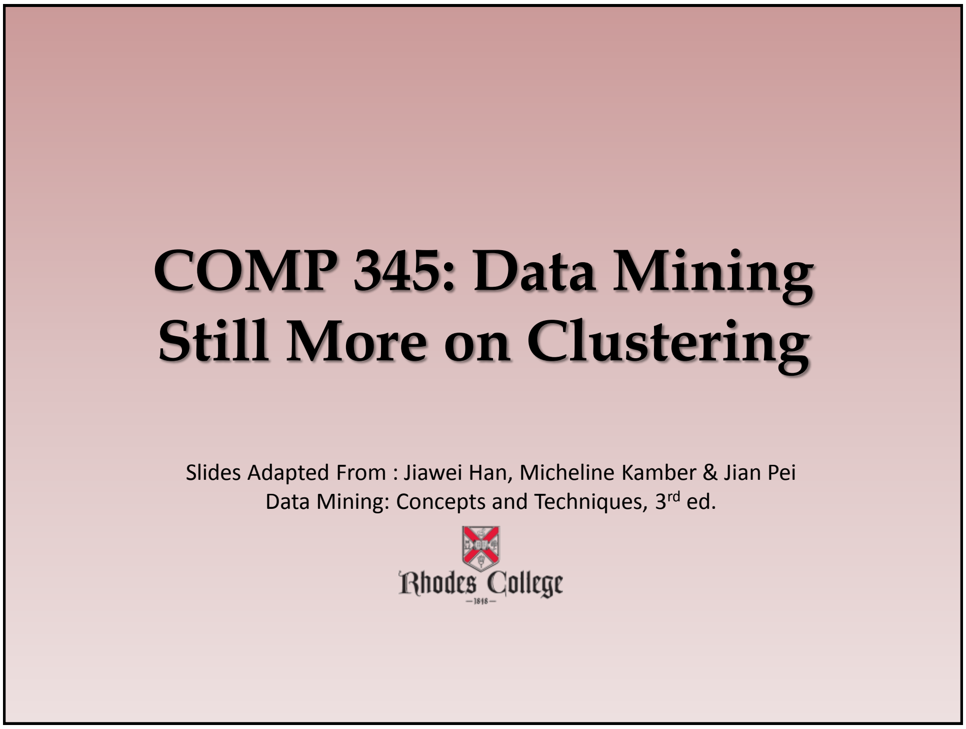
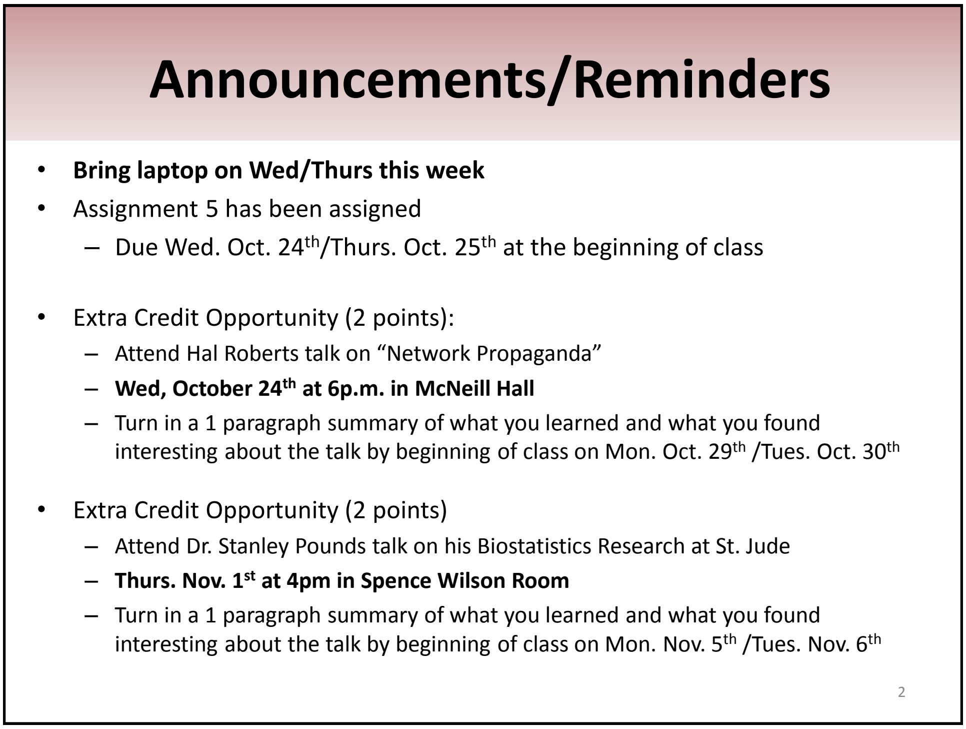


COMP 345: Data Mining
Still More on Clustering

Slides Adapted From : Jiawei Han, Micheline Kamber & Jian Pei
Data Mining: Concepts and Techniques, 3rd ed.

Rhodes College



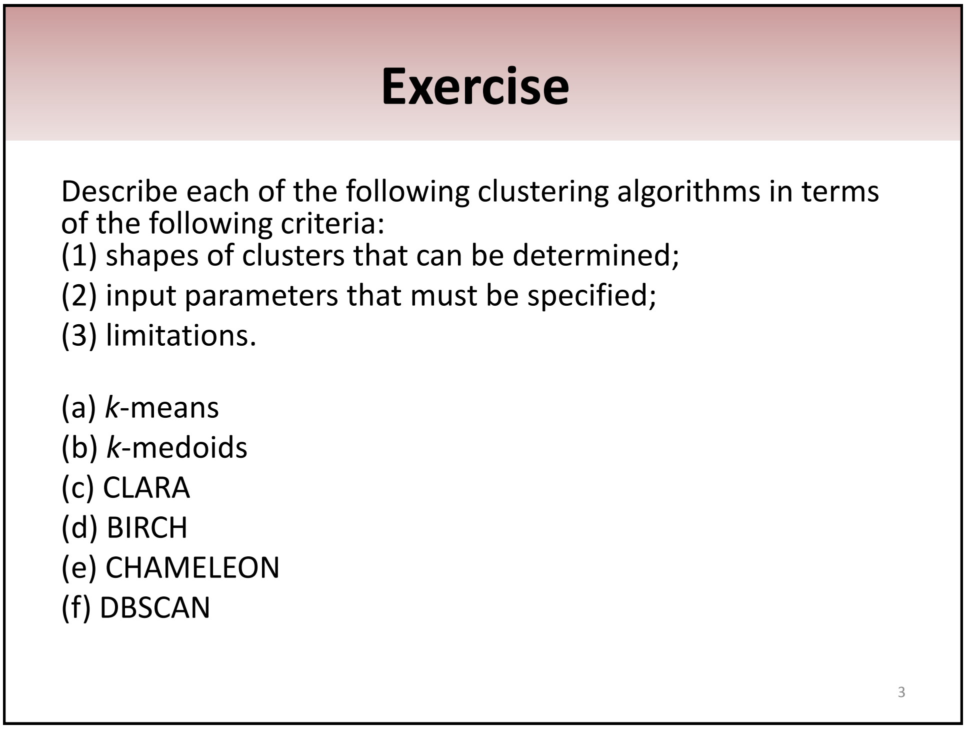
Announcements/Reminders

- **Bring laptop on Wed/Thurs this week**
- Assignment 5 has been assigned
  - Due Wed. Oct. 24th/Thurs. Oct. 25th at the beginning of class

- Extra Credit Opportunity (2 points):
  - Attend Hal Roberts talk on "Network Propaganda"
  - **Wed, October 24th at 6p.m. in McNeill Hall**
  - Turn in a 1 paragraph summary of what you learned and what you found interesting about the talk by beginning of class on Mon. Oct. 29th /Tues. Oct. 30th

- Extra Credit Opportunity (2 points)
  - Attend Dr. Stanley Pounds talk on his Biostatistics Research at St. Jude
  - **Thurs. Nov. 1st at 4pm in Spence Wilson Room**
  - Turn in a 1 paragraph summary of what you learned and what you found interesting about the talk by beginning of class on Mon. Nov. 5th /Tues. Nov. 6th

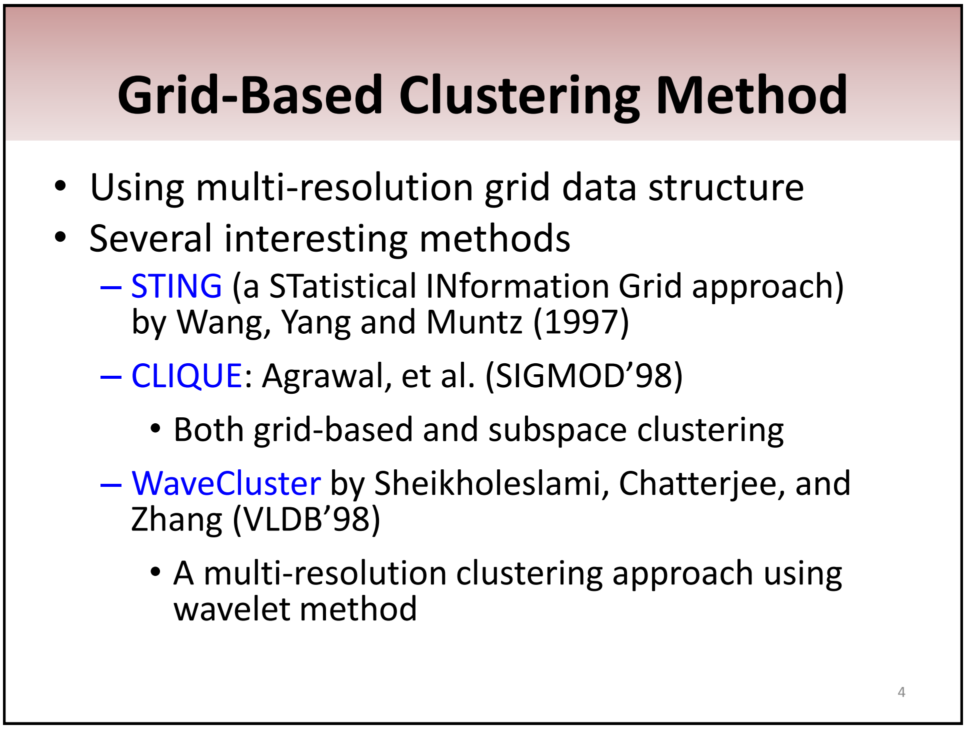
2



Exercise

Describe each of the following clustering algorithms in terms of the following criteria:
(1) shapes of clusters that can be determined;
(2) input parameters that must be specified;
(3) limitations.

(a) *k*-means
(b) *k*-medoids
(c) CLARA
(d) BIRCH
(e) CHAMELEON
(f) DBSCAN

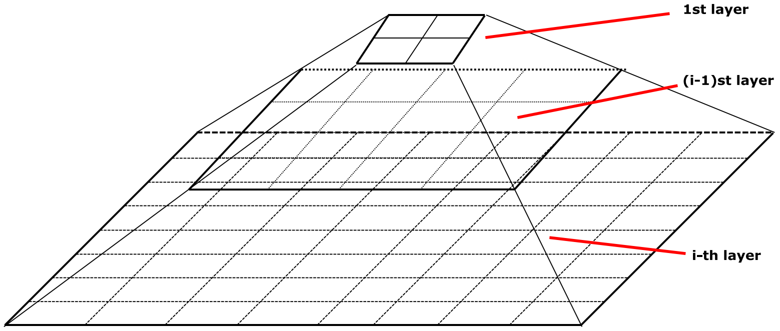


Grid-Based Clustering Method

- Using multi-resolution grid data structure
- Several interesting methods
  - STING (a STatistical INformation Grid approach) by Wang, Yang and Muntz (1997)
  - CLIQUE: Agrawal, et al. (SIGMOD’98)
    - Both grid-based and subspace clustering
  - WaveCluster by Sheikholeslami, Chatterjee, and Zhang (VLDB’98)
    - A multi-resolution clustering approach using wavelet method

STING: A Statistical Information Grid Approach

- Wang, Yang and Muntz (VLDB'97)
- The spatial area is divided into rectangular cells
- There are several levels of cells corresponding to different levels of resolution



The STING Clustering Method

- Each cell at a high level is partitioned into a number of smaller cells in the next lower level
- Statistical info of each cell is calculated and stored beforehand and is used to answer queries
- Parameters of higher level cells can be easily calculated from parameters of lower level cell
  - *count*, *mean*, *s*, *min*, *max*
  - type of distribution—*normal*, *uniform*, etc.
- Use a top-down approach to answer spatial data queries
- Start from a pre-selected layer—typically with a small number of cells
- For each cell in the current level compute the confidence interval

STING Algorithm and Its Analysis

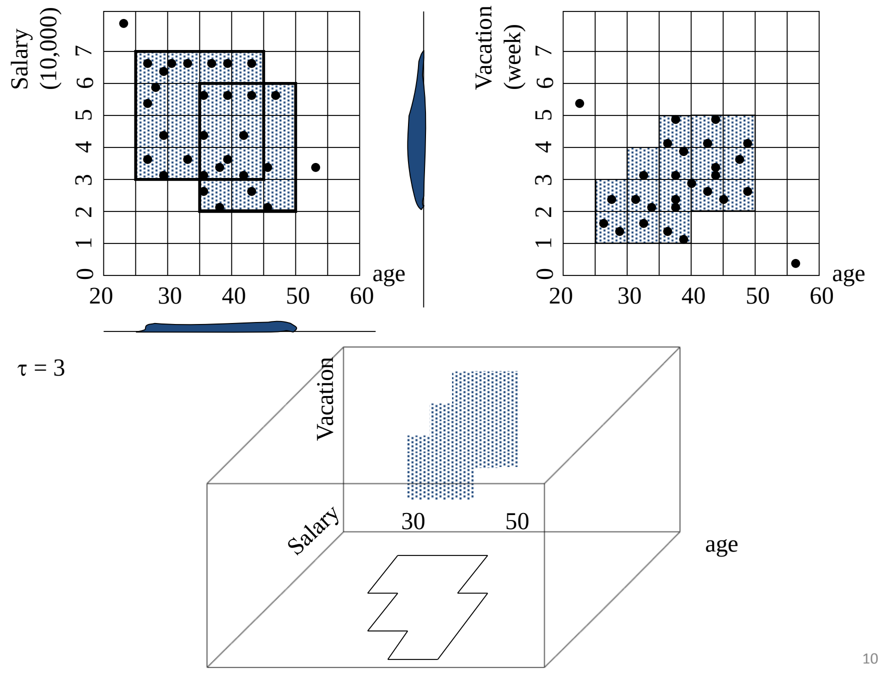
- Remove the irrelevant cells from further consideration
- When finish examining the current layer, proceed to the next lower level
- Repeat this process until the bottom layer is reached
- Advantages:
  - Query-independent, easy to parallelize, incremental update
  - *O(K),* where *K* is the number of grid cells at the lowest level
- Disadvantages:
  - All the cluster boundaries are either horizontal or vertical, and no diagonal boundary is detected

CLIQUE (Clustering In QUEst)

- Agrawal, Gehrke, Gunopulos, Raghavan (SIGMOD'98)
- Automatically identifying subspaces of a high dimensional data space that allow better clustering than original space
- CLIQUE can be considered as both density-based and grid-based
  - It partitions each dimension into the same number of equal length interval
  - It partitions an m-dimensional data space into non-overlapping rectangular units
  - A unit is dense if the fraction of total data points contained in the unit exceeds the input model parameter
  - A cluster is a maximal set of connected dense units within a subspace

CLIQUE: The Major Steps

- 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 description for the clusters
  - Determine maximal regions that cover a cluster of connected dense units for each cluster
  - Determination of minimal cover for each cluster



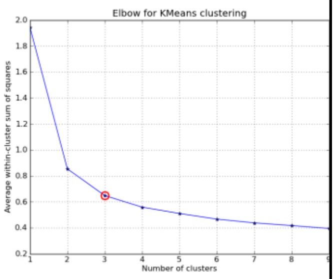
τ = 3

Strength and Weakness of *CLIQUE*

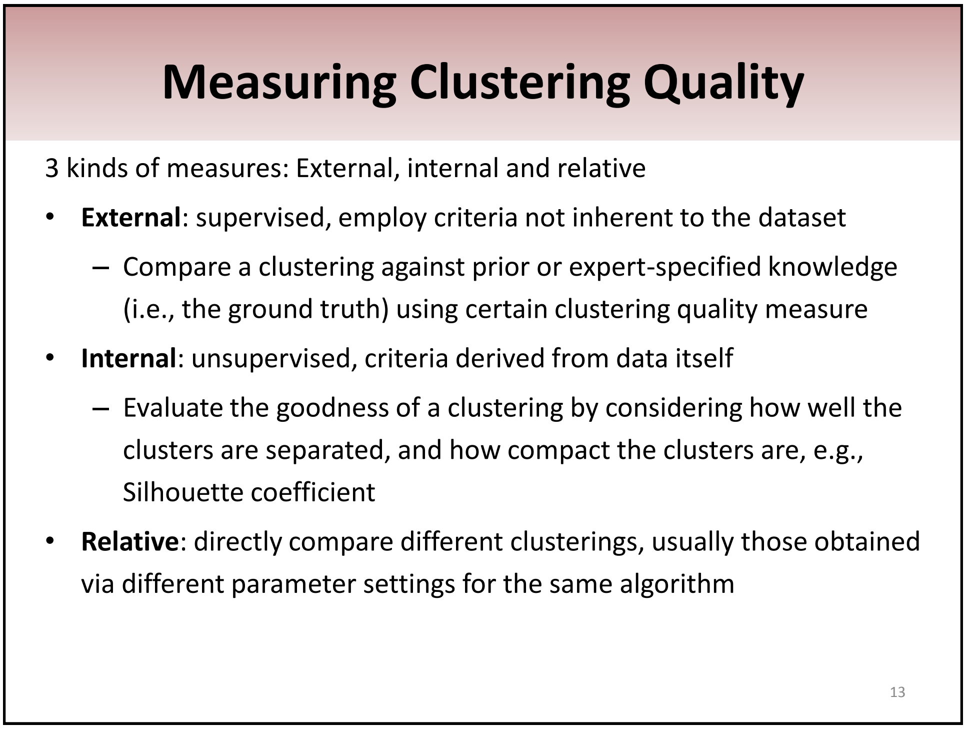
- Strength
  - *automatically* finds subspaces of the highest dimensionality such that high density clusters exist in those subspaces
  - *insensitive* to the order of records in input and does not presume some canonical data distribution
  - scales *linearly* with the size of input and has good scalability as the number of dimensions in the data increases
- Weakness
  - The accuracy of the clustering result may be degraded at the expense of simplicity of the method

Determine the Number of Clusters

- **Empirical method**
  - # of clusters: k ≈√n/2 for a dataset of n points, (e.g., n = 200, k = 10)
- **Elbow method:** Use the turning point in the curve of sum of within cluster variance w.r.t the # of clusters



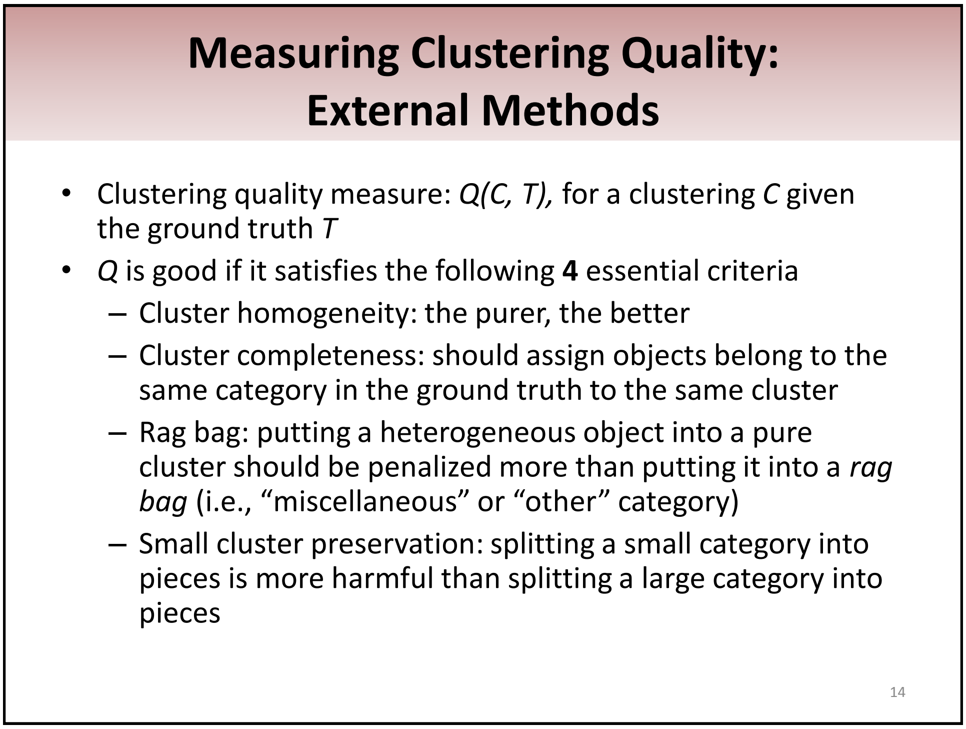
- **Cross validation method**
  - Divide a given data set into *m* parts
  - Use *m* – 1 parts to obtain a clustering model
  - Use the remaining part to test the quality of the clustering
    - E.g., For each point in the test set, find the closest centroid, and use the sum of squared distance between 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



Measuring Clustering Quality

3 kinds of measures: External, internal and relative

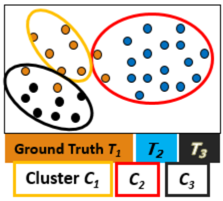
- **External**: supervised, employ criteria not inherent to the dataset
  - Compare a clustering against prior or expert-specified knowledge (i.e., the ground truth) using certain clustering quality measure
- **Internal**: unsupervised, criteria derived from data itself
  - Evaluate the goodness of a 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

- Clustering quality measure: *Q(C, T),* for a clustering *C* given the ground truth *T*
- *Q* is good if it satisfies the following **4** essential criteria
  - Cluster homogeneity: the purer, the better
  - Cluster completeness: should assign objects belong to the same category in the ground truth to the same cluster
  - Rag bag: 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

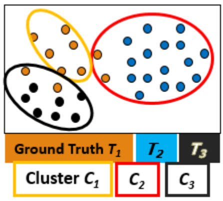
Some Commonly Used External Measures



- Matching-based measures
  - Purity, maximum matching, F-measure
- Entropy-Based Measures
  - Conditional entropy, normalized mutual information (NMI), variation of information
- Pair-wise 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

15

Matching-Based Measures (I): Purity

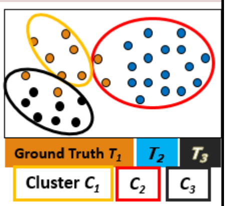


Purity: Quantifies the extent that cluster C_i contains points only from one (ground truth) partition: $purity_i = \frac{1}{n_i} \max_{j=1}^{k}\{n_{ij}\}$

- 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: $purity_1$ = 30/50; $purity_2$ = 20/25; $purity_3$ = 25/25; *purity* = (30 + 20 + 25)/100 = 0.75

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

Matching-Based Measures (II): F-measure



- **Precision**: The fraction of points in C_i from the majority partition (i.e., the same as purity), where j_i is the partition that contains the maximum # of points from C_i
  - Ex. For the green table $prec_i = \frac{1}{n_i}\max_{j=1}^{k}\{n_{ij}\} = \frac{n_{ij_i}}{n_i}$
    - $prec_1$ = 30/50; $prec_2$ = 20/25; $prec_3$ = 25/25
- **Recall**: The fraction of points in partition shared in common with cluster C_i, where $recall_i = \frac{n_{ij_i}}{|T_{j_i}|} = \frac{n_{ij_i}}{m_{j_i}}$
  - Ex. For the green table
    - $recall_1$ = 30/35; $recall_2$ = 20/40; $recall_3$ = 25/25

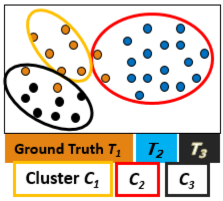
C\T	T_1	T_2	T_3	Sum
C_1	0	20	30	50
C_2	0	20	5	25
C_3	25	0	0	25
m_j	25	40	35	100

- **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

Entropy-Based Measure (I): Conditional Entropy

- Entropy of clustering C: $H(\mathcal{C}) = -\sum_{i=1}^{r} p_{C_i}\log p_{C_i}$ $\quad p_{C_i} = \frac{n_i}{n}$ *the prob. of cluster C*
- Entropy of partitioning T: $H(\mathcal{T}) = -\sum_{j=1}^{k} p_{T_j}\log p_{T_j}$
- Entropy of T w.r.t. cluster C_i: $H(\mathcal{T}|C_i) = -\sum_{j=1}^{k}(\frac{n_{ij}}{n_i})\log(\frac{n_{ij}}{n_i})$
- Conditional entropy of T w.r.t. clustering C:

$$H(\mathcal{T}|\mathcal{C}) = -\sum_{i=1}^{r}(\frac{n_i}{n})H(\mathcal{T}|C_i) = -\sum_{i=1}^{r}\sum_{j=1}^{k} p_{ij}\log(\frac{p_{ij}}{p_{C_i}})$$

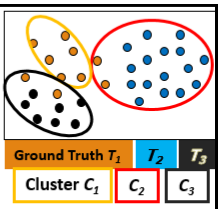


  - The more a cluster's members are split into different partitions, the higher the conditional entropy
  - For a perfect clustering, the conditional entropy value is 0, where the worst possible conditional entropy value is log k

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

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

Entropy-Based Measure (II): Normalized mutual information (NMI)



- Mutual information: quantify the amount of shared info between the clustering C and partitioning T:

$$I(\mathcal{C},\mathcal{T}) = -\sum_{i=1}^{r}\sum_{j=1}^{k} p_{ij} \log\left(\frac{p_{ij}}{p_{C_i} \cdot p_{T_j}}\right)$$

It measures the dependency between the observed joint probability p_{ij} of C and T, and the expected joint probability p_{Ci} * p_{Tj} under the independence assumption

When C and T are independent, $p_{ij} = p_{Ci}$ * p_{Tj}, I(C, T) = 0. However, there is no upper bound on the mutual information

- Normalized mutual information (NMI)

$$NMI(\mathcal{C},\mathcal{T}) = \sqrt{\frac{I(\mathcal{C},\mathcal{T})}{H(\mathcal{C})} \cdot \frac{I(\mathcal{C},\mathcal{T})}{H(\mathcal{T})}} = \frac{I(\mathcal{C},\mathcal{T})}{\sqrt{H(\mathcal{C}) \cdot H(\mathcal{T})}}$$

Value range of NMI: [0,1]. Value close to 1 indicates a good clustering

19

Pairwise Measures: Four Possibilities for Truth Assignment

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

- **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

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

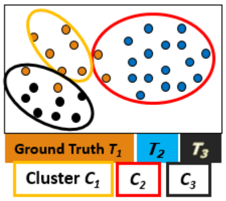
$$FP = \sum_{i=1}^{r}\binom{n_i}{2} - TP \qquad TN = N - (TP + FN + FP) = \frac{1}{2}\left(n^2 - \sum_{i=1}^{r} n_i^{2} - \sum_{j=1}^{k} m_j^{2} + \sum_{i=1}^{r}\sum_{j=1}^{k} n_{ij}^{2}\right)$$

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



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

Summary

- Cluster analysis groups objects based on their similarity and has wide applications
- Measure of similarity can be computed for various types of data
- Clustering algorithms can be categorized into partitioning methods, hierarchical methods, density-based methods, grid-based methods, and model-based methods
- K-means and K-medoids algorithms are popular partitioning-based clustering algorithms
- Birch and Chameleon are interesting hierarchical clustering algorithms, and there are also probabilistic hierarchical clustering algorithms
- DBSCAN, OPTICS, and DENCLU are interesting density-based algorithms
- STING and CLIQUE are grid-based methods, where CLIQUE is also a subspace clustering algorithm
- Quality of clustering results can be evaluated in various ways