled from the same underlying distribution as \mathbf{D} should be similar or “stable”
- ❑ Typical approach:
 - ❑ Find good parameter values for a given clustering algorithm
- ❑ Example: Find a good value of k , the correct number of clusters
- ❑ A **bootstrapping approach** to find the best value of k (judged on stability)
 - ❑ Generate t samples of size n by sampling from \mathbf{D} with replacement
 - ❑ For each sample \mathbf{D}_i , run the same clustering algorithm with k values from 2 to k_{max}
 - ❑ Compare the distance between all pairs of clusterings $C_k(\mathbf{D}_i)$ and $C_k(\mathbf{D}_j)$ via some distance function
 - ❑ Compute the expected pairwise distance for each value of k
 - ❑ The value k^* that exhibits the least deviation between the clusterings obtained from the resampled datasets is the best choice for k since it exhibits the most stability



Other Methods for Finding K, the Number of Clusters

□ Empirical method

□ # of clusters: $k \approx \sqrt{n/2}$ for a dataset of n points (e.g., $n = 200$, $k = 10$)

□ **Elbow method:** Use the turning point in the curve of the sum of within cluster variance with respect to the # of clusters

□ Cross validation method

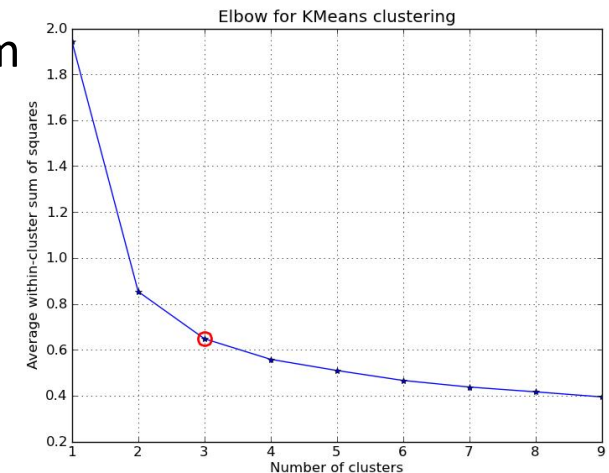
□ Divide a given data set into m parts

□ Use $m - 1$ parts to obtain a clustering model

□ Use the remaining part to test the quality of the clustering

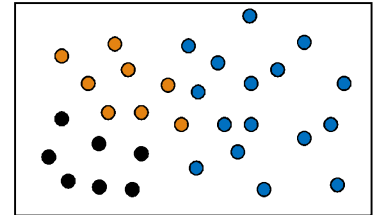
□ For example, for each point in the test set, find the closest centroid, and use the sum of squared distance between all points in the test set and the closest centroids to measure how well the model fits the test set

□ For any $k > 0$, repeat it m times, compare the overall quality measure w.r.t. different k 's, and find # of clusters that fits the data the best




Clustering Tendency: Whether the Data Contains Inherent Grouping Structure

- Assessing the **suitability of clustering**
 - (i.e., whether the data has any inherent grouping structure)
- Determining **clustering tendency** or **clusterability**
 - **A hard task** because there are so many different definitions of clusters
 - E.g., partitioning, hierarchical, density-based, graph-based, etc.
 - Even fixing cluster type, still hard to define an appropriate null model for a data set
- Still, there are some **clusterability assessment methods**, such as
 - **Spatial histogram**: Contrast the histogram of the data with that generated from random samples
 - **Distance distribution**: Compare the pairwise point distance from the data with those from the randomly generated samples
 - **Hopkins Statistic**: A sparse sampling test for spatial randomness



Chapter 10. Cluster Analysis: Basic Concepts and Methods

- ❑ Cluster Analysis: An Introduction
- ❑ Partitioning Methods
- ❑ Hierarchical Methods
- ❑ Density- and Grid-Based Methods
- ❑ Evaluation of Clustering
- ❑ Summary 

Summary

- ❑ Cluster Analysis: An Introduction
- ❑ Partitioning Methods
- ❑ Hierarchical Methods
- ❑ Density- and Grid-Based Methods
- ❑ Evaluation of Clustering