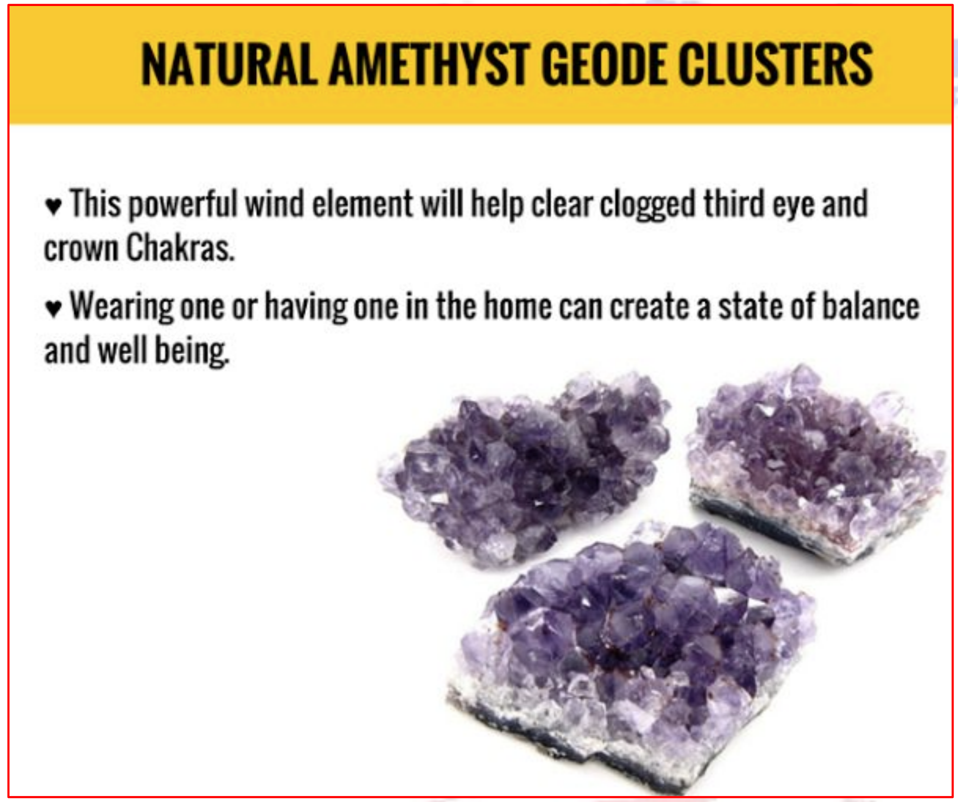


Chapter 10. Cluster Analysis: Evaluation

P. Flynn, M. Jiang

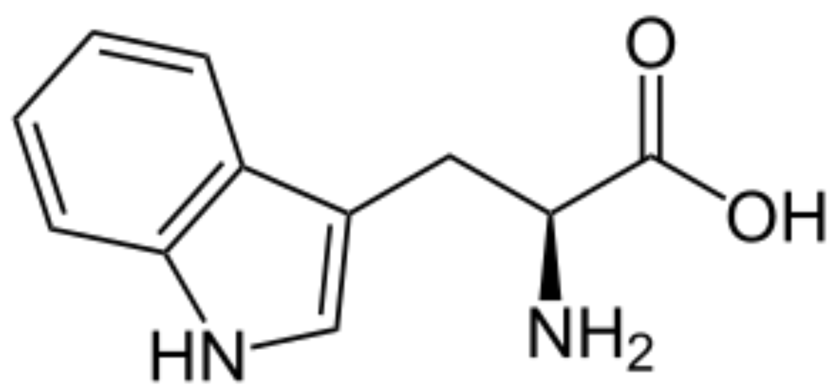
CSE 40647/60647 Data Science Fall 2017

Introduction to Data Mining

Data source: United Nations

HW4 statistics

- Min 56
- Max 100
- Mean 89.6
- Median 96
- Mode 98
- Standard deviation 13.52



Cluster Analysis

- Cluster Analysis: An Introduction
- Partitioning Methods
- Density-based Methods
- **Evaluation of Clustering**

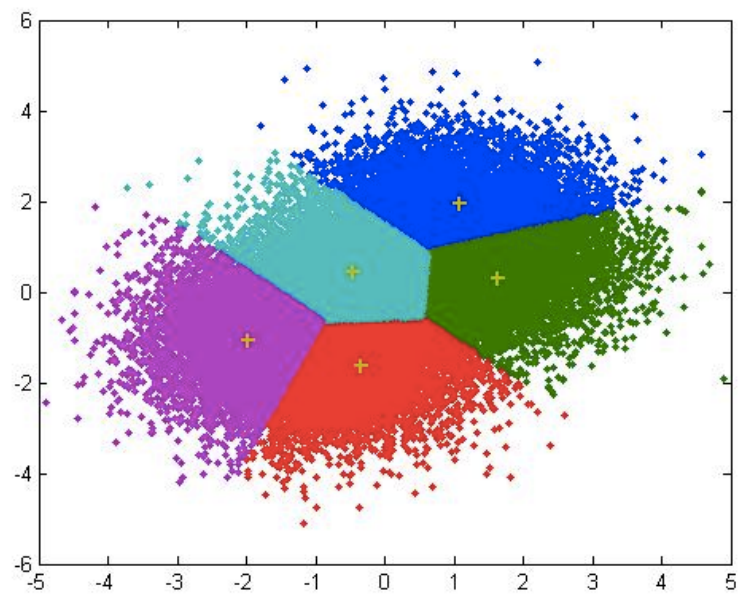


Figure from: "Efficient K-Means Clustering using JIT", MathWorks site

Clustering Methodology

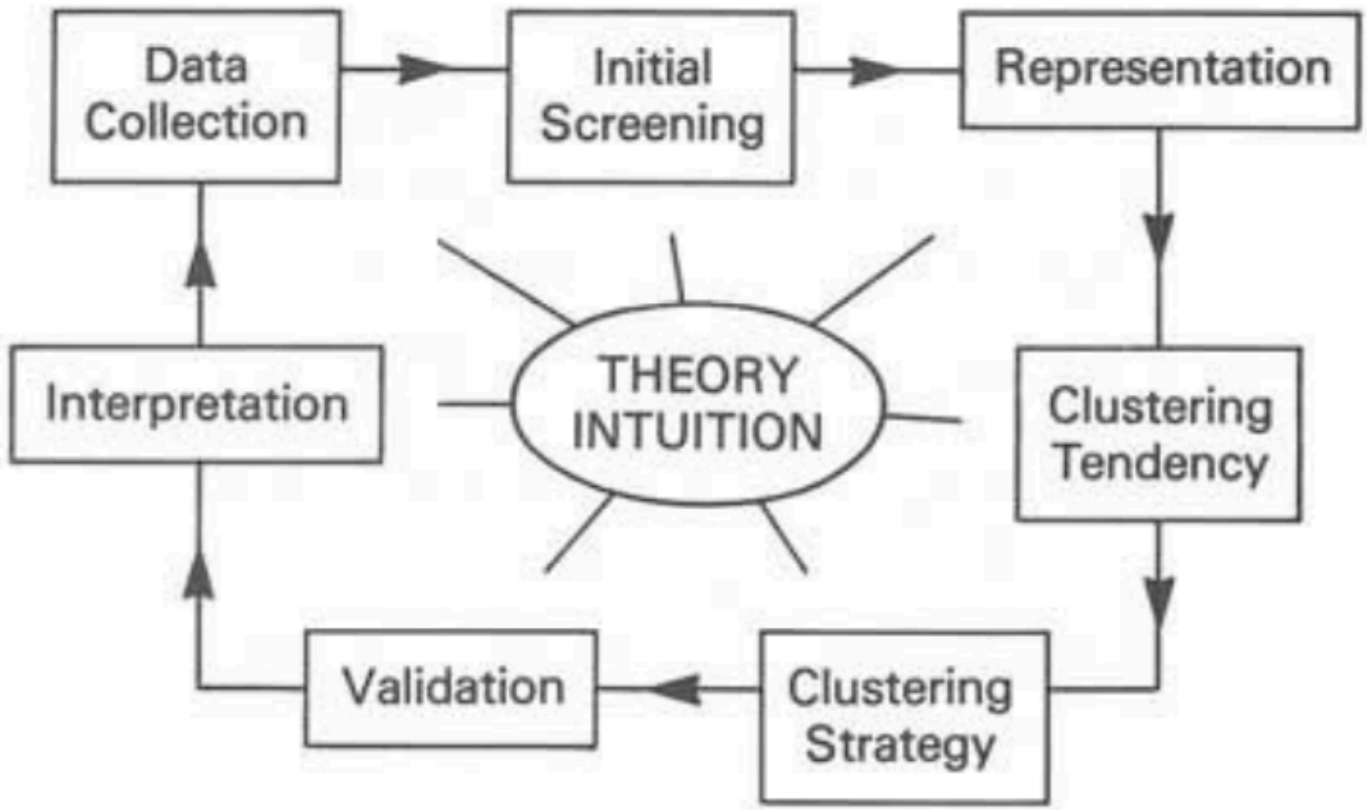
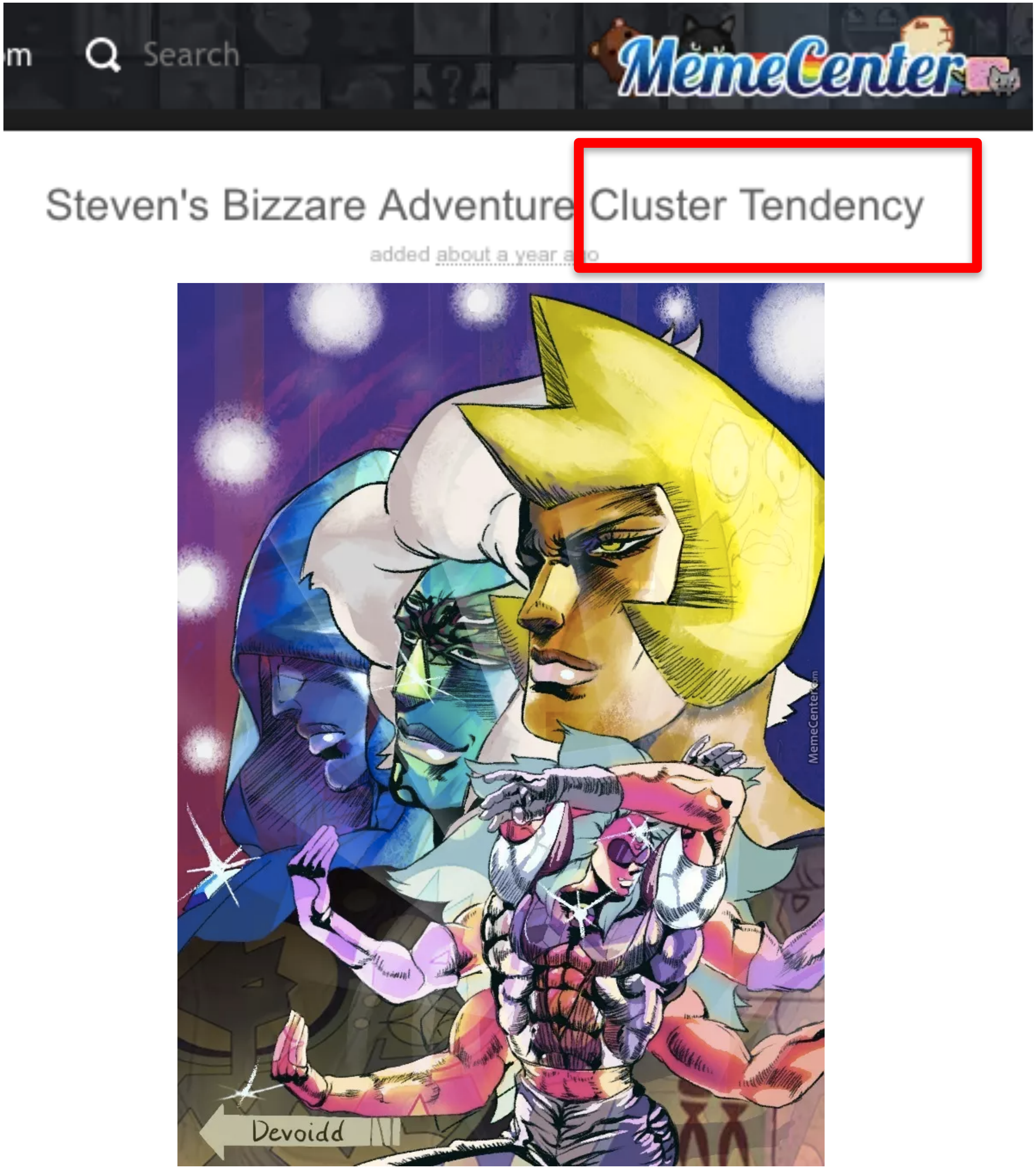


Figure 3.35 Clustering methodology.

From Jain and Dubes, *Algorithms for Clustering Data*, 1988.

Clustering Validation and Assessment

- Cluster Validation
  - Evaluating the goodness of a given clustering
  - Does it reflect structure in the data?
  - "Quantitative and objective"
- Clustering stability
  - Sensitivity of the clustering result to tunable parameters, e.g., # of clusters
- Cluster tendency
  - Are there clusters in this data?


m
Search
MemeCenter
Steven's Bizzare Adventure Cluster Tendency
added about a year ago
MemeCenter.com
Devoidd

Cluster Tendency

- “Are there clusters in this data?”
- “Can I **reject** a **hypothesis** that the data are all generated from a random process that does not have cluster structure?” => tests of hypotheses of randomness
- “Spatial statistics” and threshold values for tests
- Book: Hopkins statistic

Hopkins statistic

- Output value: 1 means highly clustered, 0 means uniformly distributed
- X is the data set with N points in d dimensions
- Consider a sample of size m << n with members x_i, i=1…m
- Generate a set Y of m points uniformly randomly distributed over the same spatial window as X
- Let u_i = distance between y_i and its nearest neighbor in X
- Let w_i = distance between x_i and its nearest neighbor in X
- Then $$H = \frac{\sum_{i=1}^{m} u_i^d}{\sum_{i=1}^{m} u_i^d + \sum_{i=1}^{m} w_i^d}$$

Hopkins statistic ctd

- If H > 0.75, clustering tendency exists "at a 90% confidence level"
- If H is approx 0.5 then the data are probably uniformly distributed

How many clusters?

- Ad hoc: for *n* points, guess $\sqrt{\frac{n}{2}}$ for the number of clusters (??)
- A bit unsatisfying…

- Recall clustering squared error for k clusters:

$$\min \; E_k^2 = \sum_{i=1}^{k} e_i^2 \qquad e_i^2 = \sum_{j=1}^{n_i} \left\| \vec{x}_j^{(i)} - \vec{m}^{(i)} \right\|^2$$

How many clusters? ctd

- "elbow" method: plot a cluster validity index like clustering error versus number of clusters k and choose the number k* that shows a "corner" in the curve
  - Tension between "more clusters -> smaller SSE" **and** "more clusters -> flat SSE"
  - Don't look for the minimum of the curve, because it is at k* = N

Cluster Validation

- Want
  - Yes/No answer to “is this a good clustering?”… or
  - A “score” for how good it is
- No commonly recognized best suitable measure in practice
- **Three criteria**
  - **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, squared-error
  - **Relative**: Directly compare different clusterings, usually those obtained via different parameter settings for the same algorithm

Cluster validity: Internal measures

- Squared error!

$$\min \ E_k^2 = \sum_{i=1}^{k} e_i^2 \qquad e_i^2 = \sum_{j=1}^{n_i} \left\| \vec{x}_j^{(i)} - \vec{m}^{(i)} \right\|^2$$

- Generalizations, e.g. Dunn index

$$DI_m = \frac{\min_{i,j} \delta(C_i, C_j)}{\max_{1 \le k \le m} \Delta_k}$$

Min dist between clusters

Max size of any cluster

Internal measures ctd.

- Silhouette coefficient (book)
- ’o’ is an item in the data set (belons to a cluster)
- a(o) = avg. dist. Between o and all other items in o’s cluster
- b(o) = minimum avg distance between o and the other clusters
- Then

$$s(0) = \frac{b(o) - a(o)}{\max\{a(o), b(o)\}}$$

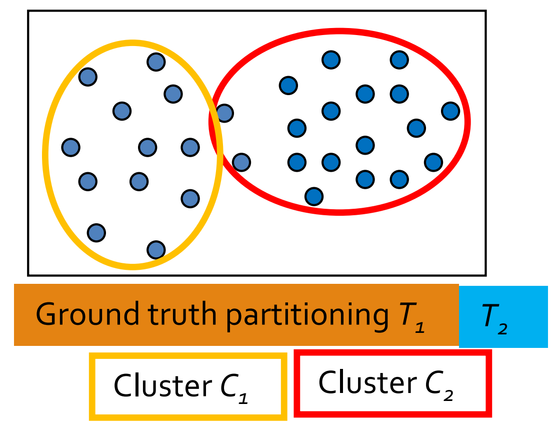
- Bigger is better (-1 <= s(o) <= 1)
- Can average over all data to get avg coef for data set

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

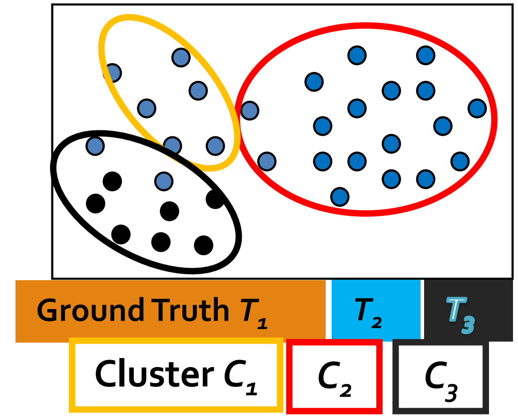
Commonly Used External Measures

- **Matching-based measures**
  – Purity, maximum matching, F-measure
- **Pairwise measures**
  – Four possibilities: True positive (TP), FN, FP, TN
  – Jaccard coefficient, Rand statistic, Fowlkes-Mallow measure
- **Entropy-Based Measures**
  – Conditional entropy
  – Normalized mutual information (NMI)
  – Variation of information



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
  - Maximum weight matching: Pair-wise
  - Ex2. (green) *match* = *purity* = 0.75; (orange) *match* = 0.65 > 0.6

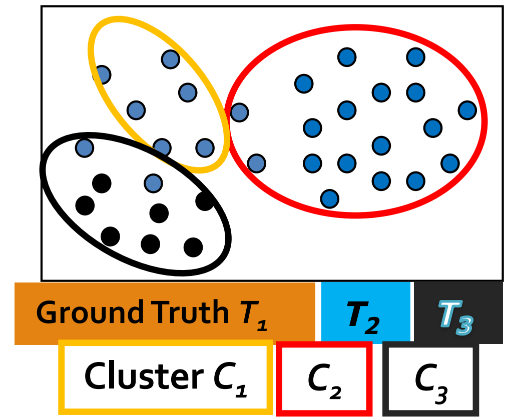


C\T	T_1	T_2	T_3	Sum
C_1	0	20	30	50
C_2	0	20	5	25
C_3	25	0	0	25
m_j	25	40	35	100

C\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
  - Ex. For the green table
    - $prec_1 = 30/50$; $prec_2 = 20/25$; $prec_3 = 25/25$
- **Recall**: The fraction of point in partition shared in common with cluster C_i, where $m_{j_i} = |T_{j_i}|$
  - 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

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

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

Pairwise Measures: Four Possibilities for Truth Assignment

- **Four possibilities** based on the agreement between cluster label and partition label
  - *TP*: true positive—Two points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ 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 $\boldsymbol{x}_i$

  - *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_total*
  - Symmetric; perfect clustering: Rand = 1
- Fowlkes-Mallow Measure:
  - Geometric mean of precision and recall

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

$C\backslash T$	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

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

Entropy-Based Measures (I): Conditional Entropy

- **Entropy of clustering** C: $H(\mathcal{C}) = -\sum_{i=1}^{r} p_{C_i} \log p_{C_i}$ $\quad p_{C_i} = \frac{n_i}{n}$ (i.e., the probability of cluster C_i)
- **Entropy of partitioning** T: $H(\mathcal{T}) = -\sum_{j=1}^{k} p_{T_j} \log p_{T_j}$
- **Entropy of** T **with respect to 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 **with respect to 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 Measures (II): Normalized Mutual Information (NMI)

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

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

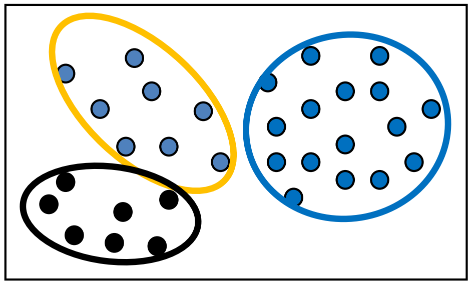
  - Measures the dependency between the observed joint probability p_{ij} of *C* and *T*, and the expected joint probability $p_{Ci} \cdot p_{Tj}$ under the independence assumption
  - When *C* and *T* are independent, $p_{ij} = p_{Ci} \cdot p_{Tj}$, $I(C, T) = 0$. However, there is no upper bound on the 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

Internal Measures: BetaCV Measure

- A trade-off in maximizing intra-cluster compactness and inter-cluster separation
- Given a clustering $C = \{C_1, \ldots, C_k\}$ 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: $W_{out} = \frac{1}{2}\sum_{i=1}^{k} W(C_i, \overline{C_i}) = \sum_{i=1}^{k-1}\sum_{j>i} W(C_i, C_j)$
  - 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**:
  - The ratio of the mean intra-cluster distance to the mean inter-cluster distance
  - The smaller, the better the clustering

$$BetaCV = \frac{W_{in} / N_{in}}{W_{out} / N_{out}}$$

Summary

- Cluster Analysis: An Introduction
- Partitioning Methods
- Density-based Methods
- Evaluation of Clustering

References: (IV) Evaluation of Clustering

- M. J. Zaki and W. Meira, Jr.. Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press, 2014
- L. Hubert and P. Arabie. Comparing Partitions. Journal of Classification, 2:193–218, 1985
- A. K. Jain and R. C. Dubes. Algorithms for Clustering Data. Printice Hall, 1988
- M. Halkidi, Y. Batistakis, and M. Vazirgiannis. On Clustering Validation Techniques. Journal of Intelligent Info. Systems, 17(2-3):107–145, 2001
- J. Han, M. Kamber, and J. Pei. Data Mining: Concepts and Techniques. Morgan Kaufmann, 3rd ed. , 2011
- H. Xiong and Z. Li. Clustering Validation Measures. in (Chapter 23) C. Aggarwal and C. K. Reddy (eds.), Data Clustering: Algorithms and Applications. CRC Press, 2014