Note

- Homework #3
  - On Lecture 6-7 (last and this ones)
  - **Due in one week** (Tuesday on 3/19)
- Project topic/papers due in one week (3/19).
  - Right before the break
  - Read the assignment thoroughly
  - Your literature survey topic and 5 selected papers must be approved by me. (**Submit them to Canvas discussion**).
  - **Late policy will be applied**. See the course web/PDF.

Note

- **Complete the Exercise #1: final session**
  - **HW Assignment: Complete FP#1 on PCA and make your final submission of your code and results (screen shots/short doc report) via Canvas by tomorrow midnight.**
- Fast Prototyping Exercise #2 on Mean Shift Segmentation starts next week
  - **HW Assignment: Read carefully and thoroughly the reference paper**: https://bidal.sfsu.edu/~kazokada/csc872/PD2.pdf
  - **Download:** https://bidal.sfsu.edu/~kazokada/csc872/DATA/Segmentation_Data.zip

Non-Parametric Statistical Modeling

CSC 872

Pattern Analysis and Machine Intelligence

References
Andrew Moore's great slides at
http://www.cs.cmu.edu/~awm/tutorials

Statistical Modeling: Review

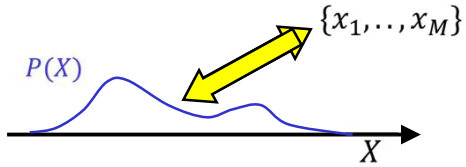
Knowledge based AI Logical Inference KB ← Logical Lang (PL, FOL)

Bayesian Framework Probabilistic Inference using Bayes rule KB ← PDFs.

is → uncertainty Posterior Prob ← Bayes rule Likelihood & Prior

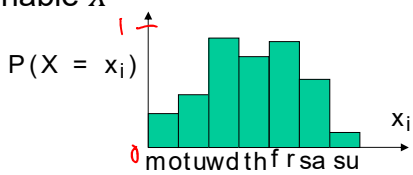
Post

- Problem
  - **Estimating probability distribution from data (Bayesian framework: needs likelihood&prior)**
- Assumption
  - **Data = Samples drawn from an unknown underlying distribution**
- Question
  - **What is the underlying distribution given these data?**

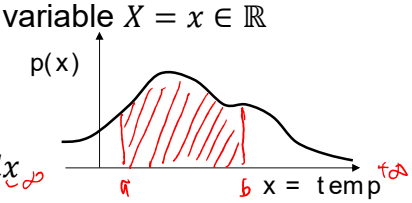


KR: Probability Distribution

- **Probability Mass Function**
  - For a discrete random variable X
  - $P(X)$: $\sum_{x_i} P(X = x_i) = 1$
  - $\forall\, x\; P(X) \geq 0$



- **Probability Density Function** PDF
  - For a continuous random variable $X = x \in \mathbb{R}$
  - $p(x)$: $\int_{-\infty}^{\infty} p(x)dx = 1$
  - $\forall x\; p(x) \geq 0$
  - $P(a < X < b) = \int_a^b p(x)dx$



What for?

- Bayesian X (= inference, classification etc)
- Computing Expectation
  - $E[X] = \sum_i x_i P(x_i) = \mu$ (population mean)
  - $E[f(X)] = \sum_i f(x_i)P(x_i)$
  - $E[aX + Y + b] = aE[X] + E[Y] + b$ (linear operator)
- Solving Maximum Likelihood Estimation Numerically
- Anomaly Detection
  - Sort events by associated probability (anomaly=small prob)
- Clustering/Segmentation/VectorQuantization!!!

Non-Parametric Modeling

- **Most simple way** for representing distribution
- Density Estimation
- Basically Two Steps
  1) Domain quantization (Binning)
  2) Measure Frequency / Density (Counting)

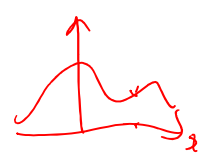
- Useful **when no known prior information about functional property** of the distribution

We don't know the shape of the type of PDFs.

PF: Basic Density Estimation

- Given a sample $\boldsymbol{x}$, a **density estimator** $\boldsymbol{M}$ can tell you how likely each data item is:

$$\hat{P}(\mathbf{x}|M)$$

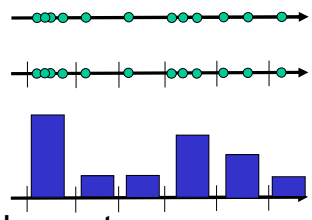


- Given a sample set $\{\boldsymbol{x_1}, .., \boldsymbol{x_N}\}$, find a density estimator M that most accurately estimate the likelihood function

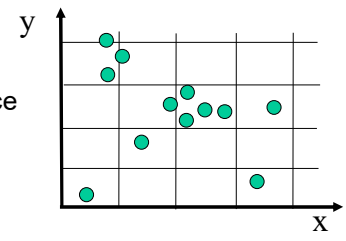
- You can model this with constructing a **normalized histogram** from $\{\boldsymbol{x_1}, .., \boldsymbol{x_N}\}$

PF: Histogram as a Density Estimator

- Given a dataset $\{\boldsymbol{x}_1, .., \boldsymbol{x}_N\}$
- Quantization of data space
- Frequency counts

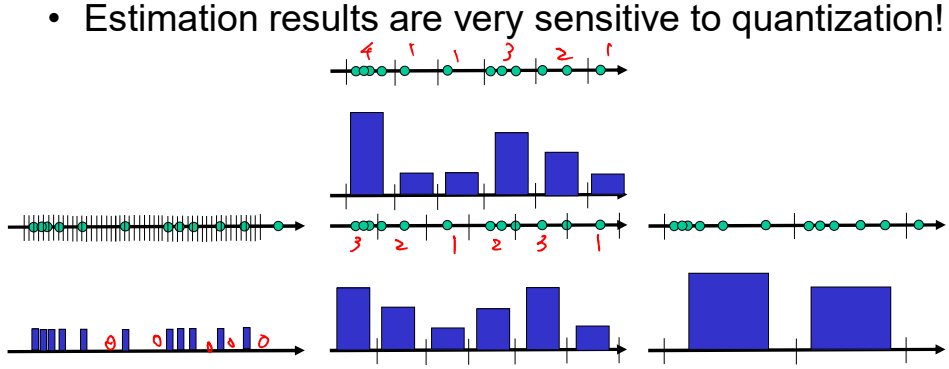


- **Normalization** = divide by total counts
- Modeling **Probability Mass Function** $P(x)$
- Modeling **Joint Distribution**
  - $P(x, y, z, \ldots)$
  - Frequency counting in N-D feature space



Issues: Bin Size

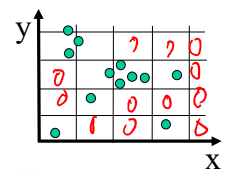
- Estimation results are very sensitive to quantization!



- For continuous RV, quantization may be difficult
- You cannot cover entire domain accurately

Issues: Differentiability

- For some quantization, you may have some bins having zero values… So what?



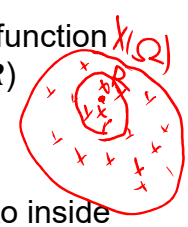
- You cannot compute dataset density

$$\hat{P}(dataset|M) = \hat{P}(x_1 \wedge x_2 \wedge \cdots \wedge x_N|M) = \prod_{n=1}^{N} \hat{P}(x_n|M) \rightarrow 0$$

- You cannot differentiate P(x|M) at bins with zero
  - You cannot compute gradient
  - Later you will see how you don't like this..

Modeling Probability Density Function

- Probability that $\boldsymbol{x}$, drawn from unknown density function $\boldsymbol{p(x)}$, fall inside some region $\boldsymbol{R}$ ($\boldsymbol{V}$ is volume of $\boldsymbol{R}$)



$$P = \int_R p(\mathbf{x}')d\mathbf{x}' \simeq p(\mathbf{x})V$$

mean value theorem

- Given $\boldsymbol{N}$ points, probability that $\boldsymbol{K}$ of them fall into inside region $\boldsymbol{R}$ follows the binomial law. Thus

$$E[K/N] = P$$

- Since the variance vanishes as $N \to \infty$, we approximate

$$P \simeq K/N$$

- With above results (two approximations) we have

$$p(\mathbf{x}) \simeq \frac{K}{NV}$$

Biship: p50-55

PF: Kernel Density Estimation KDE

- Set region $\boldsymbol{R}$ as a hypercube of edge-length $\boldsymbol{h}$ $V = h^d$

center loc. width ← (bandwidth)

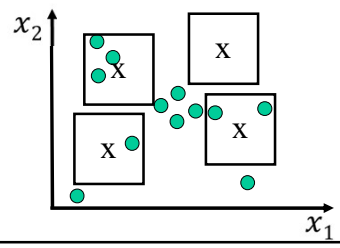
- ***Kernel function*** $\boldsymbol{H}(\boldsymbol{x}, \boldsymbol{h})$ is used to count data points lying inside the hypercube (also known as Parzen window)

$\{x_1, \ldots x_N\}$

$$H(\boldsymbol{x} - \boldsymbol{x}_n, h) = \begin{cases} 1 & \frac{|x_j - x_{nj}|}{h} < 0.5 \quad j = 1, .., d \\ 0 & otherwise \end{cases}$$

$$\boldsymbol{x} = (x_1, .., x_j, .., x_d)$$

sample x_n is inside the hypercube



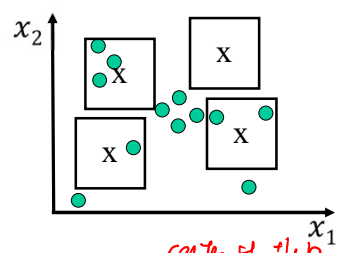
So a kernel is a windowing function!!!

PF: Kernel Density Estimation cond

- Counting with the kernel

$$K = \sum_{n=1}^{N} H\left(\frac{|\boldsymbol{x} - \boldsymbol{x}_n|}{h}\right)$$

$p(x) \cong \frac{K}{NV}$



- **Kernel Density Estimate (KDE)**

center of the R

$$\hat{p}(\boldsymbol{x}) \cong \frac{K}{NV} = \frac{1}{Nh^d} \sum_{n=1}^{N} H\left(\frac{|\boldsymbol{x} - \boldsymbol{x}_n|}{h}\right)$$

size of R

Is KDE different from Histogram?

- Both do simple counting but…..

- histogram defines the bins in advance
- YES, KDE removes this quantization step all once!
- Bin size = Hypercube size = *h*
- Ok. This seems more flexible but what about the differentiability?
- **NO!** still discontinuous. H() may not be able to be differentiated at everywhere.

A Good News

- You can generalize the original KDE to a smooth function simply by choosing a right kernel function !!!

- What should we use then?

$H(x-x_i, h) = \begin{cases} 1 \\ 0 \end{cases}$

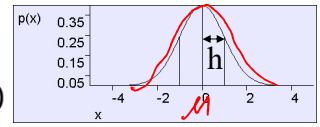
Gaussian

among others …

KR: Gaussian (Normal) PDF

- Univariate Gaussian PDF

$$p(x;\mu,h) = N(\mu,h) = \frac{1}{\sqrt{2\pi h^2}}\exp\left(-\frac{(x-\mu)^2}{2h^2}\right)$$



→ N-D space

determinant

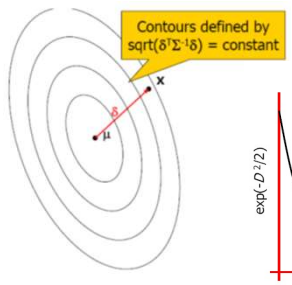
- Multivariate Gaussian PDF

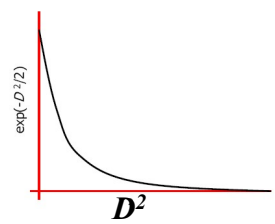
Mahalanobis Distance

$$p(\boldsymbol{x};\mu,\Sigma) = |2\pi\Sigma|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\mu)^T\Sigma^{-1}(\boldsymbol{x}-\mu)\right)$$

covariance

1) Vector $\boldsymbol{x}$
2) Define $\delta = \boldsymbol{x} - \mu$
3) Count number of contours crossed of ellipsoid formed by Σ^{-1}
4) D = Count = $sqrt(\delta^T\Sigma^{-1}\delta)$ = **Mahalanobis distance**
5) $\exp(-D^2/2)$
6) $\boldsymbol{x}$ close to μ in the squared Mahalanobis distance gets higher weight!





17

Why Gaussian?

- We like Gaussian because
  - It is *very useful* to understand basic PAMI concepts
  - *Smooth function*: derivative of exp is exp!
  - Fourier transform of Gaussian is Gaussian
  - Gaussian marginal, conditional are Gaussian
  - Linear transform of Gaussian is still Gaussian
- OK. there are much more to this
  - More on https://www.autonlab.org/_media/tutorials/gaussian14.pdf

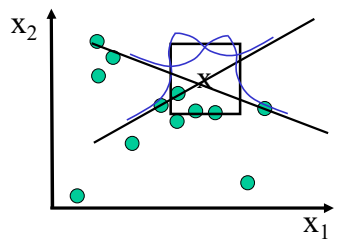
18

KDE with Gaussian

- Now we have smooth density estimator without the quantization hassle!!!

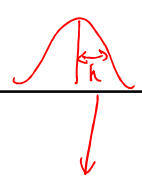
$$\hat{p}(\boldsymbol{x}) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{|2\pi\Sigma|^{1/2}}\exp(-(\boldsymbol{x}-\boldsymbol{x}_n)^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{x}_n))$$

K



It's a weighted sum!

Any Catch?



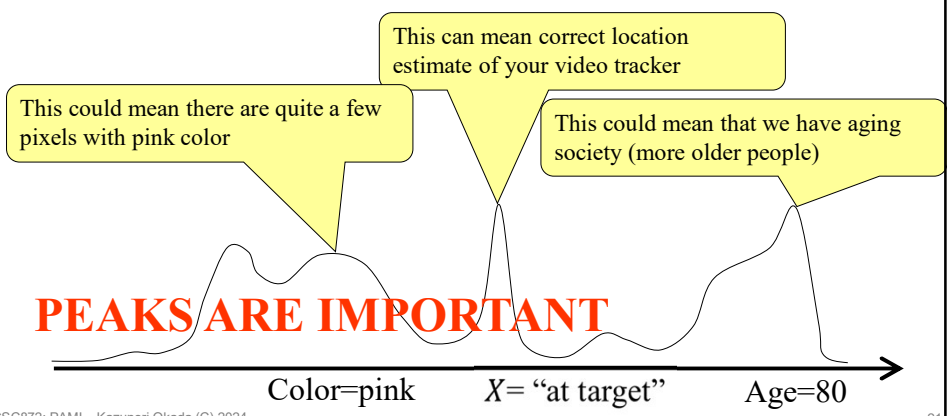
- No quantization hassle but
- You still need to tune the **bandwidth**!
- Which is another research topic…

both space & time.

- **And its complexity is high**
- You need to keep all data points
- For every location, you also need to evaluate every data points
- So it is a lot of computation with lots of data…

PF: Finding Peaks of PDF

- We want to find a peak of PDF because
  - Statistical Modes
  - (Locally) Maximum Likelihood Estimate



PS: Hill-Climbing

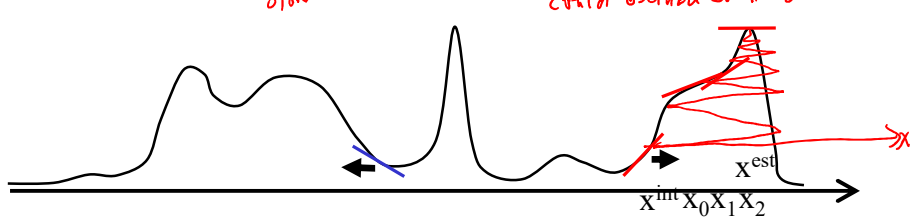
- Find a peak iteratively from an initial point $\boldsymbol{x}^{int}$
- Local peak defined as $\Delta\, \boldsymbol{p}(\boldsymbol{x}) = \mathbf{0}$ (horizontal tangent!)

$$\mathbf{x}^* = \mathrm{argmax}_{\mathbf{x}} p(\mathbf{x}|\mathbf{x}_1, .., \mathbf{x}_N)$$

- Hill-Climbing by **Gradient Ascent**
  - Iterative update by $\boldsymbol{x}_{n+1} = \boldsymbol{x}_n + \eta\, \Delta\, \boldsymbol{p}(\boldsymbol{x}_n)$ with small positive $\boldsymbol{\eta}$
  - It could oscillate & diverge, depending on $\boldsymbol{\eta}$
  - Small $\boldsymbol{\eta}$ sure-convergence; large $\boldsymbol{\eta}$ faster convergence

slow

could oscillate or diverge.

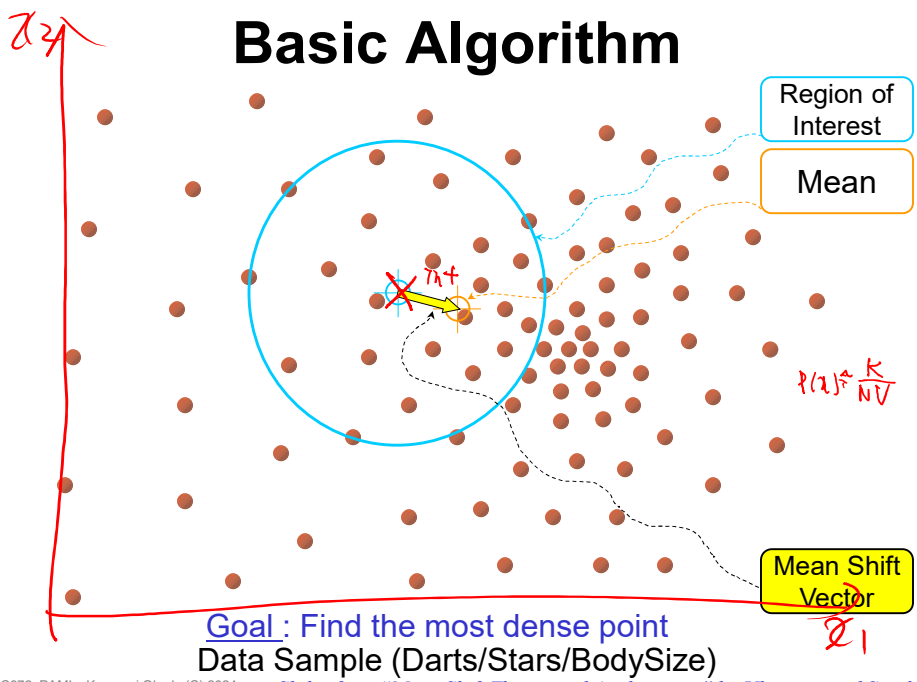


PS: Mean Shift

- Adaptive step-size gradient ascent for PDFs constructed by KDE
  - **no η!**
- Provably convergent
  - **no oscillation and no divergence!**
- Implicit construction of KDE
  - It is local hill-climbing so there is no need to construct KDE for entire domain at once
  - **Local density is computed only when needed**

Fukunaga '75
Cheng '95
Comaniciu '99

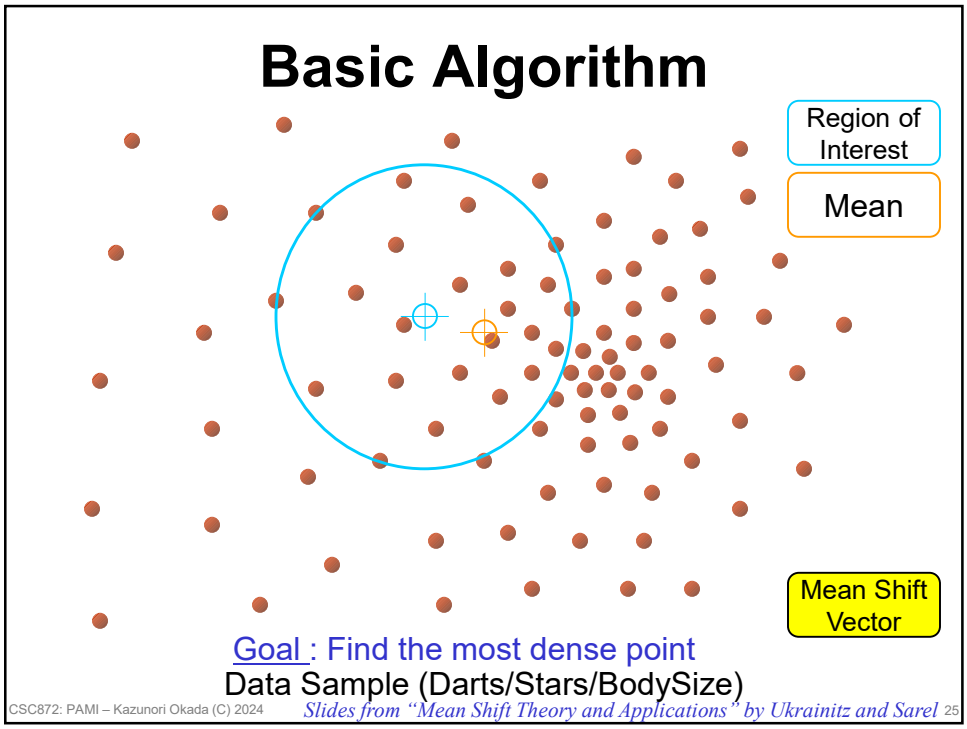
Basic Algorithm



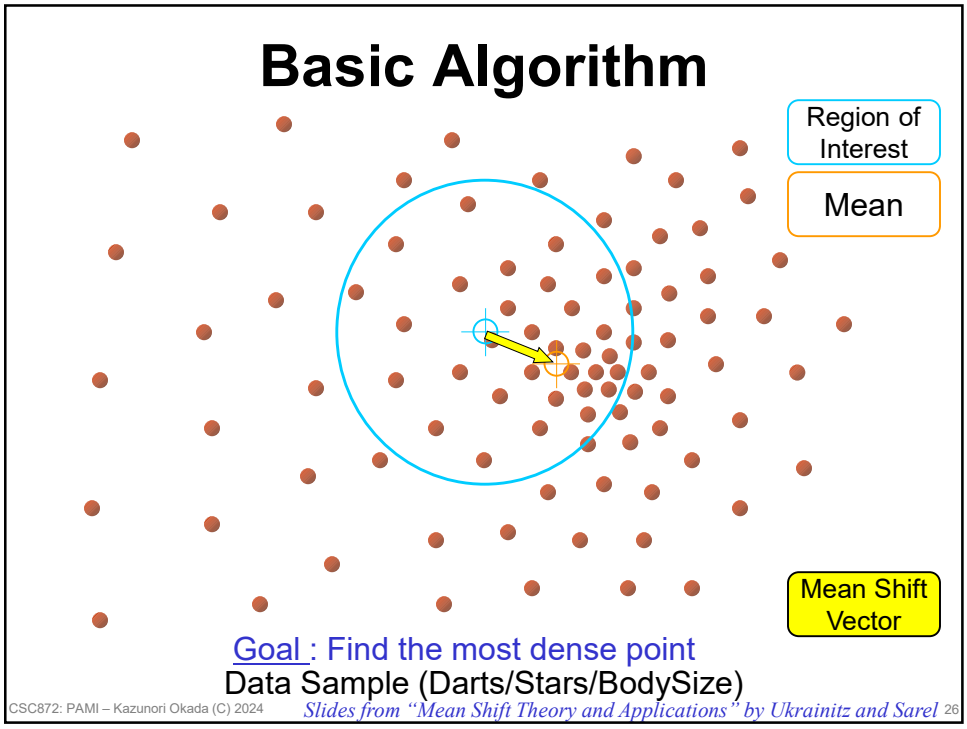
Goal : Find the most dense point

Data Sample (Darts/Stars/BodySize)

Slides from "Mean Shift Theory and Applications" by Ukrainitz and Sarel

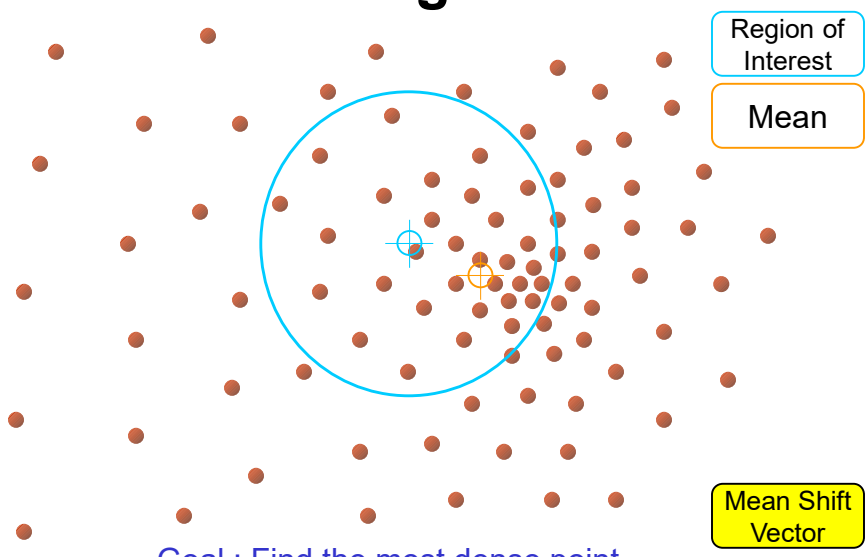

Basic Algorithm
Region of Interest
Mean
Mean Shift Vector
Goal : Find the most dense point
Data Sample (Darts/Stars/BodySize)
CSC872: PAMI – Kazunori Okada (C) 2024
Slides from "Mean Shift Theory and Applications" by Ukrainitz and Sarel
25

25


Basic Algorithm
Region of Interest
Mean
Mean Shift Vector
Goal : Find the most dense point
Data Sample (Darts/Stars/BodySize)
CSC872: PAMI – Kazunori Okada (C) 2024
Slides from "Mean Shift Theory and Applications" by Ukrainitz and Sarel
26

26

Basic Algorithm

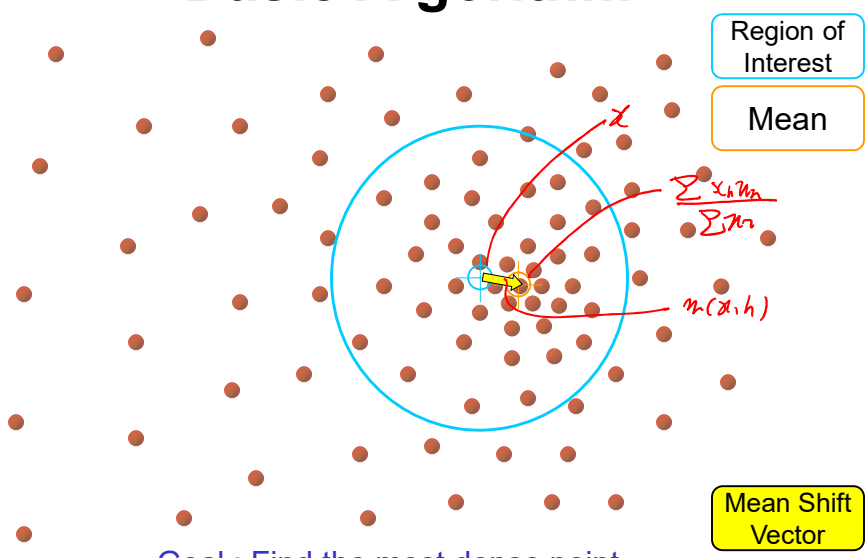


Goal : Find the most dense point

Data Sample (Darts/Stars/BodySize)

 Slides from "Mean Shift Theory and Applications" by Ukrainitz and Sarel

Basic Algorithm

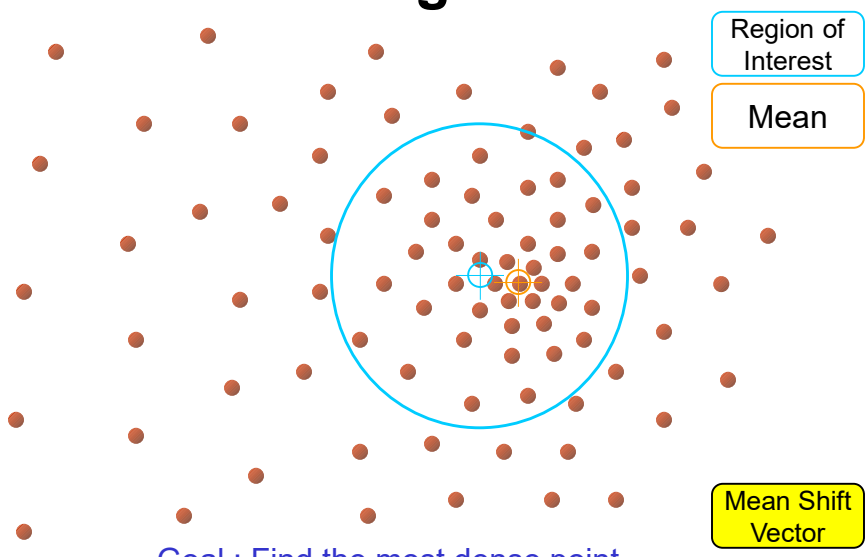


Goal : Find the most dense point

Data Sample (Darts/Stars/BodySize)

 Slides from "Mean Shift Theory and Applications" by Ukrainitz and Sarel

Basic Algorithm

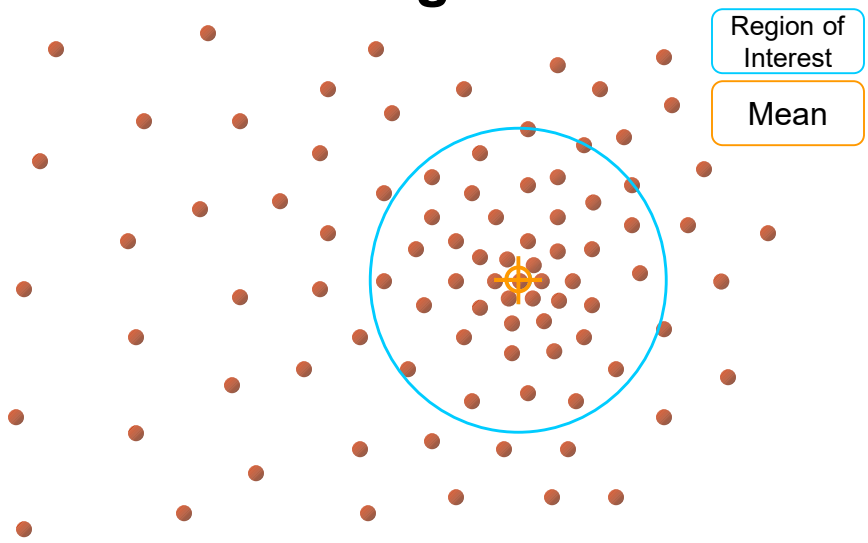


Goal : Find the most dense point

Data Sample (Darts/Stars/BodySize)

 Slides from "Mean Shift Theory and Applications" by Ukrainitz and Sarel

Basic Algorithm



Goal : Find the most dense point

Data Sample (Darts/Stars/BodySize)

 Slides from "Mean Shift Theory and Applications" by Ukrainitz and Sarel

PS: Mean Shift Algorithm

New mean

- Mean Shift Vector given samples $\boldsymbol{x_1}, .., \boldsymbol{xi}, .., \boldsymbol{x_N}$

center loc.

$$\mathbf{m}(\mathbf{x}, h) = \frac{\sum_{i=1}^{N} \mathbf{x}_i g(\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\|^2)}{\sum_{i=1}^{N} g(\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\|^2)} - \mathbf{x}$$

size R

current point

vector norm $\|\mathbf{x}\| = \sqrt{x_1^2 + .. + x_d^2}$

- Epanechnikov Kernel $g(\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\|^2) = \{ \begin{matrix} C & \|\mathbf{x} - \mathbf{x}_i\| \leq h \\ 0 & otherwise \end{matrix}$
- Gaussian Kernel

$$g(\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\|^2) = \exp(-\frac{\|\mathbf{x}-\mathbf{x}_i\|^2}{h^2})$$

- Mean Shift Procedure (Iterate this from $\boldsymbol{x_1}$ to $\boldsymbol{x_k}^{\boldsymbol{mle}}$)

$$\mathbf{x}_{n+1} = \mathbf{m}(\mathbf{x}_n, h) + \mathbf{x}_n$$

- Stopping Criteria

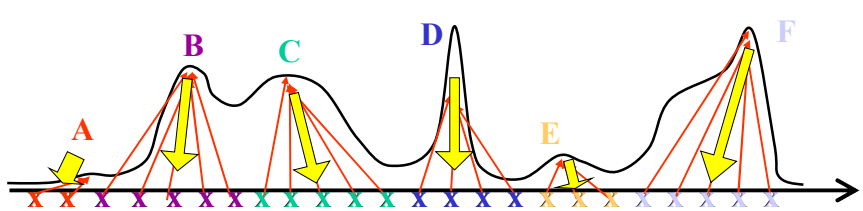
$$\frac{\|\mathbf{m}(x,h)\|^2}{h^2} < TH^2$$

Some ideas for how to derive these will be given in the paper you will read.

PF: Relation to Clustering

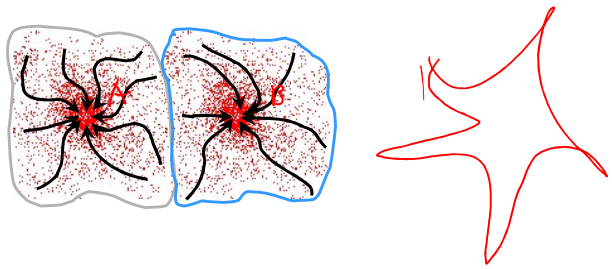
- MS is a solution for adaptive clustering!
- No need to know how many clusters are there in data *a priori*

1) For all data points, perform the MS procedure, recording their convergence to nearby modes
2) Label each mode by grouping
3) Re-assign each data points with mode labels



Mean Shift Clustering in 2D

- Advantages
  - No need to know the number of clusters a priori
  - The shape of cluster does not need to be regular



PF: Is it learning?

- Just summarizing data?
  - Descriptive Statistics
  - Inferential Statistics
- What designer must choose?
  - Quantization / Bin Size / Kernel Bandwidth
  - Types of estimation
- What must be derived from data?
  - Data Count / Frequency / Density

Summary

- Non-Parametric Statistical Modeling
  - Work with arbitrary distributions
  - Histogram
  - Kernel Density Estimation
  - Mean Shift Mode Seeking
  - Clustering
  - Memory and Time Complexity…
- Next
  - How to make use of prior knowledge of distribution?
  - Parametric Modeling!
  - HW3 & Project Topic due next week
  - FP#2 on Mean Shift Segmentation: **READ THE PAPER & download images!!!**