DENSITY ESTIMATION

Shingchern D. You

Density estimation

- Parametric (mentioned previously)
  - Given a model, estimate parameters with MLE or MAP
  - Can be used for classification & regression
- Semiparametric
  - Mixture density (such as **GMM**)
  - Trained with **Expectation-maximization (EM)** algorithm
- Nonparametric
  - Histogram
  - K-NN

Why density estimation

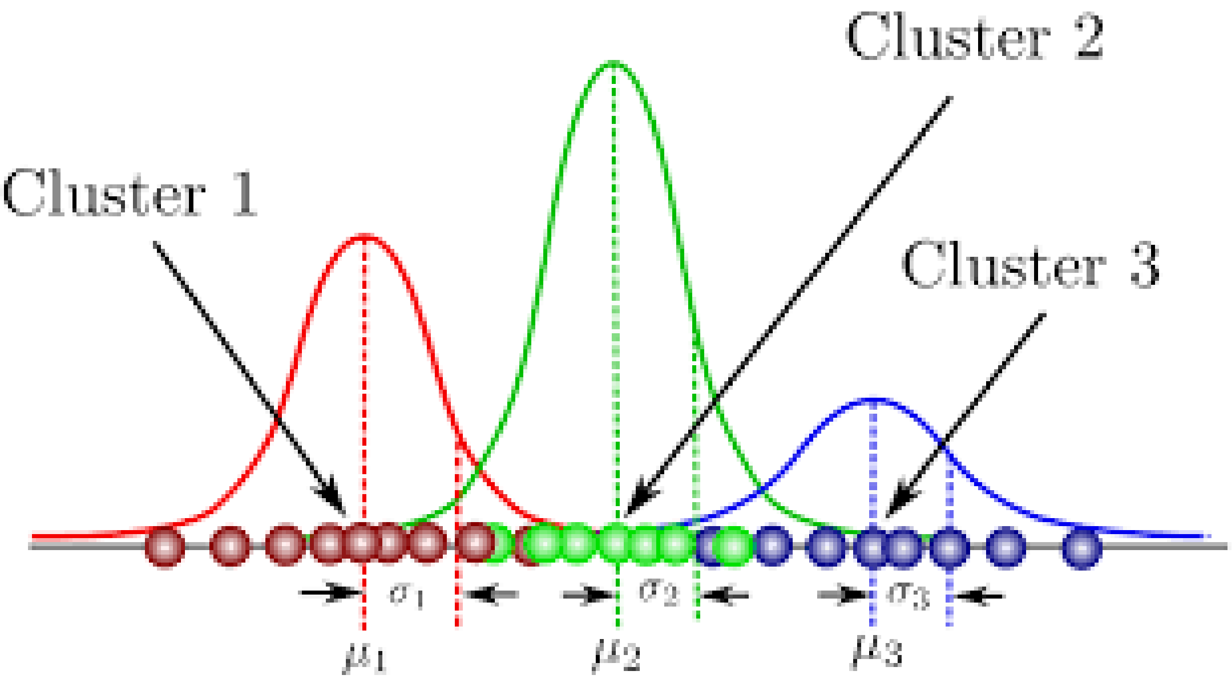
- Recall that we need to compute probability to use Bayes classifiers
- Density means probability density function, used to compute probability

Parametric estimation recap

- The term “parametric” in probability (statistics) means the distribution of the problem is know
- Without knowing anything, we assume
  - Equal probability for discrete case
  - Gaussian for continuous case
- Used model
  - Independent RV: Naïve Bayes classifier
  - Non-independent RV: Bayes classifier

Mixture of density

- A sample is randomly chosen from three clusters
- Each cluster has its own density function
- https://towardsdatascience.com/gaussian-mixture-models-explained-6986aaf5a95



Introduction to GMM

- A **Gaussian mixture model** is a weighted sum of Gaussian densities $g(\boldsymbol{x}_{(\boldsymbol{i})}|\boldsymbol{\mu}_j, \Sigma_j)$ given by (after some math steps omitted)

$$f_y(\boldsymbol{x}_{(\boldsymbol{i})}|\theta) = \sum_j \alpha_j g(\boldsymbol{x}_{(\boldsymbol{i})}|\boldsymbol{\mu}_j, \Sigma_j)$$

where $\alpha_1 + \cdots + \alpha_m = 1$ and $\boldsymbol{x}_{(\boldsymbol{i})}$ is an outcome (observation) from a multidimensional RV $\mathbf{y}$

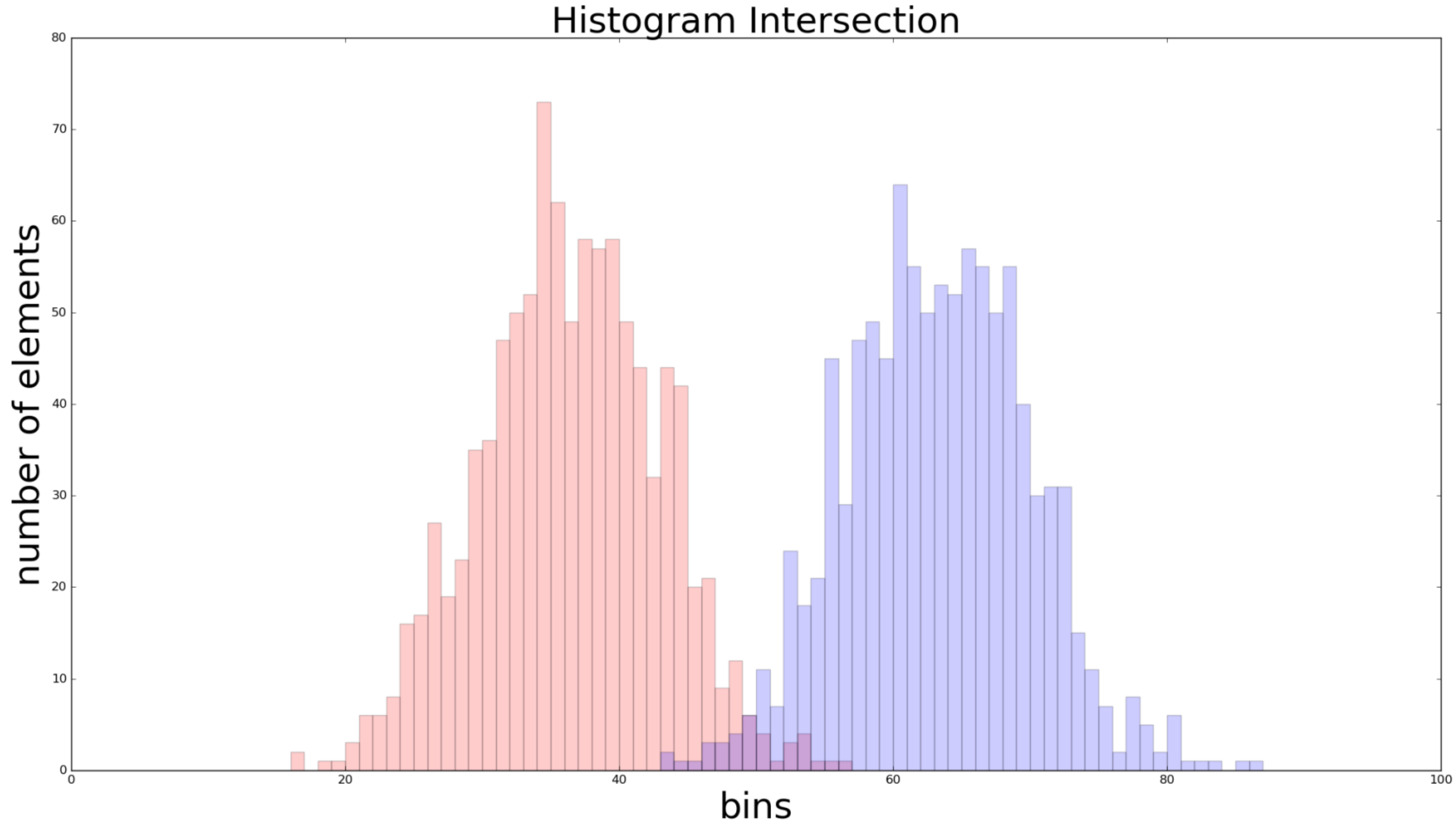
- It can approximate any pdf (with sufficiently large m)

Introduction to GMM

- GMM is an extension of Gaussian model
  - Gaussian model has only **one** mixture
- GMM can be used as
  - Classifier
  - Soft clustering
- Model complexity
  - Naïve Bayes $\leq$ Bayes $\leq$ GMM

Histogram example

- Not related to our problem (https://mpatacchiola.github.io/blog/2016/11/12/the-simplest-classifier-histogram-intersection.html)



Histogram

- How to determine # of bins (no obvious approach)
- Problem of origin
  - Bin width = 1
  - Dataset: 0.99, 0.98, 1.01, 1,99, 2.02
  - Bin 1 (0 ~ 0.99): 2, center @ 0.5
  - Bin 2 (1.00 ~ 1.99): 2, center @ 1.5
  - Bin 3 (2.00 ~ 2.99): 1, center @ 2.5
- But, obviously there are only two clusters

Histogram Naïve estimator

- Let x_i be a data point with total of N points

$$\hat{p}(x) = \frac{\#\{x - h/2 < x_i \leq x + h/2\}}{Nh}$$

- Center the bin to x in $\hat{p}(x)$
- Can plot a curve if x is a variable

Histogram kernel estimator

- We use a rectangle window function previously, i.e., $\#\{x - h/2 < x_i \leq x + h/2\}$
- It is possible to use a kernel function instead

$$\hat{p}(x) = \frac{1}{Nh} \sum_{i=1}^{N} g\left(\frac{x - x_i}{h}\right)$$

where $g(\cdot)$ is a kernel function

K-NN

- K-NN can be use for
  - Density estimate (exercise)
  - Classification
  - Regression

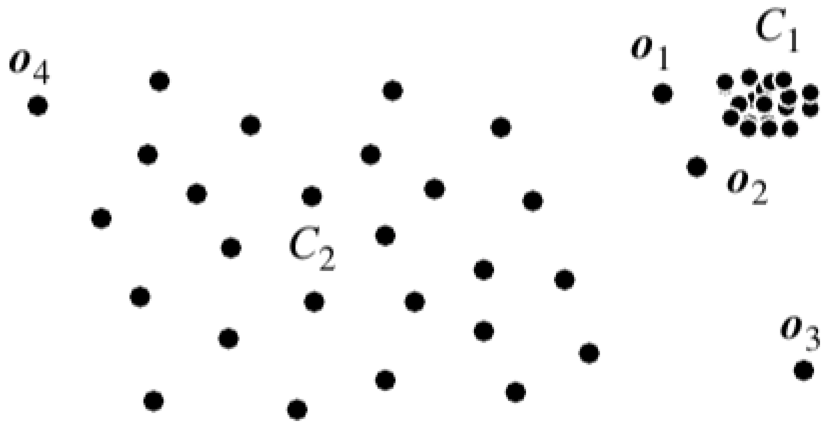
Outlier detection

- Similar term: Anomaly detection
- Used to detect special cases, such as credit fraud
  - Identity theft
  - Card lost
  - etc

https://www.eastwestbank.com/ReachFurther/en/News/Article/Credit-Card-Fraud-The-Three-Words-You-Never-Want-to-Hear

Outlier detection

- Distance based detection (only o3) versus density based detection (able to detect o1 & o2)
  - Local density ~ local distribution



https://towardsdatascience.com/density-based-algorithm-for-outlier-detection-8f278d2f7983

Outlier detection

- Outliner detection (list only a few)
  - Probabilistic approach
  - Factor analysis
  - LOF (Local outlier factor) – a density-based approach
  - Autoencoder
- Read the following for the idea of LOF:
  https://towardsdatascience.com/density-based-algorithm-for-outlier-detection-8f278d2f7983