BIOSTAT 527

4/17/23

Lecture 7

NP and Hierarchical
clustering

L IV.1 - Hierarchical
LV SVM

(lecture given on UW tablet)

Lecture IV – Non-parametric clustering

Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

Think of next topics for rest of 527

- LIV.1 Hierarchical
- LV SVM Kernel
- HW2 Wednesday

1. Paradigms for clustering ✓
2. Methods based on non-parametric density estimation ✓

   Cluster tree (level sets)

   DB scan
3. Model-based: Dirichlet process mixture models

Reading AoNPS Ch.: –, HTF Ch.: 14.3 Murphy Ch.: 11.[1], 11.2.1-3, 11.3, Ch 25

The Nugent-Stuetzle algorithm

Algorithm Nugent-Stuetzle

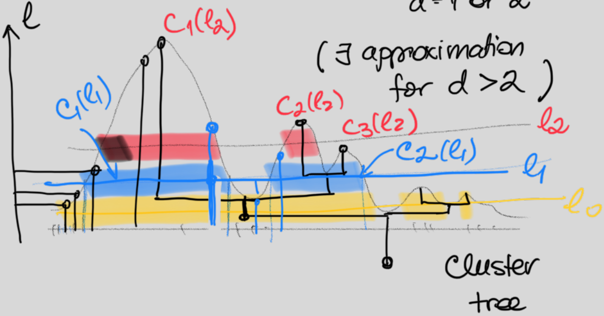
Input Data $\mathcal{D} = \{x_i\}_{i=1:n}$, kernel $K(z)$

1. Compute KDE $f(x)$ for chosen h → $f(x^i)$ $i=1:n$ sort by $f(x^i)$
2. for levels $0 < l_1 < l_2 < \ldots < l_r < \ldots < l_R \geq \sup_x f(x)$
   1. find **level set** $L_r = \{x \mid f(x) \geq l_r\}$ of f
   2. if L_r **disconnected** then each connected component is a cluster → $(C_{r,1}, C_{r,2}, \ldots C_{r,K_r})$

Output clusters $\{(C_{r,1}, C_{r,2}, \ldots C_{r,K_r})\}_{r=1:R}$

tractable only in $d=1$ or 2

($\exists$ approximation for $d>2$)



Remarks

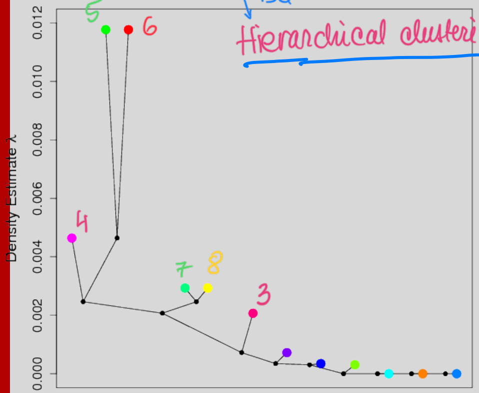
- every cluster $C_{r,k} \subseteq$ some cluster $C_{r-1,k'}$
- therefore output is hierarchical clustering
- some levels can be pruned (if no change, i.e. $K_r = K_{r-1}$)
- algorithm can be made recursive, i.e. efficient
- finding level sets of f tractable only for $d = 1, 2$
- for larger d, $L_r = \{x_i \in \mathcal{D} \,|\, f(x_i) \geq l_r\}$
- to find connected components
  - for $i \neq j \in L_r$
    if $f(tx_i + (1-t)x_j) \geq l_r$ for $t \in [0, 1]$
    then $k(i) = k(j)$

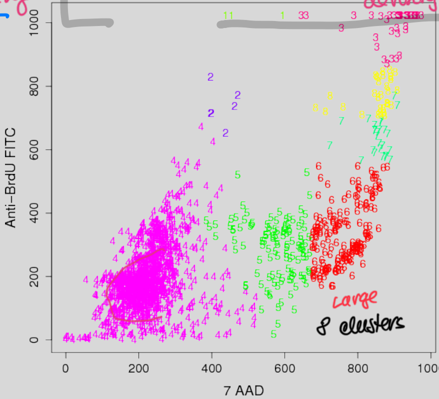
  } heuristic for $d > 2$
- confidence intervals possile by resampling

Cluster tree with 13 leaves (8 clusters, 5 artifacts)

isa Hierarchical clustering

Parenthesis on plotting
· default markers → to see outliers!
· really small dots → to see density!





(from ?)

Rebecca Nugent

Chaudhuri-Dasgupta Algorithm

- Uses k-**nearest neighbor** graphs (filtration)
- Parameters k (nearest neigbhors) and $\alpha \in [1, 2]$
- for $r \geq 0$, $G_r = (V_r, E_r)$ with
  - $x_i \in V_r$ iff **distance to** k**-nn of** $x_i \leq r$
  - $(x_i, x_j) \in E_r$ iff $||x_i - x_j|| \leq \alpha r$

Consistency Theorem For any ϵ (separation parameter) and δ (confidence), $\alpha \in [\sqrt{2}, 2]$ (graph density), if $k = C \log^2(1/\delta) \frac{d \log n}{\epsilon^2}$
for any two clusters C, C' in cluster tree, there exists a level r so that $C \cap \mathcal{D}, C' \cap \mathcal{D}$ are clusters at level r

note $k \sim d \log n$

NOT small

NP anything

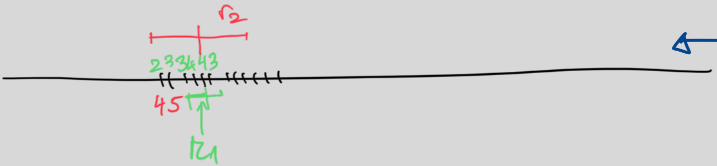
k-NN must have large enough k

GENERAL RULE OF THUMB FOR NP

The K-nn density estimator

The K-nn density estimator

- Let $B_r(x)$ be the (closed) ball of radius r centered at x
- If $|B_r(x^i) \cap \mathcal{D}| = k$ then $\hat{p}(x^i) = \frac{1}{r^n \omega_n} \frac{k}{n}$ is an estimate of the density at x^i
  - $\omega_n = \pi^{n/2} / \Gamma(n/2 + 1)$ is the volume of the unit ball in $\mathbb{R}^n$
  - intuitively, the ball of radius r contains k/n probability mass
  - Note that the density $\hat{p}$ is not required to integrate to 1

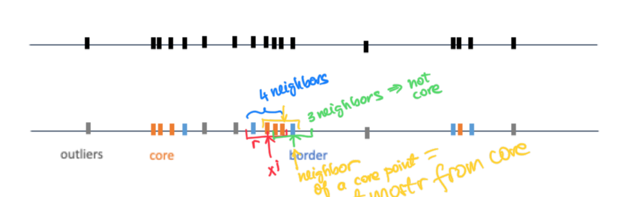


from next page

Marina Meila: L IV NP Clustering

DBScan (Heuristic)

- Introduced with no proof, but widely used. Implicitly based on the K-nn estimator
- Parameters r radius, m minimum number points
- Definitions **core** $Q = \{x^i \in \mathcal{D},$ with $B_r(x^i) \cap \mathcal{D}| \geq m\}$
- **border** $B = \{x^i \in \mathcal{D} \setminus Q,$ so that $x^i \in B_r(x^j), x^j \in Q\}$
- **outliers (noise)** $O = \mathcal{D} \setminus (Q \cup B)$



r = neighborhood radius
m = 4

m ↑ fewer core
more clusters
?

r ↑ ⇒ same m ⇒ more core
fewer clust ?

- Algorithm idea
- Construct directed graph $\mathcal{G}$ with edges (i, j) where $x^i \in Q, j \in B_r(x^i)$
- The graph edges between core points are undirected/symmetric, the other are from core to border
- Clusters are determined by the connected components of the graph restricted to Q.

efficient!

- The border points are assigned to a cluster containing x^j so that $x^i \in B_r(x^j), x^j \in Q$ Note that this assignment is not unique!
- Heuristic algorithm estimates r, m

[Supplement: Chaudhuri-Dasgupta Algorithm]

Consistency Theorem For any ϵ (separation parameter) and δ (confidence), $\alpha \in [\sqrt{2}, 2]$ (graph density), if $k = C \log^2(1/\delta) \frac{d \log n}{\epsilon^2}$
for any two clusters C, C' in cluster tree, there exists a level r so that $C \cap \mathcal{D}, C' \cap \mathcal{D}$ are clusters at level r

- r depends on $\lambda =$"bridge" between C, C' (and $\sigma > 0$ "tube" width)

$$r^d \omega_d \lambda = \frac{k}{\mathrm{n}} + \ldots \textit{confidence term}$$

- it follows that the needed sample size n at level λ

$$n = \mathcal{O}\left(\frac{d}{\lambda \epsilon^2 (\sigma/2)^d \omega_d} \log \frac{d}{\lambda \epsilon^2 (\sigma/2)^d \omega_d} \right)$$

- this **sample complexity** n is almost tight
- for $\alpha < \sqrt{2}$ sample complexity is exponential in d

- New results **[Kent, B. P., Rinaldo, A. and Verstynen, T. 2013]**

- Remark: algorithm(s) can be applied in any metric space

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

(Leskovec Data mining)

- **Agglomerative** (bottom up): n^3
  - Initially, each point is a cluster
  - Repeatedly combine the two "nearest" clusters into one

  → greedy algorithm

- **Divisive** (top down):
  - Start with one cluster and recursively split it

Dendrogram

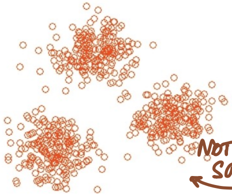
root (1 cluster of n points)

cost of merging or splitting

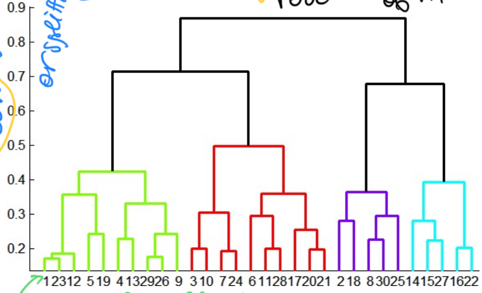
NOT ROBUST K-single linkage

– distance btw points <=> linkage

* cost e.g. $\sum_{i,j \text{ not in same cluster}} \|x^i - x^j\|^2 \iff$ K-means



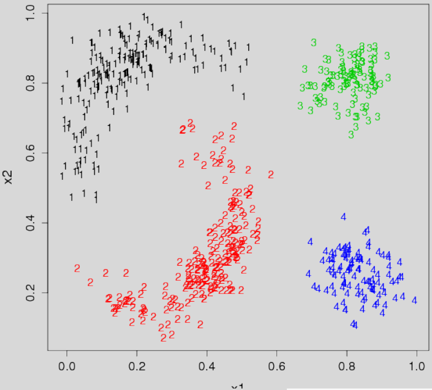
NOT same data



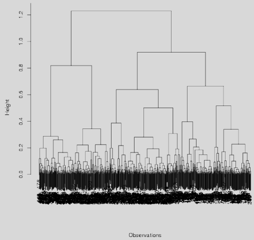
"leaves" n clusters (singletons)

What is hierarchical clustering?

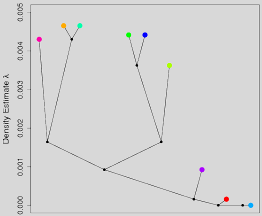
- Clusters have cluster structure
- Represented by
  - **Dendrogram**
  - **Cluster Tree**
    (only from KDE)



Dendrogram



Cluster Tree

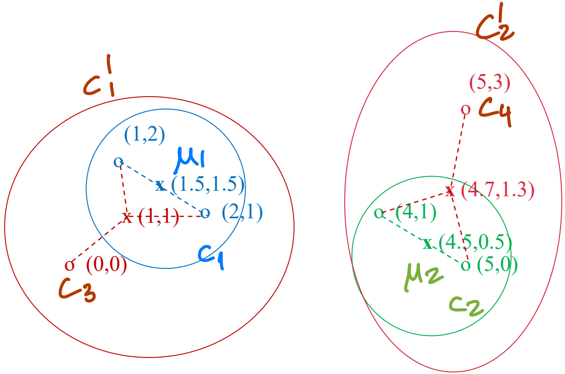


Hierarchical clustering – Overview

(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

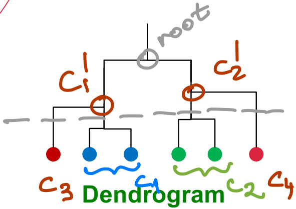


$C_1' = C_1 \cup C_3$

$C_2' = C_2 \cup C_4$

$\text{root} = C_1' \cup C_2'$

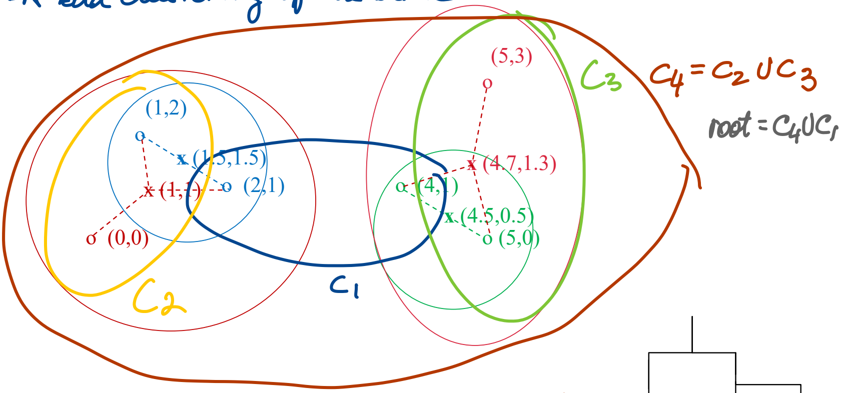
Data:
o … data point
x … centroid



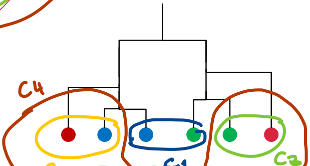
Dendrogram

Example: Hierarchical clustering

A bad clustering of the same data



Data:
o … data point
x … centroid



Dendrogram

Example: Hierarchical clustering

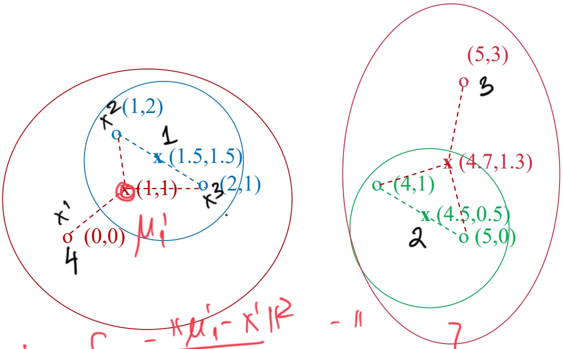
Mix with K clusters

$K=4$

likelihood $= \frac{1}{\sigma_0^2 2\pi} e^0 \times 2 +$

$+ \frac{1}{\sigma_1^2 2\pi}\left[e^{-\frac{\|\mu_1-x^2\|^2}{2\sigma_1^2}} + e^{-\frac{\|\mu_1-x^3\|^2}{2\sigma_1^2}}\right]$

$+ \frac{1}{\sigma_2^2}\left[\qquad \right]$



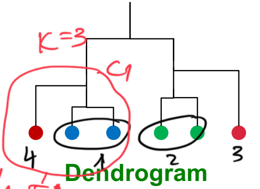
$\frac{1}{\sigma_3^2 2\pi}\left[e^{-\frac{\|\mu_1'-x^1\|^2}{2\sigma_1^2}} + e^{-\|}\right] + \dots$

Data:

o … data point

x … centroid

$\sigma_0 > 0$



μ_1', σ_1'