BIOSTAT 527

4/10/23

Lecture 5

Non-parametric clustering

HW1 link fixed
- due 4/17

HW2 on 4/19

L IV posted

OH Tue in B 321

Lecture IV – Non-parametric clustering

Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

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

What is clustering? Problem and Notation

x_i data point i

- **Informal definition** Clustering = Finding groups in data
- **Notation**

$\mathcal{D}$	=	$\{x_1, x_2, \ldots x_n\}$ a **data set**
n	=	number of **data points**
K	=	number of **clusters** ($K << n$)
Δ	=	$\{C_1, C_2, \ldots, C_K\}$ a partition of $\mathcal{D}$ into disjoint subsets
$k(i)$	=	the **label** of point i
$\mathcal{L}(\Delta)$	=	cost (loss) of Δ (to be minimized)

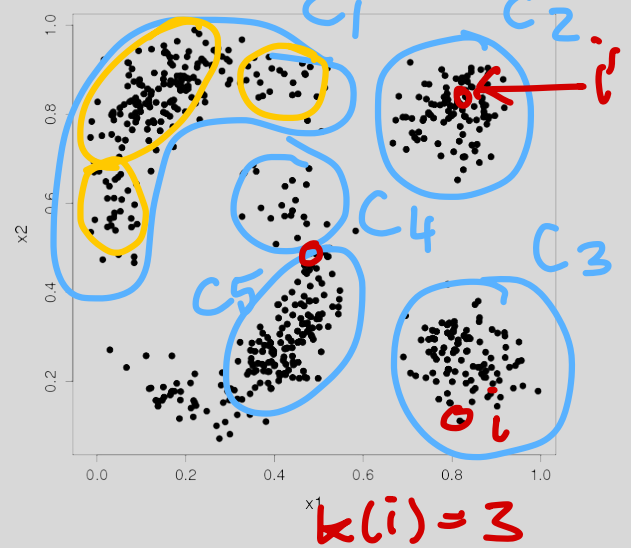
- **Second informal definition** Clustering = given n **data points**, separate them into K **clusters**
- Hard vs. soft clusterings
  - **Hard** clustering Δ: an item belongs to only 1 cluster
  - **Soft** clustering $\gamma = \{\gamma_{ki}\}_{k=1:K}^{i=1:n}$
    γ_{ki} = the **degree of membership** of point i to cluster k

$$\sum_k \gamma_{ki} = 1 \quad \text{for all } i$$

(usually associated with a probabilistic model)

C_k = k-th cluster

$C_{k(i)}$ $k(i)=2$

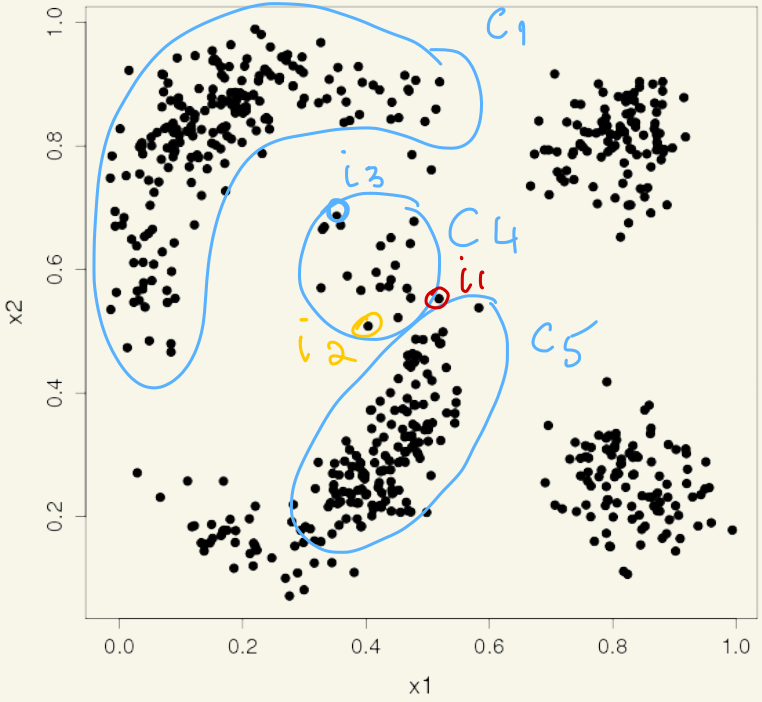


Clustering alg:

what clusters is it trying to find?

i : $\gamma_i(k)$, $k = 1:K$ degree of membership

Soft clustering



i_1 : $\gamma_4(i_1) = 0.5$

$\gamma_5(i_1) = 0.5$

$\gamma_1(i_1) = 0 = \cdots$

$\gamma_4(i_2) = 0.7$

$\gamma_5(i_2) = 0.3$

$\gamma_1(i_3) = 0.1$

$\gamma_4(i_3) = 0.85$

$\gamma_5(i_3) = 0.05$

Clustering Paradigms

Depend on type of data, type of clustering, type of cost (probabilistic or not), and constraints (about K, shape of clusters)

- Data = vectors $\{x_i\}$ in $\mathbb{R}^d$

Parametric (K known)	Cost based [hard]
	Model based [soft]
Non-parametric (K determined by algorithm)	Dirichlet process mixtures [soft]
	Information bottleneck [soft]
	Modes of distribution [hard]
	Gaussian blurring mean shift? [hard]
	Level sets of distribution [hard]

Ex: protein interaction networks

i = protein i

- Data = similarities between pairs of points $[S_{ij}]_{i,j=1:n}$, $S_{ij} = S_{ji} \geq 0$

  Similarity based clustering

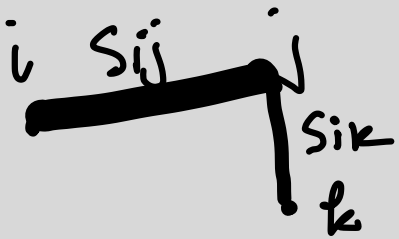
Graph partitioning	spectral clustering [hard, K fixed, cost based]
	typical cuts [hard non-parametric, cost based]
Affinity propagation	[hard/soft non-parametric]

S_{ij} = affinity for interaction

Ex. Social networks

i = person

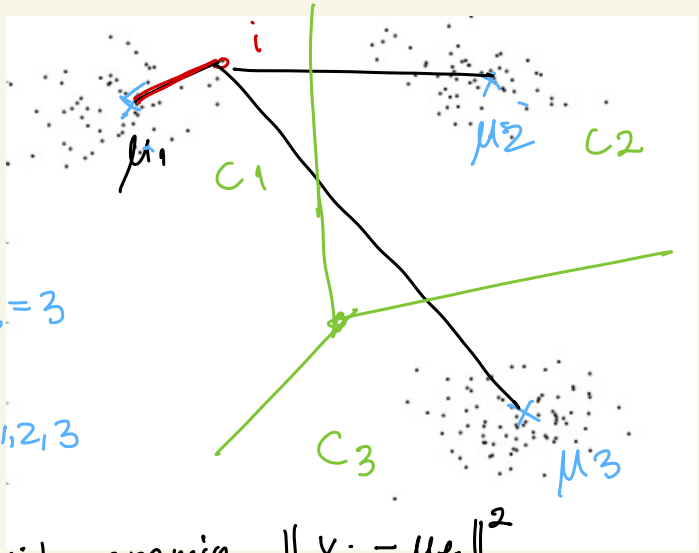
S_{ij} = coauthorship social media links



Classification vs Clustering

	Classification	**Clustering**
Cost (or Loss) $\mathcal{L}$	Expectd error	many! (probabilistic or not)
	Supervised	Unsupervised
Generalization	Performance on new data is what matters	Performance on current data is what matters
K	Known	Unknown
"Goal"	Prediction	Exploration Lots of data to explore!
Stage of field	Mature	Still young

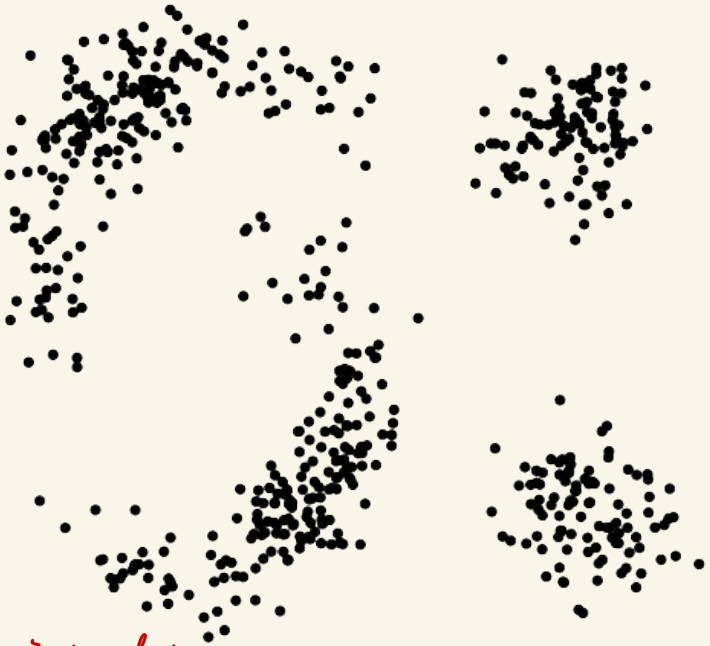
$k(i) = 1 \Leftrightarrow i \in C_1$



$$k(i) = \underset{k=1:K}{\operatorname{argmin}} \; \| x_i - \mu_k \|^2$$

Parametric clustering

- K-means ← K input (# clusters)
- EM ⇒ soft clusterings
  K given



Non-parametric clustering

K unknown ⇒ known only after clustering

K not important

smoothing parameter = kernel width h

Methods based on non-parametric density estimation

Idea The clusters are the isolated peaks in the (empirical) data density

- group points by the peak they are under
- some outliers possible
- $K = 1$ possible(no clusters)
- shape and number of clusters K determined by algorithm
- **structural parameters**
  - **smoothness** of the **density estimate**
  - what is a peak

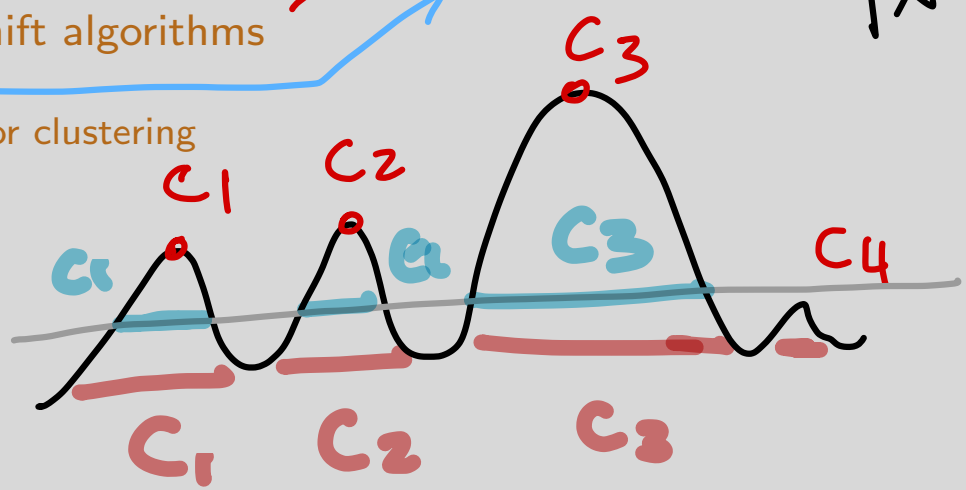
Algorithms

- peak finding algorithms Mean-shift algorithms
- level sets based algorithms
  - Nugent-Stuetzle, Support Vector clustering
- Information Bottleneck ?

peaks = clusters

connected components of y-level set

p_x



y (super) level set of p_x =

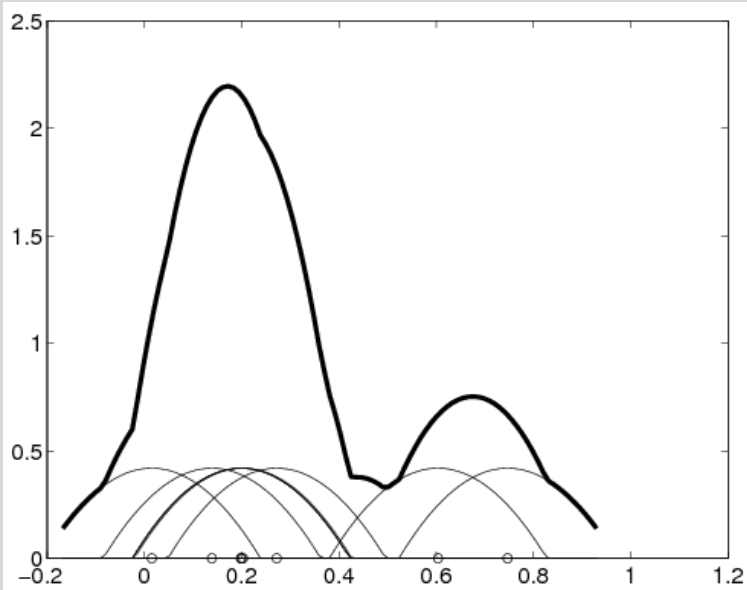
$= \{ x_i \mid p_x(x_i) \geq y \}$

Kernel density estimation

Input
- data $\mathcal{D} \subseteq \mathbb{R}^d$
- **Kernel** function $K(z)$
- parameter **kernel width** h (is a **smoothness parameter**)

Output $f(x)$ a **probability density** over $\mathbb{R}^d$

$$p_x \equiv f(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$



- f is sum of Gaussians centered on each x_i
- f is smoother (less variation) if h larger
- **caveat:** dimension d can't be too large

Mean shift algorithms

Idea find points with $\nabla f(x) = 0$

Assume $K(z) = e^{-||z||^2/2}/\sqrt{2\pi}$ Gaussian kernel

$$\nabla f(x) = -\frac{1}{nh^d}\sum_{i=1}^{n} K(\frac{x - x_i}{h})(x - x_i)/h$$

Local max of f is solution of implicit equation

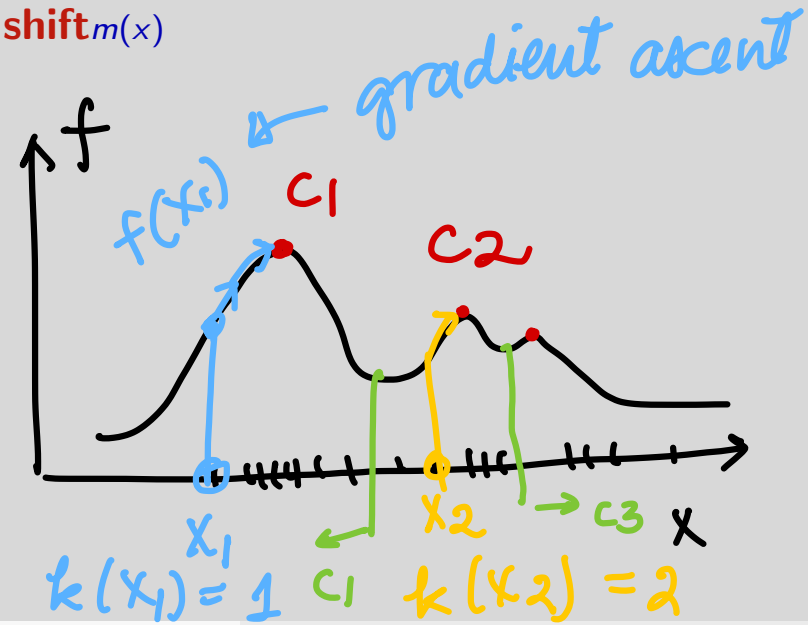
$$x = \underbrace{\frac{\sum_{i=1}^{n} x_i K(\frac{x-x_i}{h})}{\sum_{i=1}^{n} K(\frac{x-x_i}{h})}}_{\text{the mean shift } m(x)}$$

h ↓ $f = KDE_h(x_1, \ldots x_n)$

Algorithm Simple Mean Shift

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

1. for $i = 1 : n$
   1. $x \leftarrow x_i$
   2. iterate $x \leftarrow m(x)$ until convergence to m_i
2. group points with same m_i in a cluster



$$f_h(x) = \frac{1}{nh} \sum_{i=1}^{n} b\left(\frac{\|x - x_i\|}{h}\right)$$

b: $b(z) = \frac{1}{Z} e^{-\frac{z^2}{2}}$

$$b'(z) = \frac{1}{Z}(-z) e^{-\frac{z^2}{2}} = -z\, b(z)$$

$\sum w_i = 1$

$x \in \mathbb{R}^d$

$$\nabla f(x) = \frac{1}{nh} \sum_{i=1}^{n} \left(-\frac{x - x_i}{h}\right) b\left(\frac{x - x^i}{h}\right) = 0 \quad \text{for max}$$

$$\sum_{i=1}^{n} (x_i - x)\, b\left(\frac{\|x - x_i\|}{h}\right) = 0$$

$$\sum_{i=1}^{n} x_i\, b\left(\frac{\|x - x_i\|}{h}\right) = x \sum_{i=1}^{n} b\left(\frac{\|x - x_i\|}{h}\right)$$

solve for x $\Longrightarrow$

$$x = \sum_{i=1}^{n} x_i \frac{b\left(\frac{\|x - x^i\|}{h}\right)}{\sum_{i'=1}^{n} b\left(\frac{\|x - x_{i'}\|}{h}\right)}$$

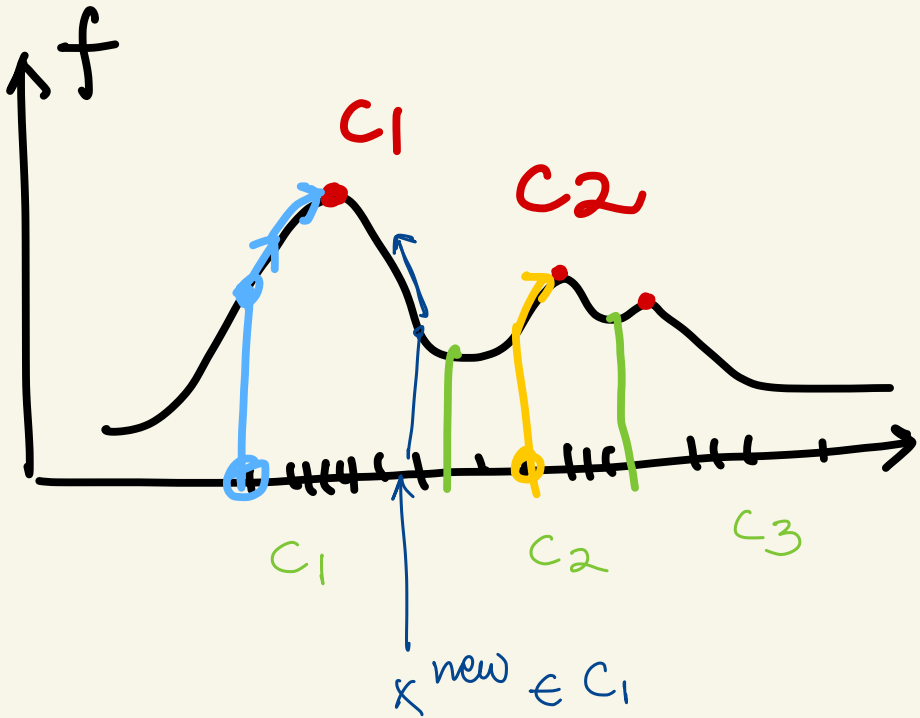
$w_i(x)$

$m(x)$ =

Condition for x = local max

weighted mean of neighbors of x

$m(x) - x$ = Mean shift at x


f
C1
C2
C1
C2
C3
x new ∈ C1