CSE6740 Computational Data Analysis

Lecture Notes

August 27, 2025

Nitya Maruthuvakudi Venkatram

Topic: Density Estimation

1 Introduction

Density estimation is used to understand the distribution of the data. It can be used to estimate the pobability desnity or likelihood of each of the points in the data, which is essential to identify outliers or perform any clustering or model fiting. We can use the samples or data to recover the distribution, since we don't have a closed form expression for the distribution.

- Use: Learn the shape of data cloud.
- Helps assess whether a point is typical (high density) or an outlier (low density).
- Applied in both supervised and unsupervised learning.

Example: Histograms are one of the most fundamental and commonly used form of density estimation (Fig 1). Histogram data can also be expressed in the form of a contour or scatter plot to be able to relatively compare two parameters (Fig 2).

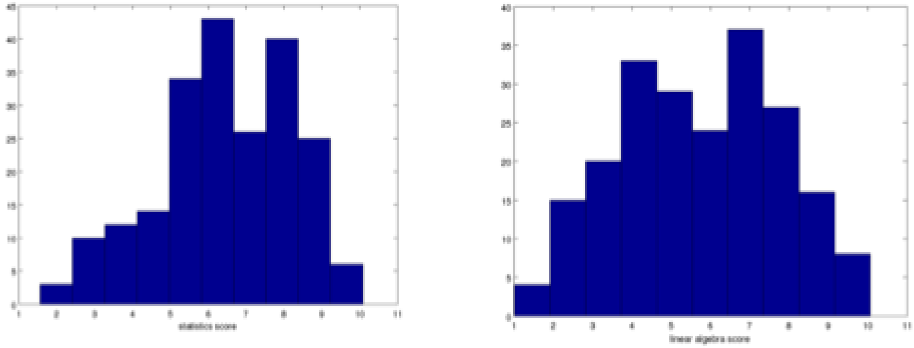


Figure 1: Histograms for 2 variables

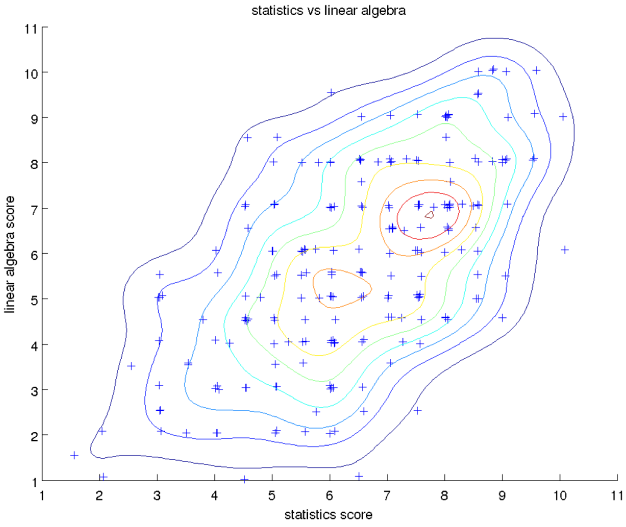


Figure 2: Contour-Scatter plot between the 2 variables

2 Parametric Models

These are models that can be described by a **fixed** and finite parameters. They can be discrete or continuous.

- Discrete Example: Bernoulli distribution (one parameter θ).
- Continuous Example: Multivariate Gaussian distribution with mean μ and covariance Σ.

2.1 Maximum Likelihood Estimator (MLE)

It is one of the most popular ways to do density extimation for paramteric models. MLE is also a technique that can be used to clean data i.e. find and remove any outliers.

Say we have n data points $D = \{x^1, x^2, x^3, ..., x^n\}$ that are draw **identically and independently** from a distribution $P^*(x)$ and we want to fit the data with a model $P(x|\theta)$ with the parameter θ, we can estimate the parameter by -

$$\hat{\theta} = \arg\max_{\theta} \sum_{i=1}^{n} \log P(x_i|\theta) \tag{1}$$

This physically means, we are finding the best value of the parameter θ that fits this data such that it has the maximum likelihood (or log likelihood).

2.1.1 MLE of a biased coin

Let θ be the probability of heads when a **biased** coin is tossed. Given we have n flips that are **identically and independently distributed**, we can find θ using MLE.

Let $D = \{x^1, x^2, x^3, ..., x^n\}, x^i \in \{0, 1\}$ and if we assume that coin flipping follows a Bernoulli distribution, we have $P(x|\theta) = \theta^x(1-\theta)^{1-x}$ i.e.

$$P(x|\theta) = \begin{cases} 1-\theta, \text{ for } x = 0 \\ \theta, \text{ for } x = 1 \end{cases}$$

The likelihod of seeing each of $x^i = L(\theta; D) = \prod_{i=1}^n P(x_i|\theta) = \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i}$ Taking the log-likelihood:

$$\ell(\theta) = \log L(\theta; D) = \sum_{i=1}^n [x_i \log\theta + (1-x_i)\log(1-\theta)]$$

Let, $n_{\text{heads}} = \sum_{i=1}^n x_i$, the number of observations with value 1. Then:

$$\ell(\theta) = n_{\text{heads}} \log\theta + (n - n_{\text{heads}})\log(1-\theta)$$

To find the maximum likelihood, take derivative w.r.t. θ and set to zero:

$$\frac{d\ell}{d\theta} = \frac{n_{\text{heads}}}{\theta} - \frac{n - n_{\text{heads}}}{1-\theta} = 0$$

$$\implies n_{\text{heads}}(1-\theta) - (n - n_{\text{heads}})\theta = 0$$

$$\implies n_{\text{heads}} - n_{\text{heads}}\theta - n\theta + n_{\text{heads}}\theta = 0$$

$$\hat{\theta} = \frac{n_{\text{heads}}}{n} = \frac{1}{n}\sum_{i=1}^n x_i$$

The MLE estimate for θ is simply the fraction of observed 1's (heads) in the data.

2.1.2 MLE for estimating parameters of a Univariate Gaussian Distribution

Consider data $D = \{x_1, x_2, \ldots, x_n\}$ assumed i.i.d. from a Gaussian distribution:

$$P(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

The likelihood function is:

$$L(\mu, \sigma^2; D) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

The log-likelihood function is:

$$\ell(\mu, \sigma^2) = \log L(\mu, \sigma^2; D) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2$$

MLE for μ

Take the derivative of ℓ with respect to μ and set it to zero:

$$\frac{\partial\ell}{\partial\mu} = -\frac{1}{2\sigma^2} \cdot 2\sum_{i=1}^n (x_i - \mu)(-1) = \frac{1}{\sigma^2}\sum_{i=1}^n (x_i - \mu)$$

$$\frac{\partial \ell}{\partial \mu} = 0 \implies \sum_{i=1}^{n}(x_i - \mu) = 0 \implies n\mu = \sum_{i=1}^{n} x_i$$

$$\boxed{\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i}$$

The MLE estimate for μ is simply the average of all possible values of the data i.e the sample mean.

MLE for σ^2

Take the derivative of ℓ with respect to σ^2:

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2} \cdot \frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

Set derivative to zero:

$$-\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i - \mu)^2 = 0$$

$$\implies -n\sigma^2 + \sum_{i=1}^{n}(x_i - \mu)^2 = 0$$

$$\boxed{\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2}$$

The MLE estimate for σ^2 is simply the average sample variance.
This way, MLE can be used to find the parameters for different parametric models.

2.2 Advantages and Disadvantages of Parametric Models

Advantages:

- It is simple and straightforward since we just use MLE to estimate parameters
- Once fitted, these are much more efficient in terms of storage and computation

Disadvantages: Prior knowledge about the data is necessary since we need to assume an distribution, hence these models rely on very strong (simplistic) distributional assumptions

3 Non Parametric Models

In order to overcome the disadvantage of parametric model, we have the non-parametric model, where this model assumption is not necessary.
Any model that cannot be described by a fixed number of paramters (finite or infinite) are non-parametric models. This does not mean there are non parameters, it just means that the number of parameters are not fixed. Examples: Histogram, Kernel Density Estimator

3.1 Histogram

Given n i.i.d samples $D = \{x^1, x^2, x^3, ..., x^n\}, x^i \in [0, 1)$, we can split these into say m bins-

$$B_1 = \left[0, \frac{1}{m}\right), B_2 = \left[\frac{1}{m}, \frac{2}{m}\right) ..., B_m = \left[\frac{m-1}{m}, 1\right)$$

and then count the number of points that fall in each bin: c_1 points in B_1, c_2 points in $B_2, ...$

The probability density function for this can be written as -

$$p(x) = \Sigma_{j=1}^{m} \frac{mc_j}{n} I \ (x \in B_j)$$

NOTE: $p(x)$ must satisfy:

1. The probabilities must be greater than or equal to zero i.e. $p(x) \geq 0$
2. The sum of probabilities should be 1 i.e $\int_\Omega p(x)dx = 1$
   Verification:

   $$\int_\Omega p(x)dx = \int_{[0,1)} \Sigma_{j=1}^{m} \frac{mc_j}{n} I \ (x \in B_j)dx$$

   Since, x has values on when it lies within the range of the bin and all the bins are disjoint,

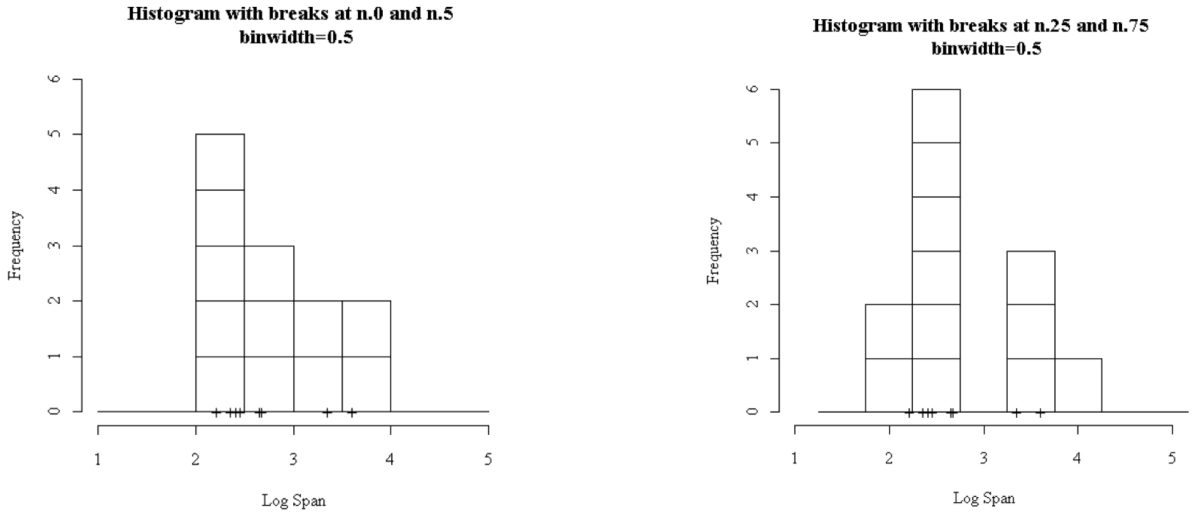
   $$\implies \Sigma_{j=1}^{m} \int_{[\frac{j-1}{m}, \frac{j}{m})} \frac{mc_j}{n} dx = \Sigma_{j=1}^{m} \frac{1}{m} \times \frac{mc_j}{n} = \Sigma_{j=1}^{m} \frac{c_j}{n} = 1$$

3.1.1 Problems with Histogram

1. The density is bin dependent i.e. the output depends on how we define the bins.
   For example, if we consider the test scores example in Fig 1 and change the bin definitions, we get completely different plots as shown in Fig 3
2. Histogram is not very simple when the data has higher dimensions.
   Say, we consider n i.i.d samples of a d dimensional data and split every $[0, 1)^d$ evenly into m^d bins. The bin size $h = \frac{1}{m}$. If $m^d > n$, then some bins will have no samples, thus it is not a good estimator then. Hence, histograms are not preferred for higher dimensional data $(d > 2)$. [Eg: if m=10, n=6, then we need atleast 1 million data samples]
3. Statistically, histogram is not the best.
   The integrated risk can be defined as:

   $$r(\hat{p}, p) = \int_\Re E_X[(\hat{p}(x) - p(x))^2]dx$$

   And this risk changes with n (with bin size $h \approx n^{\frac{-1}{3}}$ as $r(\hat{p}, p) \approx \frac{C}{n^{\frac{2}{3}}}$. This difference increase further for higher dimensional data.



(a) Binwidth = 0.5 with breaks at n.0 and n.5

(b) Binwidth = 0.5 with breaks at n.25 and n.75

Figure 3: Bin Dependency of Histograms

3.2 Kernel Density Estimator

These are an extension to histograms, where there are defined bins, instead, every single location is chosen as a possible bin.
The probability density function is given as:

$$p(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x^i - x}{h}\right)$$

Where $K(.)$nis called the ”Smoothing Kernel Function”. It can be any function, provided it satisfies certain conditions.

3.2.1 Conditions to be satisfied by the Smoothing Kernel Function

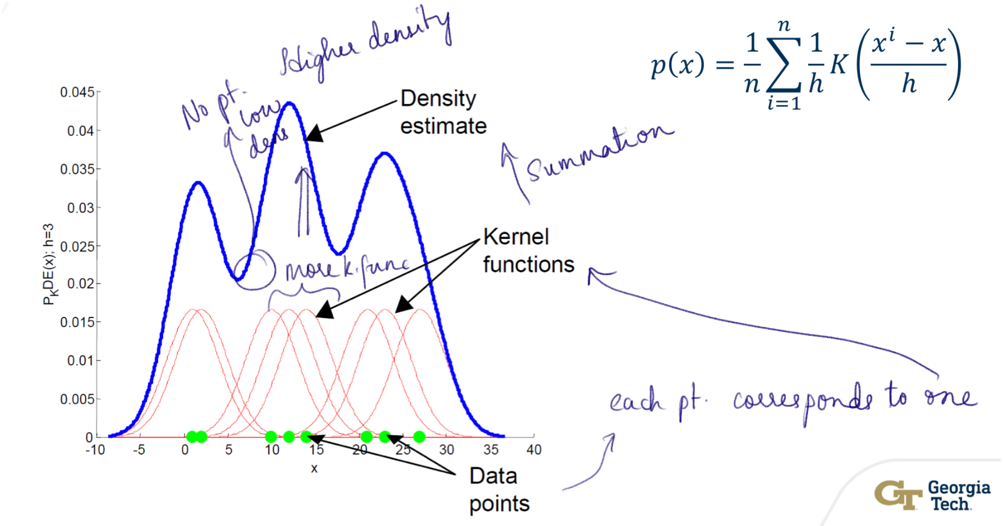
1. It must be non-negative at all points i.e. $K(u) \geq 0$
2. It must be symmetric
3. $\int K(u)du = 1$
4. $\int uK(u)du = 0$
5. $\int u^2 K(u)du < \infty$

Example: Gaussian Kernel $K(u) = \frac{1}{\sqrt{2\pi}} e^{\frac{-u^2}{2}}$

3.2.2 Important points about the Kernel Density Estimator

- The choice of $K(.)$ requires some prior knowledge about the data
- Unline parametric models, different smoothing functions can be used for different points

- The data point is considered to be at the center of the function and the function is symmetric about this point (Fig 4)
- To get the overall density estimate of the data, the individual kernal functions are added up (Fig 4)



$$p(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x^i - x}{h}\right)$$

Figure 4: Density estimation using Kernel Density Estimator

- Lower the Kernel Bandwidth h, higher the granularity in the estimate
  Silverman's rule of thumb: If using Gaussian kernel, a good choice is $h \approx 1.06\hat{\sigma}n^{\frac{-1}{5}}$, where $\hat{\sigma}$ is the standard deviation of the samples
  A better but more computationally intensive approach is:
  - Randomly split the data into two sets
  - Obtain a kernel density estimate for the first set
  - Measure the likelihood of the second set
  - Repeat over many random splits and take the average
- Depending on h, we have or not have probability at a point (Fig 4)
- Higher the number of kernels, better the density estimate (Fig 4). However, caution should practiced to not overfit

3.3 Advantages and Disadvantages of Non Parametric Models

Advantages:

- Very mild assumptions on data distribution
- Usually better models for complex data

Disadvantages: Nonparametric models (not histograms) requires storing and computing with the entire data set

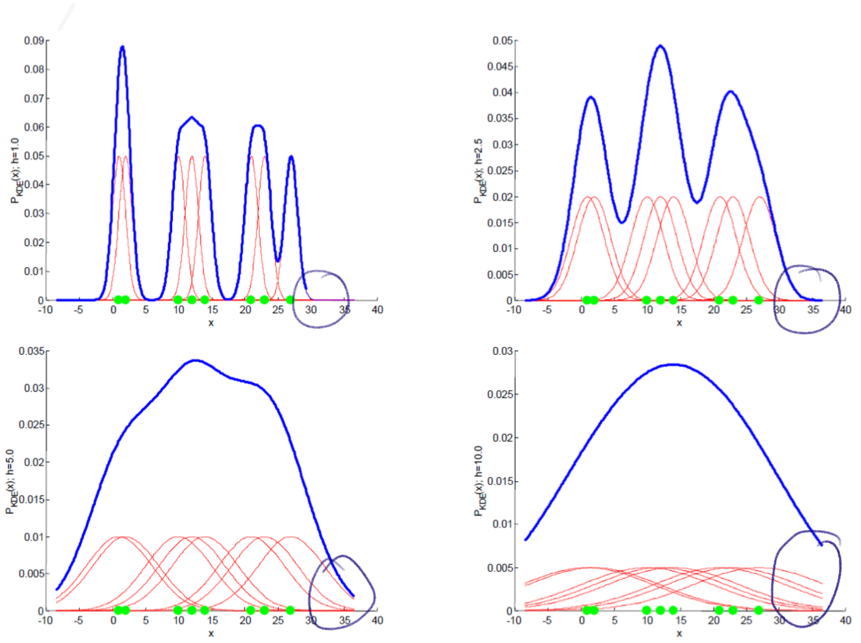


Figure 5: Effect of kernel bandwith h

4 Parametric vs Non Parametric Model

Consider the data $x \in \mathbb{R}^d$ with fixed dimension d, with n training data points $\{x^1, x^2, \ldots, x^n\}$ and partitioned into m bins in each dimension -

Aspects	**Gaussian**	**Histogram**	**KDE**
Flexible	No	Yes	Yes
Assumption	Strong	Mild	Mild
Parameter number	Fixed	Increased with m	Increased with n
Memory requirement	$d + d^2$	m^d	nd
Training computation	Closed form	Binning and counting	Nothing
Test computation	Plug in formula	Find the bin	Evaluate n functions
Statistical guarantee	Only Gaussian case	Arbitrary (worse)	Arbitrary (better)

REFERENCES

All pics and information are from the lecture and lecture slides (Lecture 4).
Used Perplexity AI for some formatting (Overleaf).