Lecture IV – Hierarchical clustering. Comparing clusterings

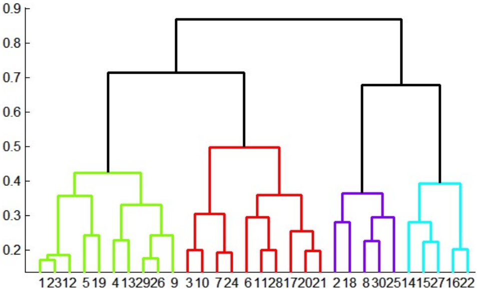
Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

STAT/BIOST 527

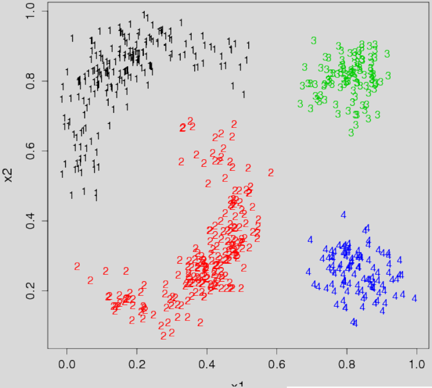
Hierarchical Methods of Clustering

- **Agglomerative** (bottom up):
  - Initially, each point is a cluster
  - Repeatedly combine the two "nearest" clusters into one
- **Divisive** (top down):
  - Start with one cluster and recursively split it

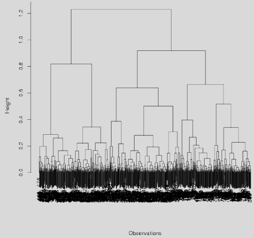


What is hierarchical clustering?

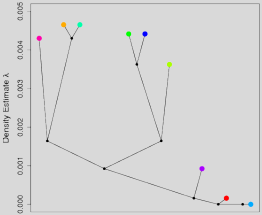
- Clusters have cluster structure
- Represented by
  - **Dendrogram**
  - **Cluster Tree** (only from KDE)



Dendrogram



Cluster Tree

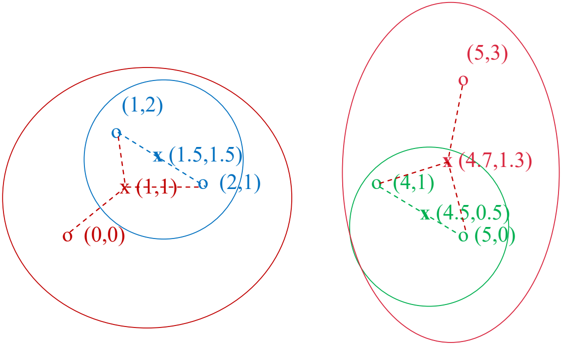


Hierarchical clustering – Overview

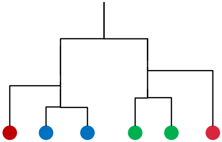
(Dendrograms)

- **Agglomerative** (bottom up)
  - **Single linkage**
    - based on Minimum Spanning Tree
    - $\mathcal{O}(n^2 \log n)$
    - sensitive to outliers
  - Heuristics – average linkage
  - **Agglomerative K-means**
    - Loss $\mathcal{L}(\Delta_K) = 0$ for $K = n$
    - When $K \leftarrow K - 1$ (two clusters merged), $\mathcal{L}$ increases
    - For $K = n, n-1, \ldots 2$, iteratively merge the 2 clusters that minimize increase of $\mathcal{L}$
    - $\mathcal{O}(n^3)$ – too expensive for big data
- **Divisive** (bottom down)
  - Recursively split $\mathcal{D}$ into $K = 2$ clusters
  - almost any clustering algorithm (e.g. K-means, min diameter)
  - notable example Spectral clustering (later)
  - Advantages
    - most important splits are first
    - can stop after only a few splits

Example: Hierarchical clustering



Data:
o … data point
x … centroid



Dendrogram

Cluster tree

- λ-**tree** Defined by the level sets of the KDE
- α-**tree** Defined by the number of points in r-ball around x_i
  - i.e. by level sets of the nearest neighbor density estimator
  - more robust [Yen-Chi Chen "Generalized cluster tree and singular measures", 2019]

Requirements for a distance

Depend on the application

- Applies to any two partitions of the same data set
- Makes no assumptions about how the clusterings are obtained
- Values of the distance between two pairs of clusterings comparable under the weakest possible assumptions
- Metric (triangle inequality) desirable
- **understandable, interpretable**

The confusion matrix

- Let $\Delta = \{C_{1:K}\}, \; \Delta' = \{C'_{1:K'}\}$
- Define $n_k = |C_k|, \; n'_{k'} = |C'_{k'}|$
- $m_{kk'} \; = |C_k \cap C'_{k'}|, \; k = 1:K, k' = 1:K'$
- note: $\sum_k m_{kk'} = n'_{k'}, \; \sum_{k'} m_{kk'} = n_k, \; \sum_{k,k'} m_{kk'} = n$
- The **confusion matrix** $M \in \mathbb{R}^{K\times K'}$ is

$$M \;=\; [m_{kk'}]_{k=1:K}^{k'=1:K'}$$

- all distances and comparison criteria are based on M
- the **normalized confusion matrix** $P = M/n$

$$p_{kk'} \;=\; \frac{m_{kk'}}{n}$$

- The **normalized cluster sizes** $p_k = n_k/n, \; p'_{k'} = n'_{k'}/n$ are the **marginals** of P

$$p_k \;=\; \sum_{k'} p_{kk'} \quad p_{k'} \;=\; \sum_k p_{kk'}$$

Matrix Representations

- matrix reprentations for Δ
  - unnormalized (redundant) representation

$$\tilde{X}_{ik} = \begin{cases} 1 & i \in C_k \\ 0 & i \notin C_k \end{cases} \quad \text{for } i = 1:n, k = 1:K$$

  - normalized (redundant) representation

$$X_{ik} = \begin{cases} 1/\sqrt{|C_k|} & i \in C_k \\ 0 & i \notin C_k \end{cases} \quad \text{for } i = 1:n, k = 1:K$$

    therefore $X_k^T X_{k'} = \delta(k, k')$, X orthogonal matrix
    X_k = column k of X
  - normalized non-redundant reprentation
    - X_K is determined by $X_{1:K-1}$
    - hence we can use $Y \in \mathbb{R}^{n \times (K-1)}$ orthogonal representation
    - intuition: Y represents a subspace (is an orthogonal basis)
    - K centers in $\mathbb{R}^d$, $d \geq K$ determine a $K-1$ dimesional subspace plus a translation

The Misclassification Error (ME) distance

- Define the **Misclassification Error (ME)** distance d_{ME}

$$d_{ME} = 1 - \max_{\pi} \sum_{k=1}^{K} p_{k,\pi(k)} \quad \pi \in \{\text{all } K\text{−permutations}\},\ K \leq K' \text{w.l.o.g}$$

- Interpretation: treat the clusterings as classifications, then minimize the classification error over all possible label matchings
- Or: nd_{ME} is the Hamming distance between the vectors of labels, minimized over all possible label matchings
- can be computed in polynomial time by **Max bipartite matching** algorithm (also known as Hungarian algorithm)
- Is a metric: symmetric, ≥ 0, triangle inequality

$$d_{ME}(\Delta_1, \Delta_2) + d_{ME}(\Delta_1, \Delta_3) \geq d_{ME}(\Delta_2, \Delta_3)$$

- easy to understand (very popular in computer science)
- $d_{ME} \leq 1 - 1/K$
- bad: if clusterings not similar, or K large, d_{ME} is coarse/indiscriminative
- recommended: for small K

The Variation of Information (VI) distance
Clusterings as random variables

- Imagine points in $\mathcal{D}$ are picked randomly, with equal probabilities
- Then $k(i), k'(j)$ are random variables
  with $Pr[k] = p_k$, $Pr[k, k'] = p_{kk'}$

STAT/BIOST 527

Incursion in information theory

- **Entropy** of a random variable/clustering $H_\Delta = -\sum_k p_k \ln p_k$
- $0 \leq H_\Delta \leq \ln K$
- Measures uncertainty in a distribution (amount of randomness)
- **Joint entropy** of two clusterings

$$H_{\Delta,\Delta'} = -\sum_{k,k'} p_{kk'} \ln p_{kk'}$$

- $H_{\Delta',\Delta} \leq H_\Delta + H_{\Delta'}$ with equality when the two random variables are independent
- **Conditional entropy** of Δ' given Δ

$$H_{\Delta'|\Delta} = -\sum_k p_k \sum_{k'} \frac{p_{kk'}}{p_k} \ln \frac{p_{kk'}}{p_k}$$

- Measures the expected uncertainty about k' when k is known
- $H_{\Delta'|\Delta} \leq H_{\Delta'}$ with equality when the two random variables are independent
- **Mutual information** between two clusterings (or random variables)

$$\begin{aligned} I_{\Delta,\Delta} &= H_\Delta + H_{\Delta'} - H_{\Delta',\Delta} \\ &= H_{\Delta'} - H_{\Delta'|\Delta} \end{aligned}$$

Marina Meila:

Incursion in information theory (2)

- Measures the amount of information of one r.v. about the other
- $I_{\Delta,\Delta} \geq 0$, symmetric. Equality iff r.v.'s independent

STAT/BIOST 527

The VI distance

- Define the **Variation of Information (VI)** distance

$$\begin{aligned} d_{VI}(\Delta, \Delta') &= H_\Delta + H_{\Delta'} - 2I_{\Delta',\Delta} \\ &= H_{\Delta|\Delta'} + H_{\Delta'|\Delta} \end{aligned}$$

- Interpretation: d_{VI} is the sum of information gained and information lost when labels are switched from $k()$ to $k'()$
- d_{VI} symmetric, ≥ 0
- d_{VI} obeys triangle inequality (is a metric)

Other properties

- Upper bound
  $d_{VI} \leq 2 \ln K_{max}$ if $K, K' \leq K_{max} \leq \sqrt{n}$
  (asymptotically attained)
- $d_{VI} \leq \ln n$ over all partitions (attained)
- Unbounded! and grows fast for small K

Marina Meila:

Other criteria and desirable properties

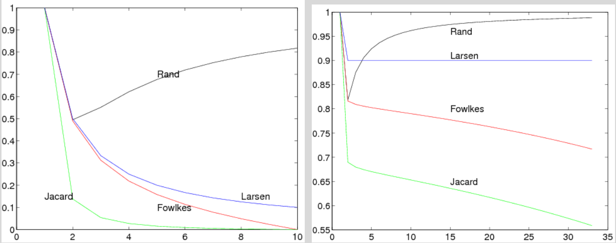
- Comparing clustering by **indices of similarity** $i(\Delta, \Delta')$
  - from statistics (Rand, adjusted Rand, Jaccard, Fowlkes-Mallows ...)
  - Normalized Mutual Information
  - range=[0,1], with $i(\Delta, \Delta') = 1$ for $\Delta = \Delta'$
  - the properties of these indices not so good
  - any index can be transformed into a "distance" by $d(\Delta, \Delta') = 1 - i(\Delta, \Delta')$
- Other desirable properties of indices and distances between clusterings
  - n-invariance
  - locality
  - convex additivity

Rand, Jaccard and Fowlkes-Mallows

- Define N_{11} = # pairs which are together in both clusterings, N_{12} = # pairs together in Δ, separated in Δ', N_{21} (conversely), N_{22} =#number pairs separated in both clusterings
- Rand index = $\frac{N_{11}+N_{22}}{\#\text{pairs}}$
- Jaccard index = $\frac{N_{11}}{\#\text{pairs}}$
- Fowlkes-Mallows = Precision$\times$ Recall

- all vary strongly with K. Thereforek, **Adjusted** indices used mostly

$$adj(i) = \frac{i - \bar{i}}{\max(i) - \bar{i}}$$



Normalized Mutual Information (NMI)

$$i_{NMI}(\Delta, \Delta') = \frac{I_{\Delta', \Delta}}{H_\Delta + H_{\Delta'}} \quad (1)$$

- Takes values between [0,1]
- No probabilistic interpretaion
- Variant $\frac{I_{\Delta', \Delta}}{H_{\Delta, \Delta'}}$