Fundamentals of AI

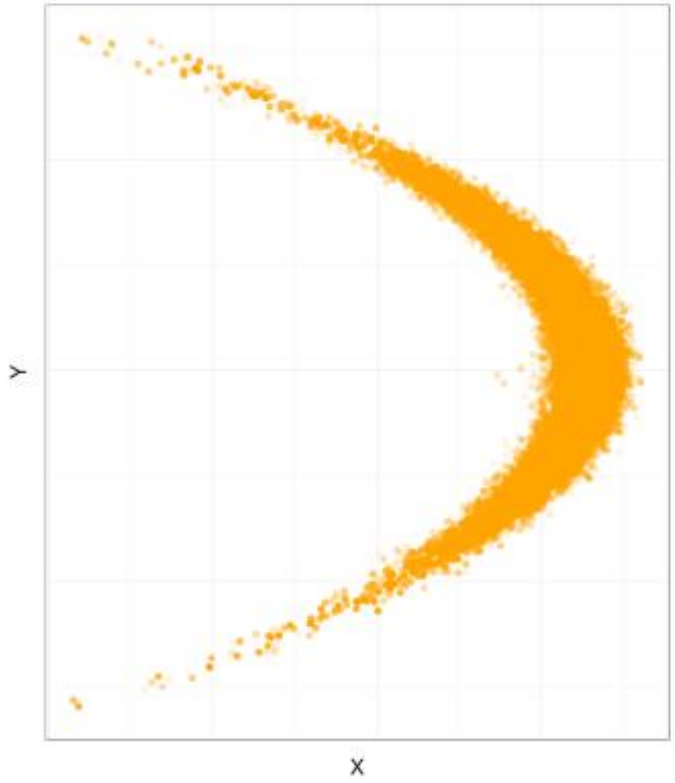
Introduction and the most basic concepts

Probability Density Function (PDF)

Joint Probability Distribution

- Probability of any combination of features to happen
- Fundamental assumption: dataset is i.i.d. (Independent and identically distributed) sample following PDF
- If we know PDF underlying our dataset then we can predict everything (any dependence, together with uncertainties)!
- Moreover, knowing PDF we can generate infinite number of similar datasets with the same or different number of points
- *Really Platonian thing!*

'Banana-shaped probability distribution'



Probability density function (PDF)

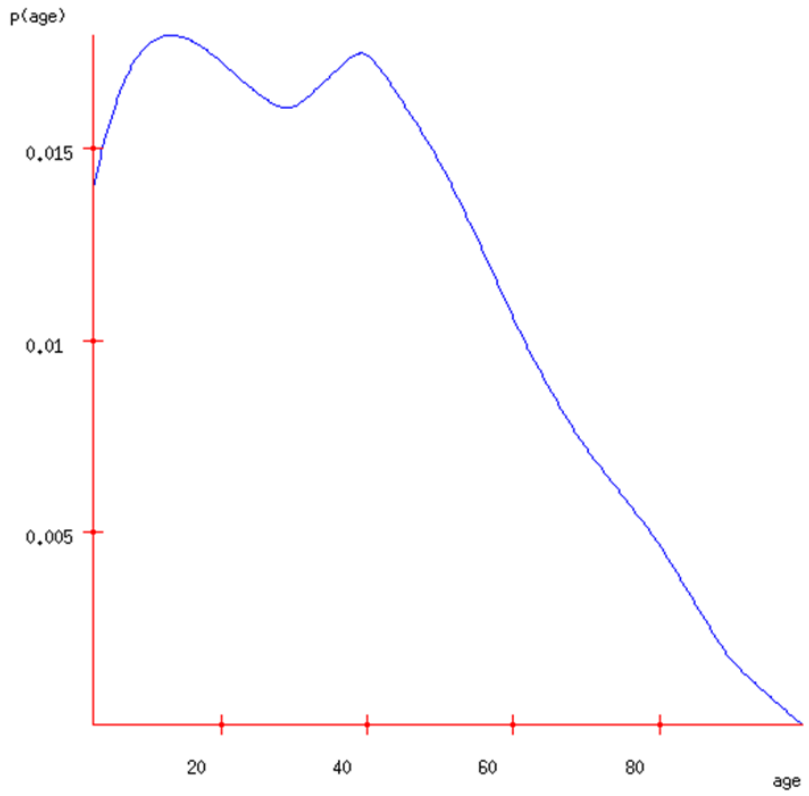
$$f(x,y) = \exp\left(-\frac{x^2}{200} - \frac{1}{2}(y + Bx^2 - 100B)^2\right)$$

Probability Density Function

- PDF is a way to define joint probability distribution for features with continuous (numerical) values
- Can immediately get us Bayesian methods that are sensible with real-valued data
- You'll need to intimately understand PDFs in order to do kernel methods, clustering with Mixture Models, analysis of variance, time series and many other things
- Will introduce us to linear and non-linear regression

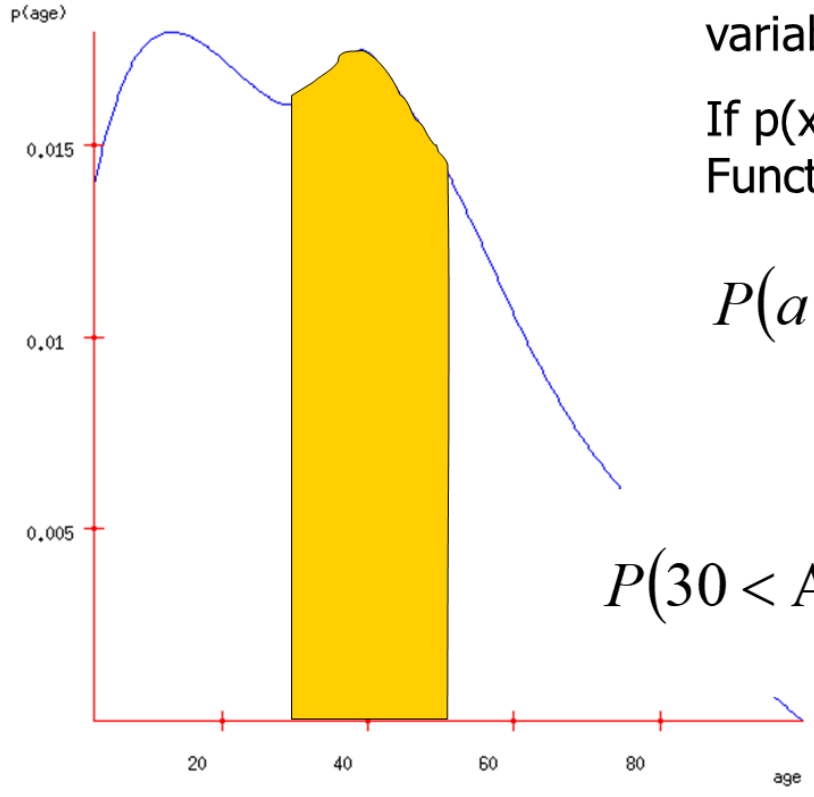
Example of a 1D PDF

A PDF of American Ages in 2000



Example of a 1D PDF

A PDF of American Ages in 2000



Let X be a continuous random variable.

If p(x) is a Probability Density Function for X then…

$$P(a < X \leq b) = \int_{x=a}^{b} p(x)dx$$

$$P(30 < \text{Age} \leq 50) = \int_{\text{age}=30}^{50} p(\text{age})d\text{age}$$

= 0.36

What's the meaning of p(x)?

If

p(5.31) = 0.06 and p(5.92) = 0.03

then

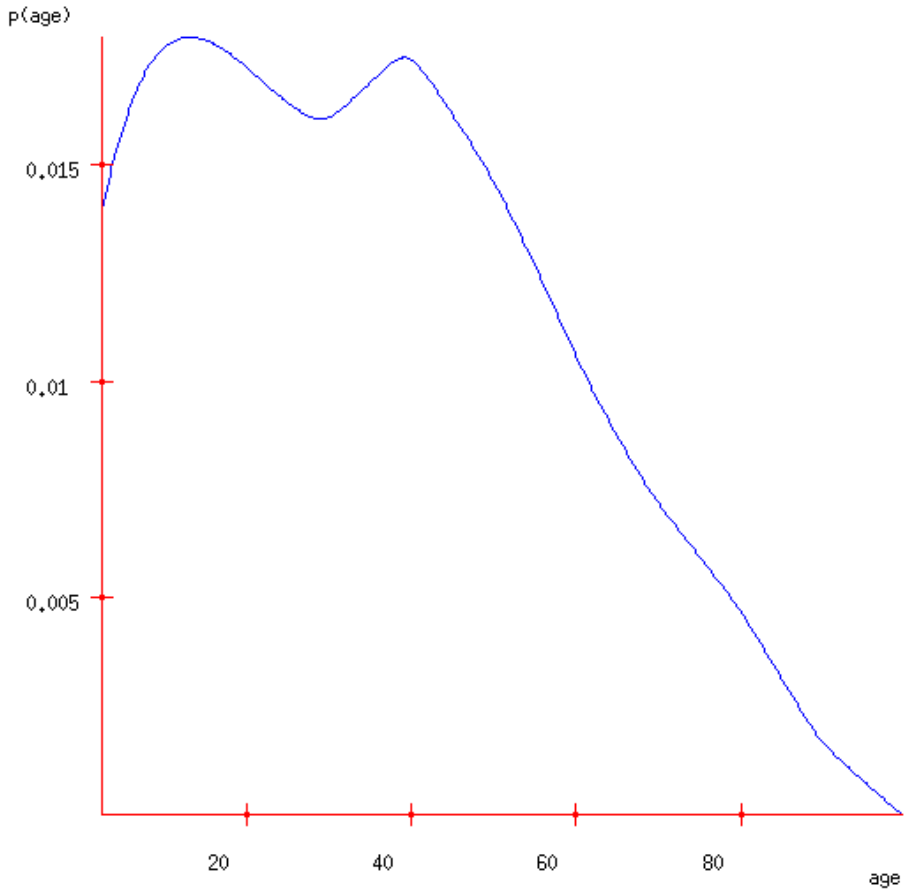
when a value X is sampled from the distribution, you are 2 times as likely to find that X is "very close to" 5.31 than that X is "very close to" 5.92.

True or False?

$$\forall x : p(x) \leq 1$$ **TRUE**

$$\forall x : P(X = x) = 0$$ **TRUE**

Expectations (aka mean value)

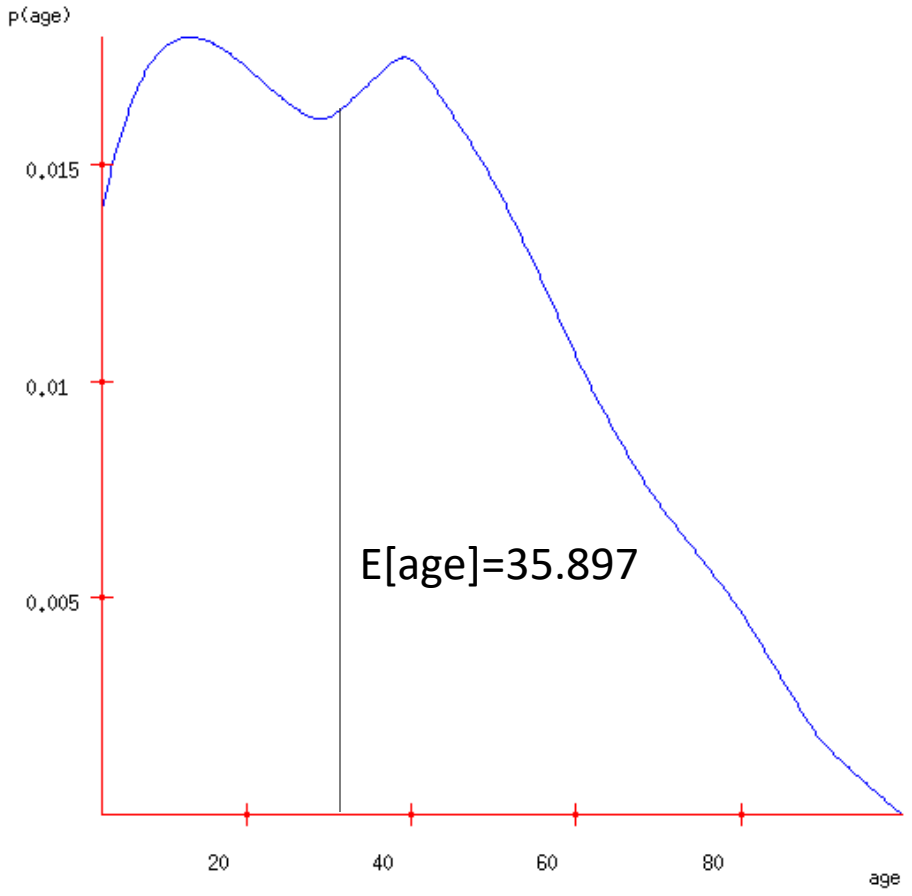


E[X] = the expected value of random variable X

= the average value we'd see if we took a very large number of random samples of X

$$= \int_{x=-\infty}^{\infty} x\, p(x)\, dx$$

Expectations



E[X] = the **expected value** of random variable X

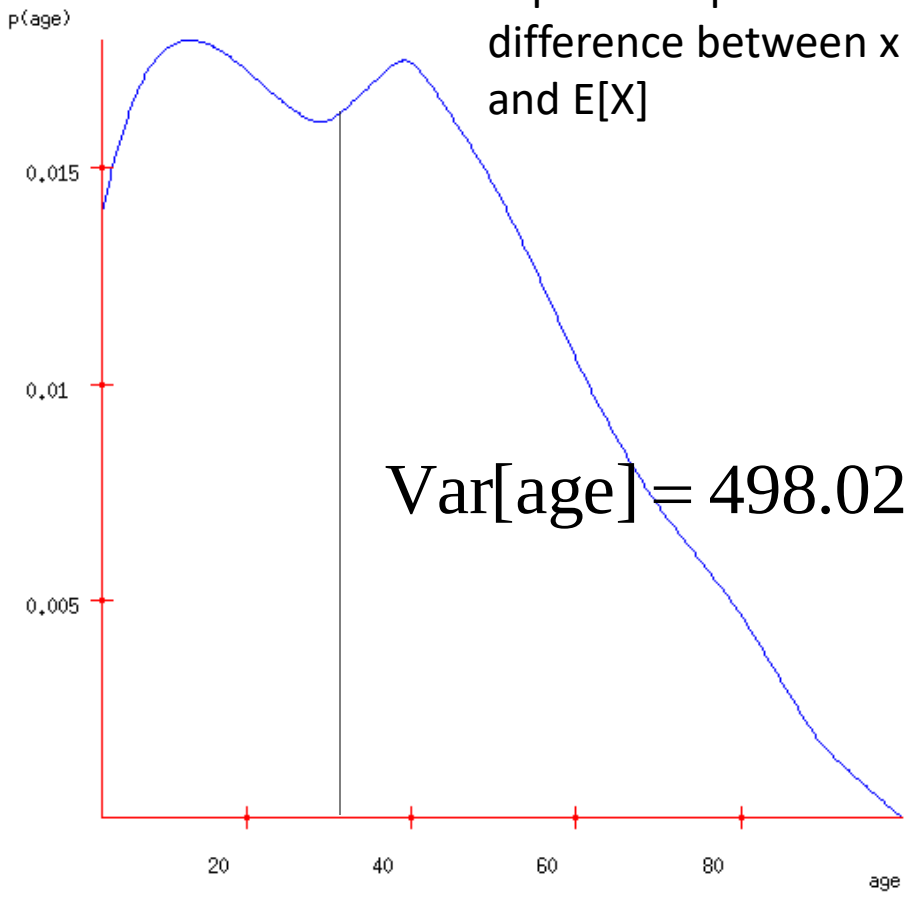
= the **average value** we'd see if we took a very large number of random samples of X

$$= \int_{x=-\infty}^{\infty} x\, p(x)\, dx$$

= the **first moment** of the shape formed by the axes and the blue curve

= the **best value** to choose if you must guess an unknown person's age and you'll be fined the square of your error

Variance

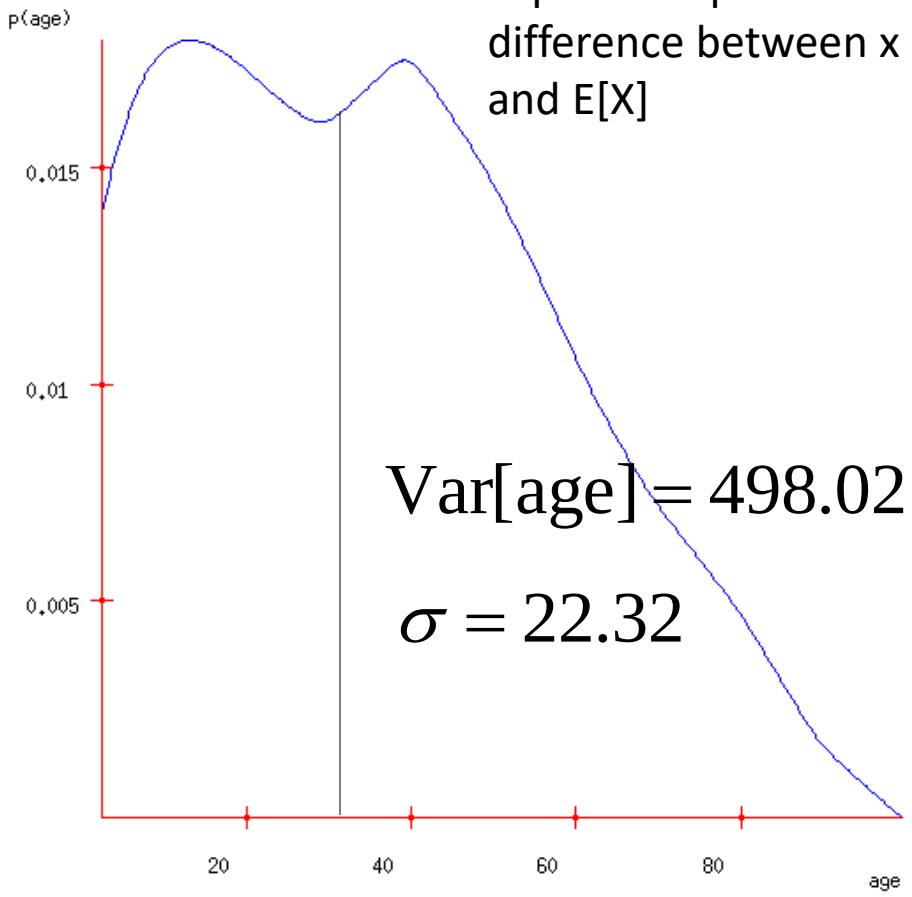


σ^2 = Var[X] = the expected squared difference between x and E[X]

$$\sigma^2 = \int_{x=-\infty}^{\infty} (x-\mu)^2 \, p(x)\, dx$$

= amount you'd expect to lose if you must guess an unknown person's age and you'll be fined the square of your error, and assuming you play optimally

Standard Deviation



σ^2 = Var[X] = the expected squared difference between x and E[X]

$$\text{Var}[\text{age}] = 498.02$$

$$\sigma = 22.32$$

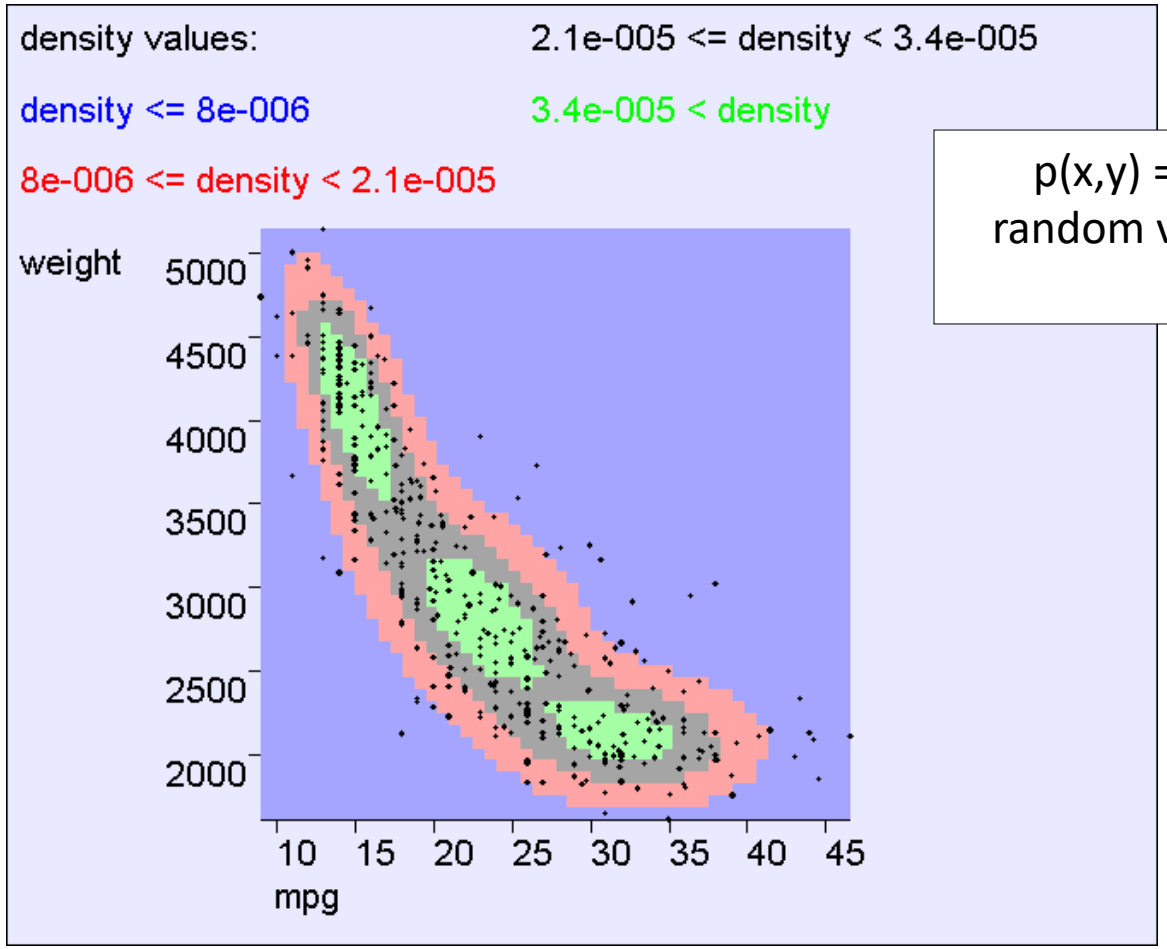
$$\sigma^2 = \int_{x=-\infty}^{\infty} (x-\mu)^2 \, p(x)\, dx$$

= amount you'd expect to lose if you must guess an unknown person's age and you'll be fined the square of your error, and assuming you play optimally

σ = Standard Deviation = "typical" deviation of X from its mean

$$\sigma = \sqrt{\text{Var}[X]}$$

In 2 dimensions

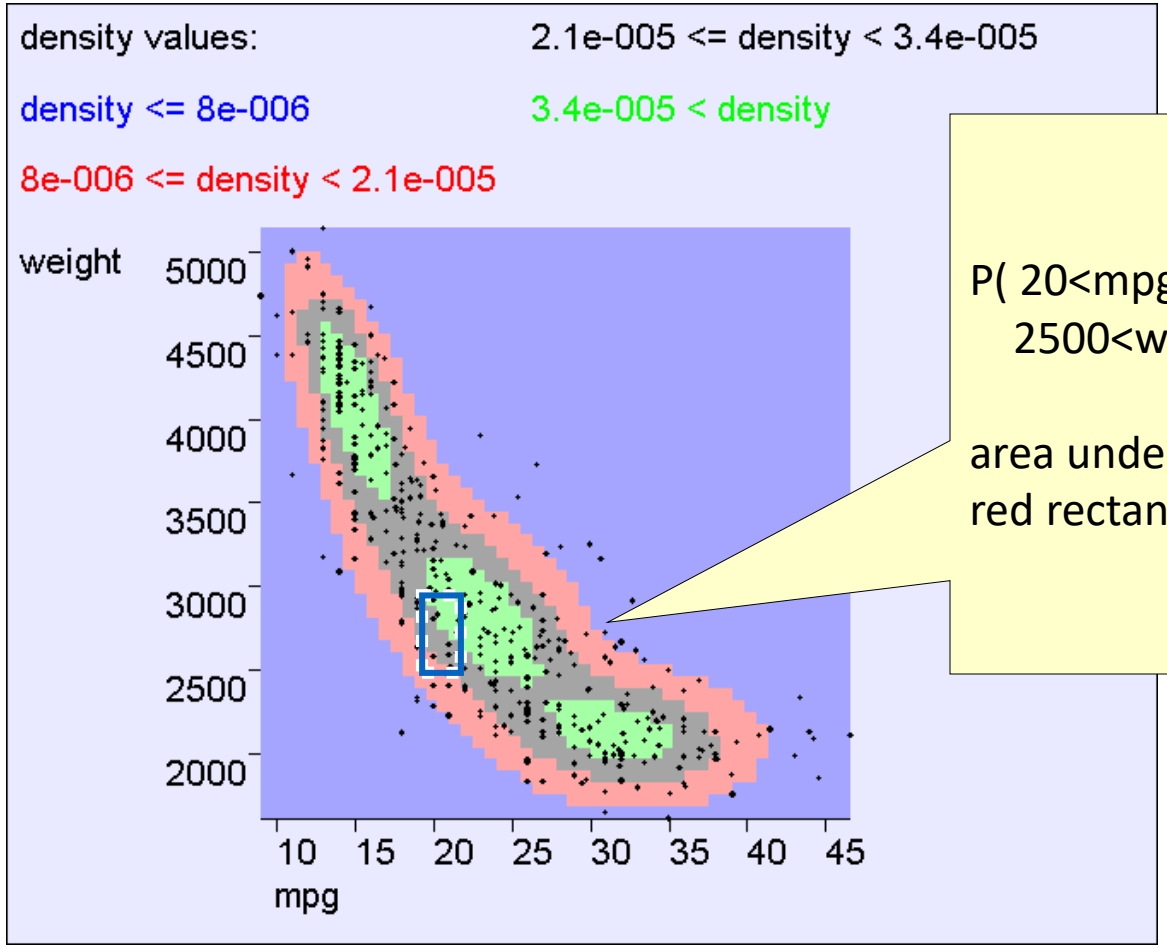


p(x,y) = probability density of random variables (X,Y) at location (x,y)

In 2 dimensions

Let *X,Y* be a pair of continuous random variables, and let R be some region of (X,Y) space…

$$P((X,Y) \in R) = \iint_{(x,y)\in R} p(x,y)dydx$$

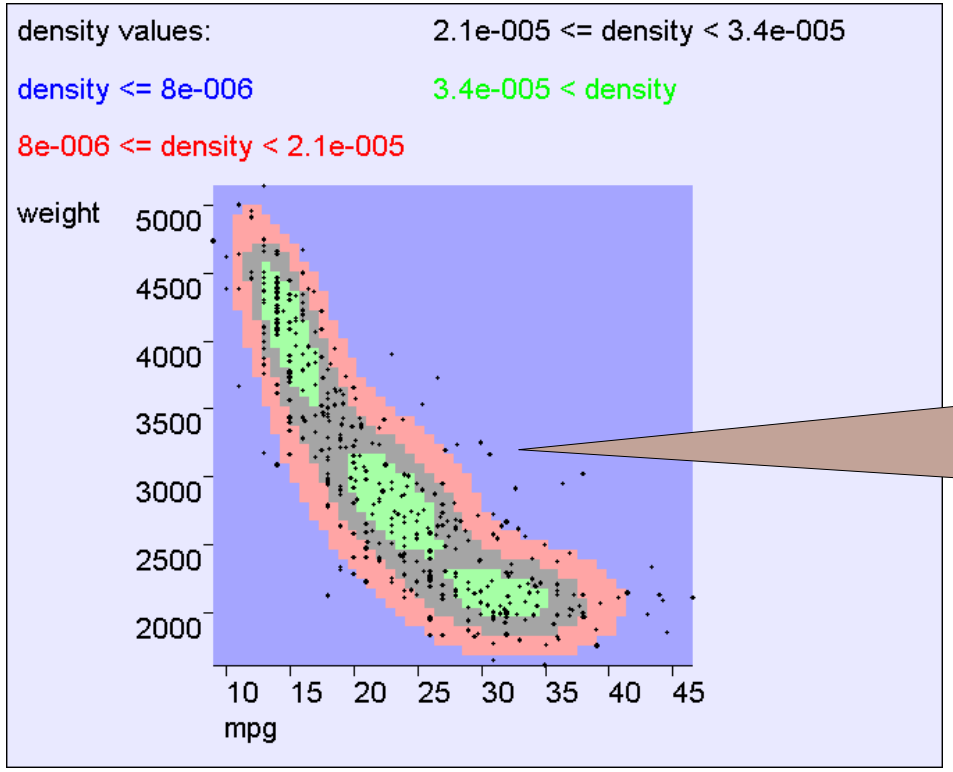


P(20<mpg<30 and
2500<weight<3000) =

area under the 2-d surface within the red rectangle

Independence

$$X \perp Y \text{ iff } \forall x, y : p(x, y) = p(x)\,p(y)$$

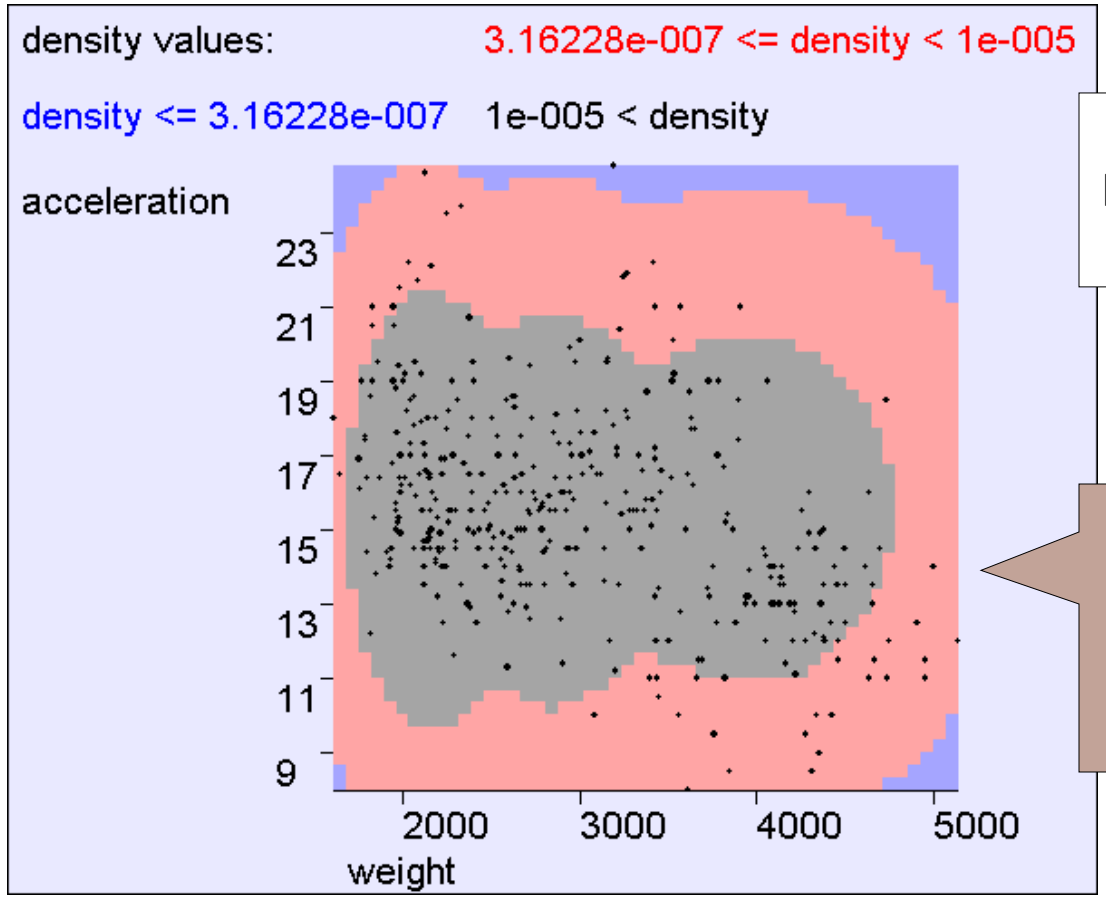


If X and Y are independent then knowing the value of X does not help predict the value of Y

mpg,weight NOT independent

Independence

$$X \perp Y \text{ iff } \forall x, y : p(x, y) = p(x)\,p(y)$$

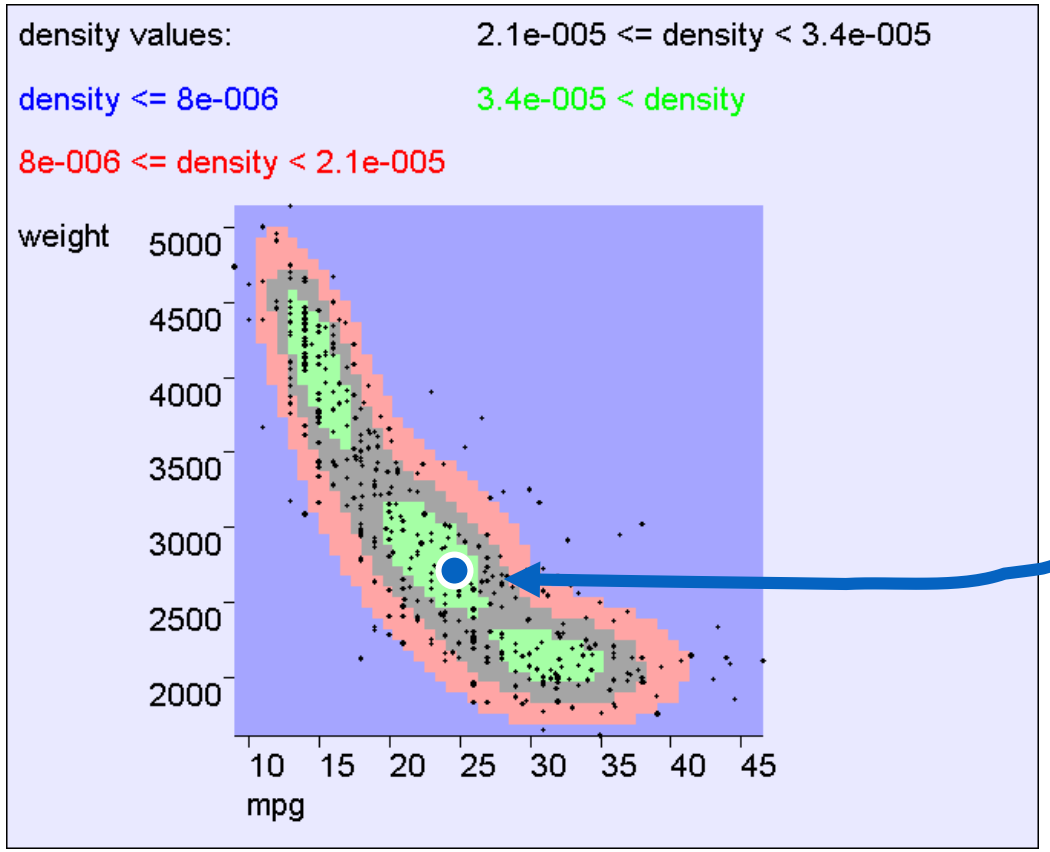


If X and Y are independent then knowing the value of X does not help predict the value of Y

the contours say that acceleration and weight are independent

Multivariate Expectation

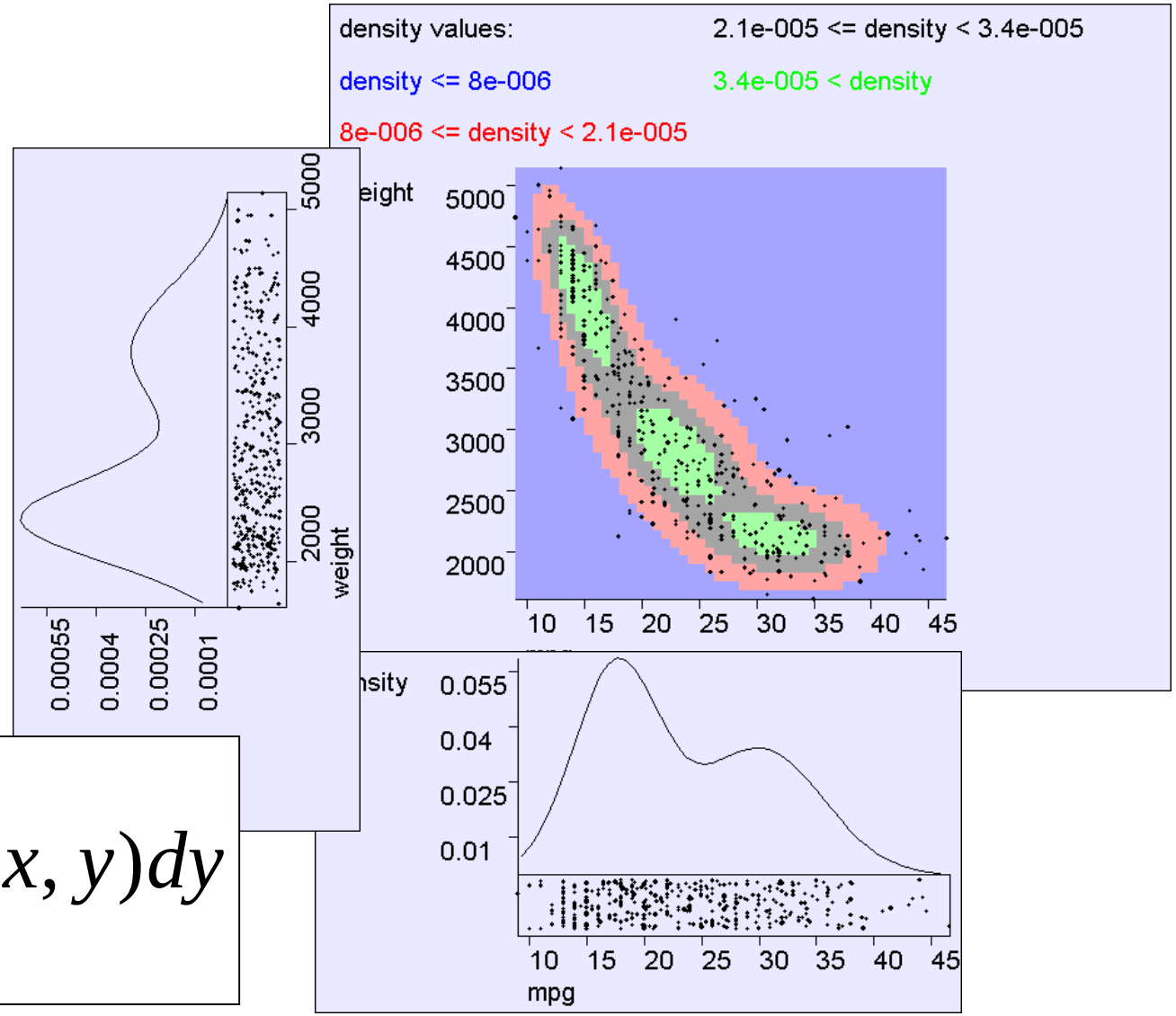
$$\boldsymbol{\mu}_{\mathbf{X}} = E[\mathbf{X}] = \int \mathbf{x}\ p(\mathbf{x}) d\mathbf{x}$$



E[mpg,weight] = (24.5,2600)

The centroid of the cloud

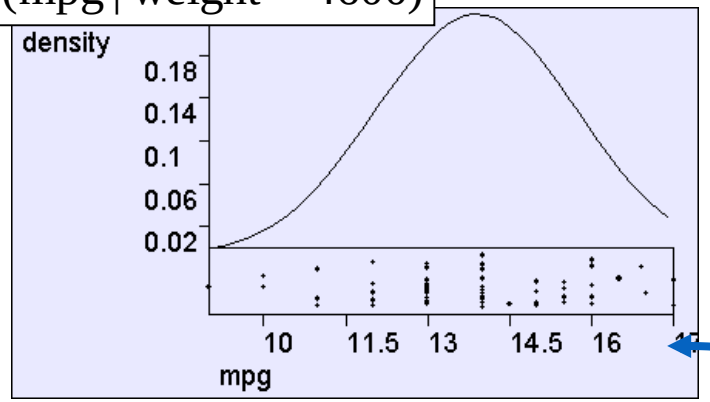
Marginal Distributions



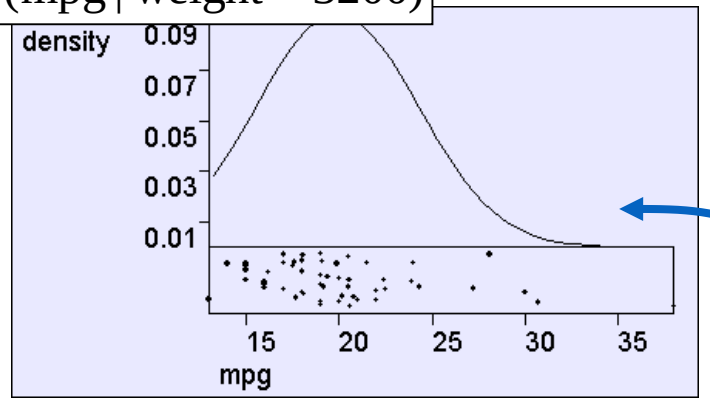
$$p(x) = \int_{y=-\infty}^{\infty} p(x, y)dy$$

Conditional Distributions

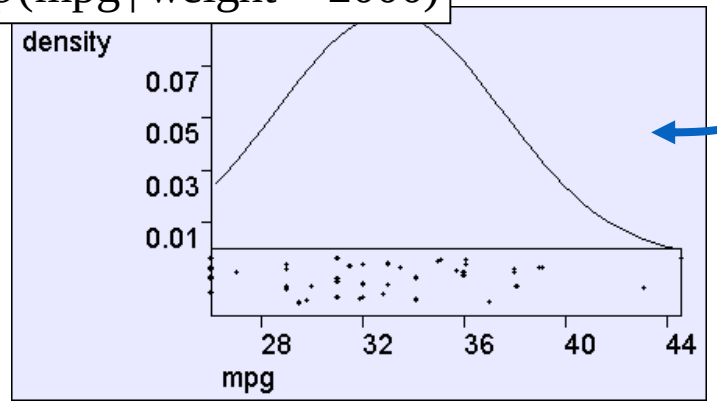
$p(\text{mpg} \mid \text{weight} = 4600)$

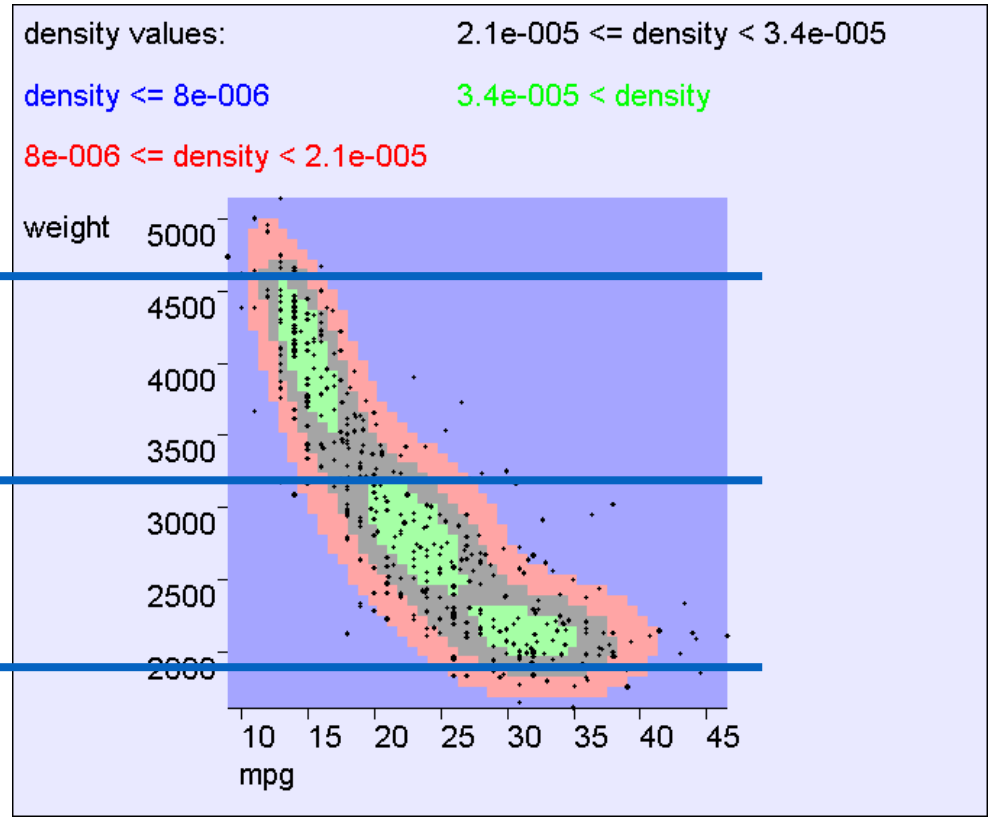


$p(\text{mpg} \mid \text{weight} = 3200)$



$p(\text{mpg} \mid \text{weight} = 2000)$



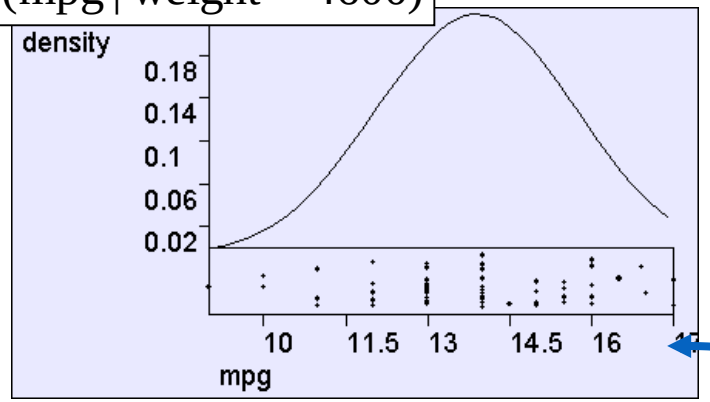


$$p(x \mid y) =$$

$$\text{p.d.f. of } X \text{ when } Y = y$$

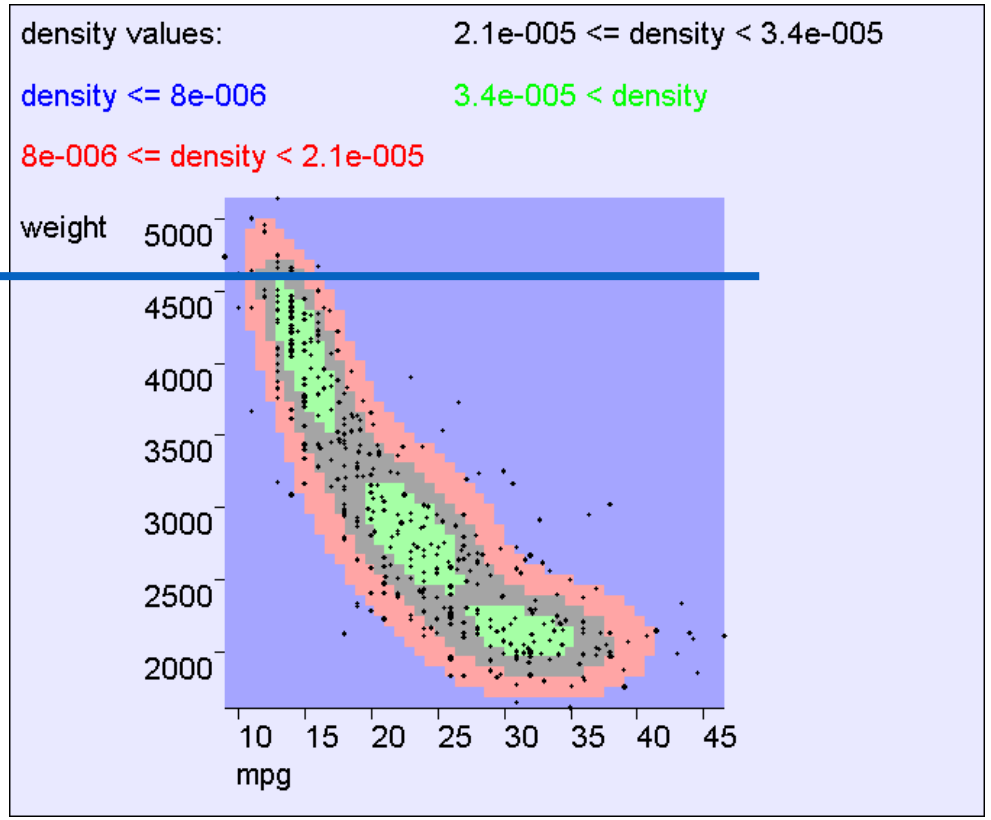
Conditional Distributions

$p(\text{mpg} \mid \text{weight} = 4600)$



$$p(x \mid y) = \frac{p(x, y)}{p(y)}$$

Why?



$$p(x \mid y) =$$

$$\text{p.d.f. of } X \text{ when } Y = y$$

Gaussian (normal) distribution

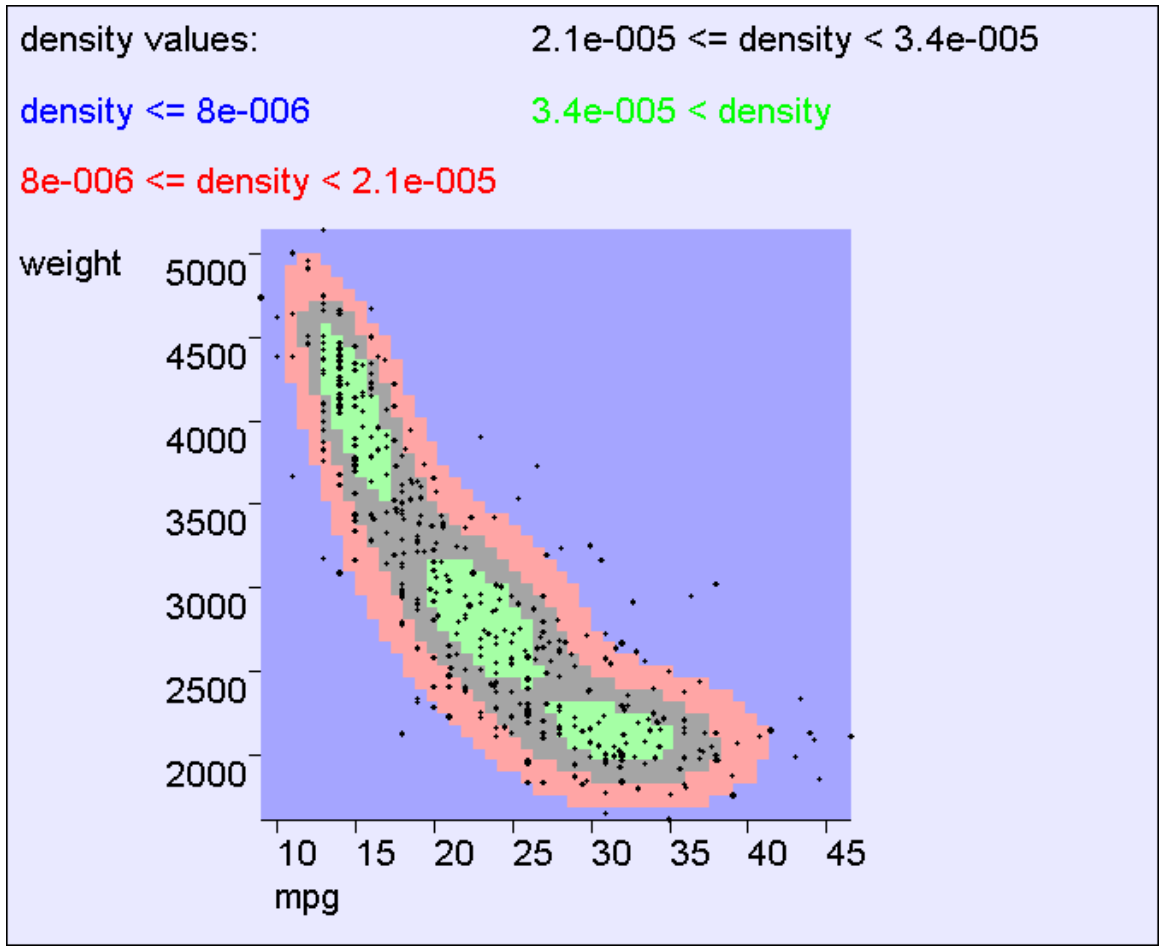
- The most used PDF
- Most of the classical statistical learning theory is based on Gaussians
- Connection to the mean-squared loss
- Connection with linearity
- Connection with Euclidean space
- Connection to a mean of (many) independent variables
- Distribution with the largest entropy among all distributions with unit variance
- Mixture of Gaussians can approximate (almost) everything

Gaussian (normal) distribution

- The most used PDF
- Most of the classical statistical learning [illegible] on Gaussians
- Connection to the mean-squared [illegible]
- Connection with linearity
- Connection with Eucl[illegible]
- Connection to [illegible] independent variables
- Distributi[illegible]gest entropy among all distributions with unit varia[illegible]
- Mix[illegible]ssians can approximate (almost) everything

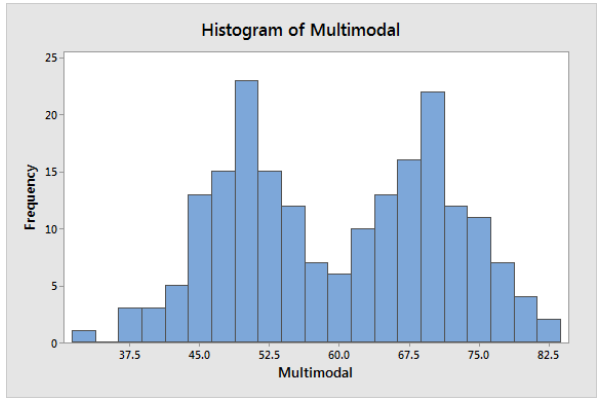
Please revise your knowledge about multivariate Gaussian distribution

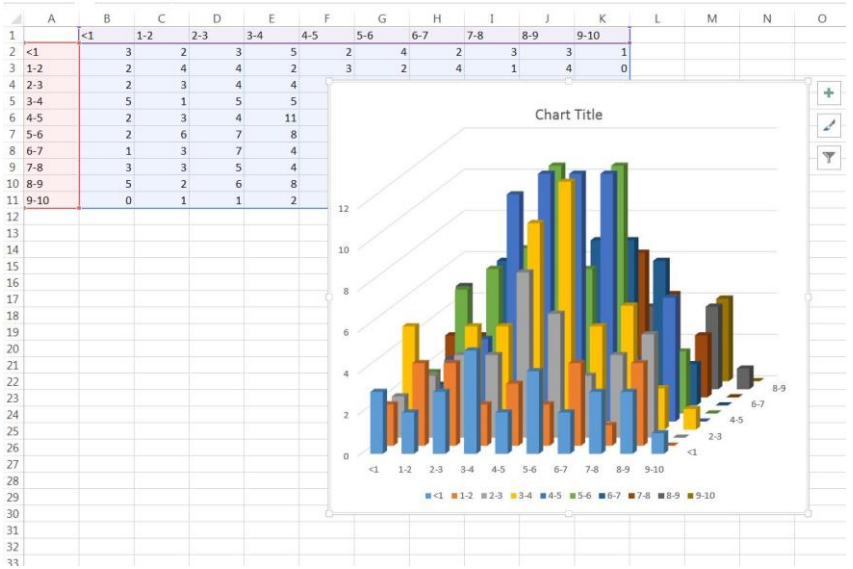
The dataset is a finite set of points.
The PDF is continuous. How this is possible?



Learning PDF from data

- Part of unsupervised machine learning
- Histograms and multi-dimensional histograms



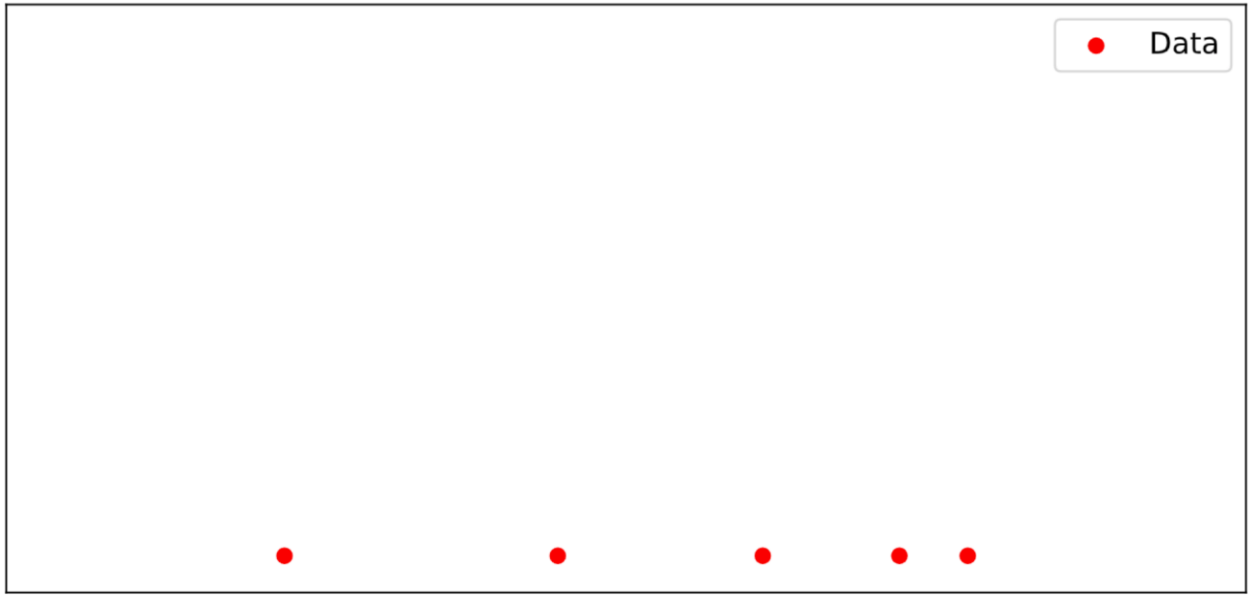


- Naïve Bayes : P(X,Y,Z,T) = P(X)P(Y)P(Z)P(T)
- Bayesian networks, graphical models
- Kernel density estimate

Estimating PDF from data: Kernel Density Estimate

On every data point x_i, we place a kernel function K. The kernel density estimate is

$$\hat{f}(x) = \frac{1}{N} \sum_{i=1}^{N} K(x - x_i).$$

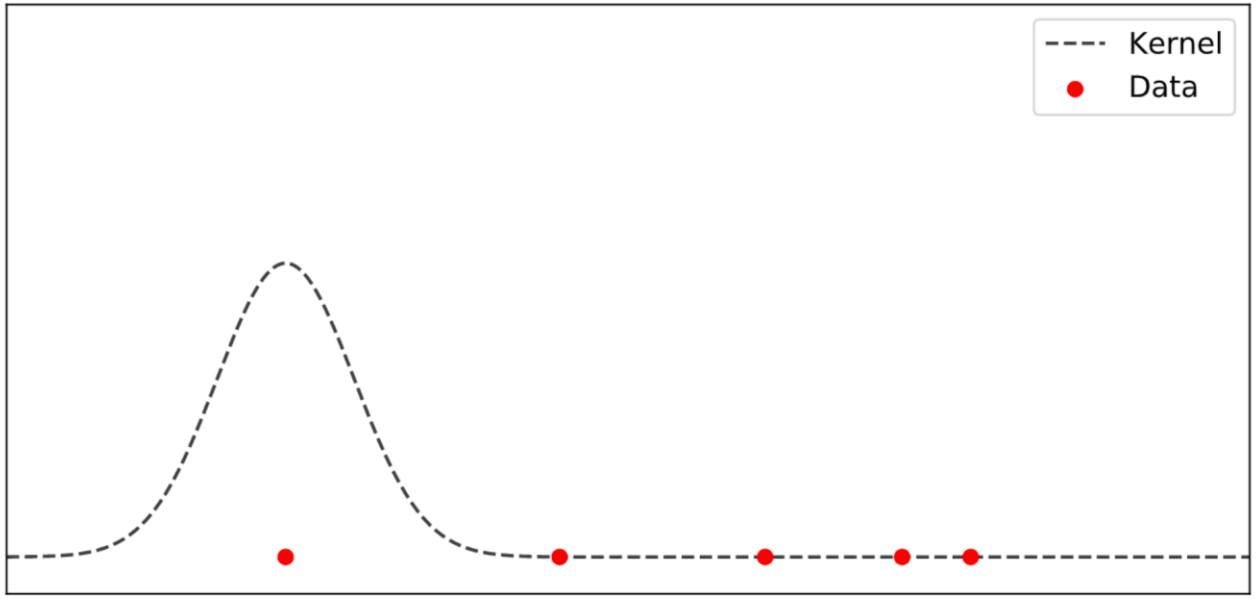


https://www.youtube.com/watch?v=gPWsDh59zdo

Estimating PDF from data: Kernel Density Estimate

On every data point x_i, we place a kernel function K. The kernel density estimate is

$$\hat{f}(x) = \frac{1}{N} \sum_{i=1}^{N} K(x - x_i).$$

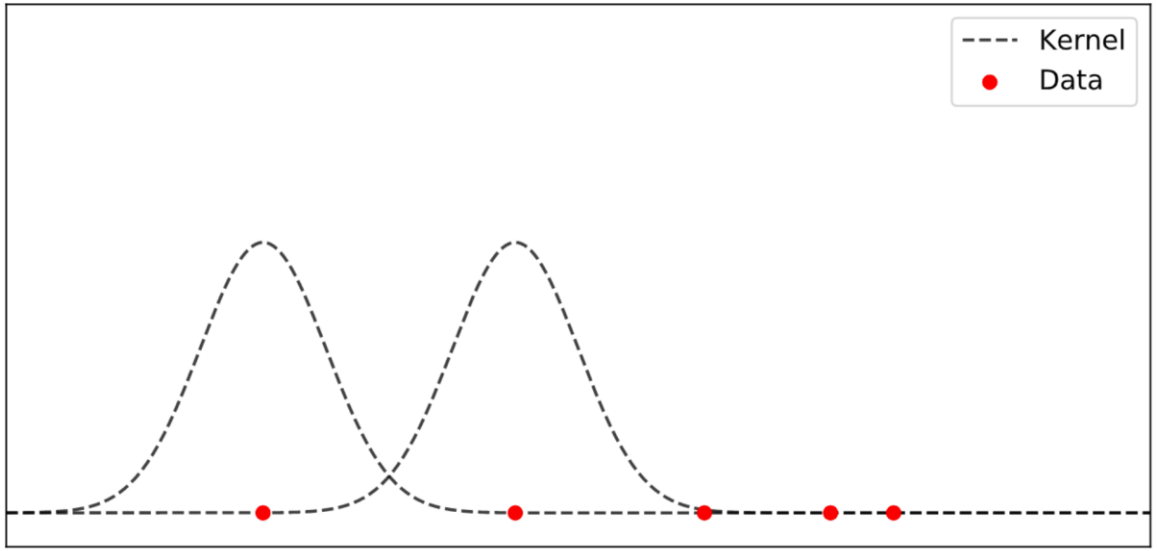


https://www.youtube.com/watch?v=gPWsDh59zdo

Estimating PDF from data: Kernel Density Estimate

On every data point x_i, we place a kernel function K. The kernel density estimate is

$$\hat{f}(x) = \frac{1}{N} \sum_{i=1}^{N} K(x - x_i).$$

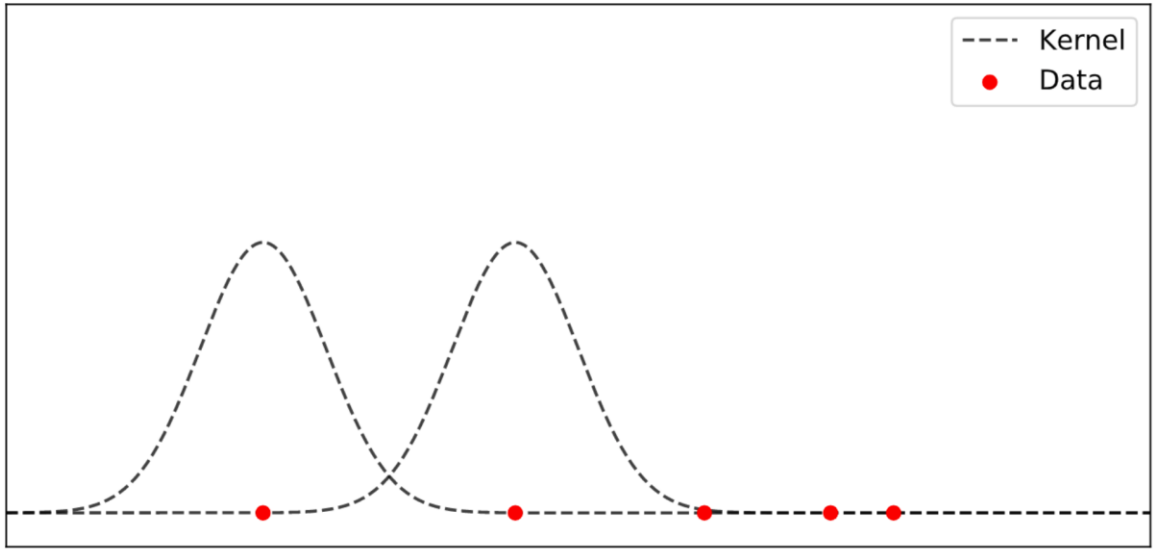


https://www.youtube.com/watch?v=gPWsDh59zdo

Estimating PDF from data: Kernel Density Estimate

On every data point x_i, we place a kernel function K. The kernel density estimate is

$$\hat{f}(x) = \frac{1}{N} \sum_{i=1}^{N} K(x - x_i).$$

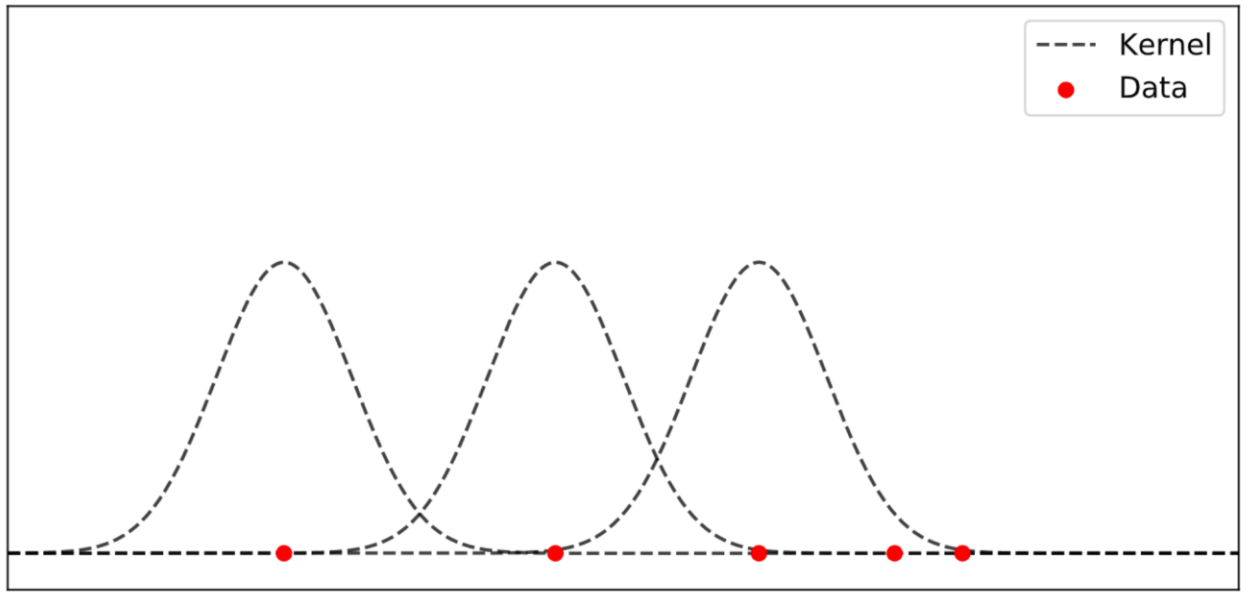


https://www.youtube.com/watch?v=gPWsDh59zdo

Estimating PDF from data: Kernel Density Estimate

On every data point x_i, we place a kernel function K. The kernel density estimate is

$$\hat{f}(x) = \frac{1}{N} \sum_{i=1}^{N} K(x - x_i).$$

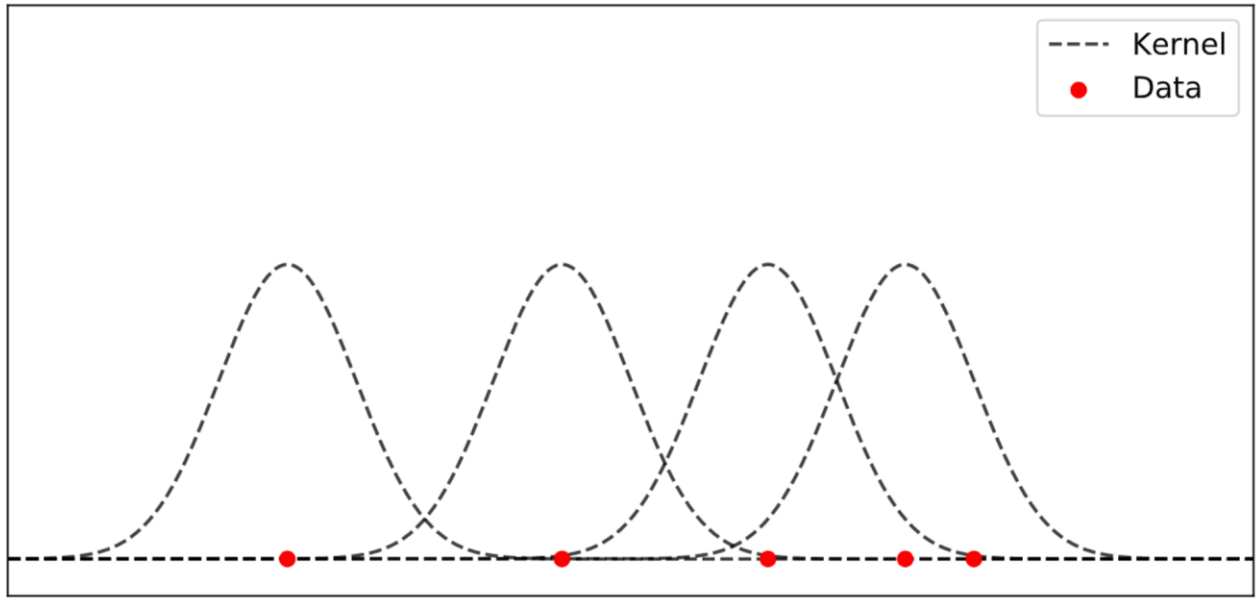


https://www.youtube.com/watch?v=gPWsDh59zdo

Estimating PDF from data: Kernel Density Estimate

On every data point x_i, we place a kernel function K. The kernel density estimate is

$$\hat{f}(x) = \frac{1}{N} \sum_{i=1}^{N} K(x - x_i).$$

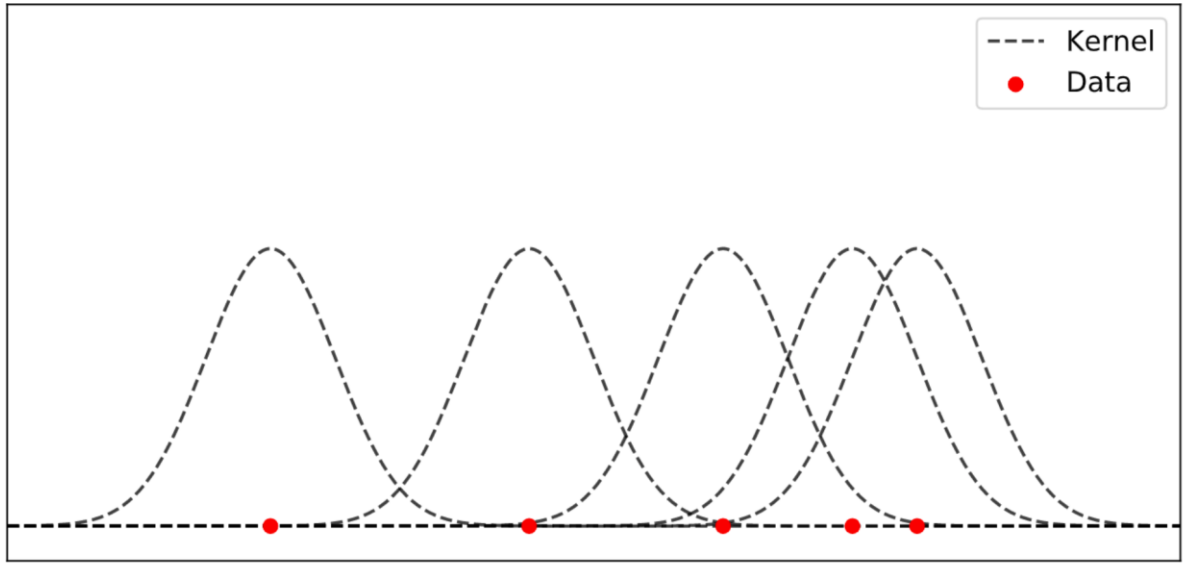


https://www.youtube.com/watch?v=gPWsDh59zdo

Estimating PDF from data: Kernel Density Estimate

On every data point x_i, we place a kernel function K. The kernel density estimate is

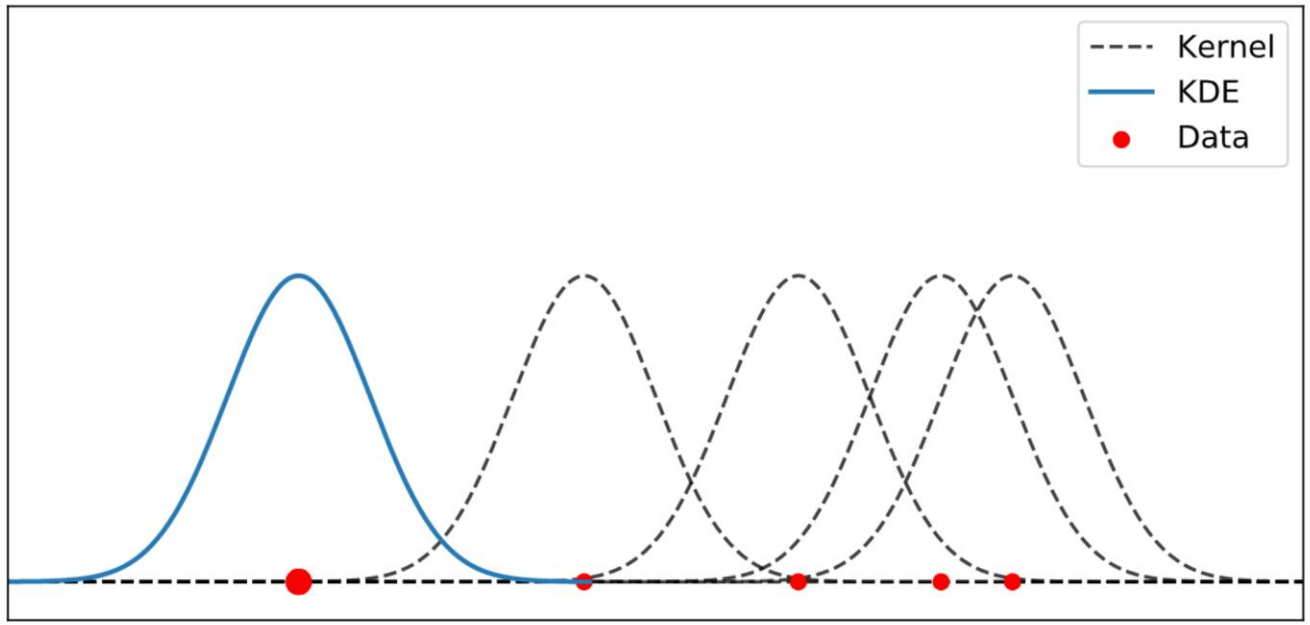
$$\hat{f}(x) = \frac{1}{N} \sum_{i=1}^{N} K(x - x_i).$$



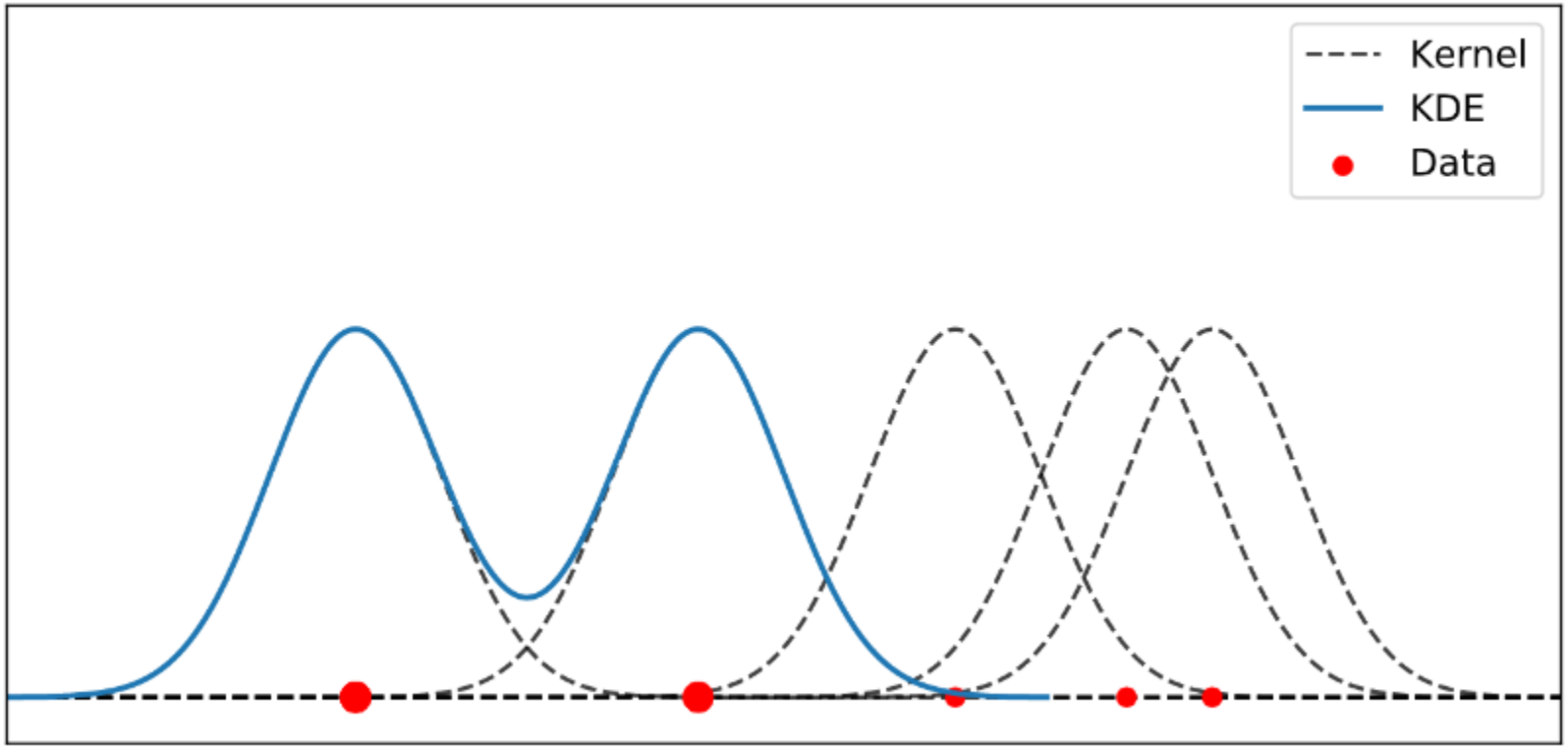
Estimating PDF from data: Kernel Density Estimate

On every data point x_i, we place a kernel function K. The kernel density estimate is

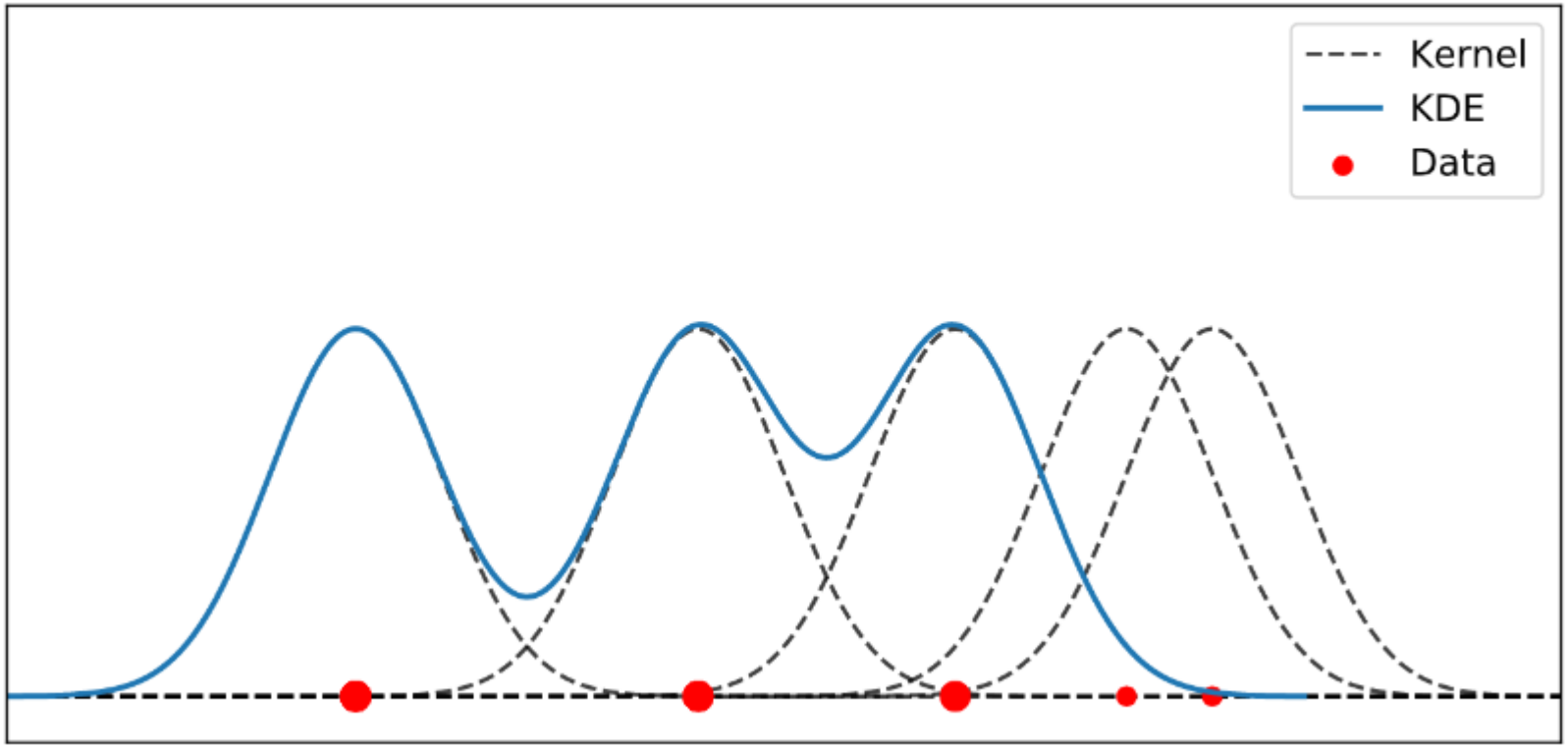
$$\hat{f}(x) = \frac{1}{N} \sum_{i=1}^{N} K(x - x_i).$$



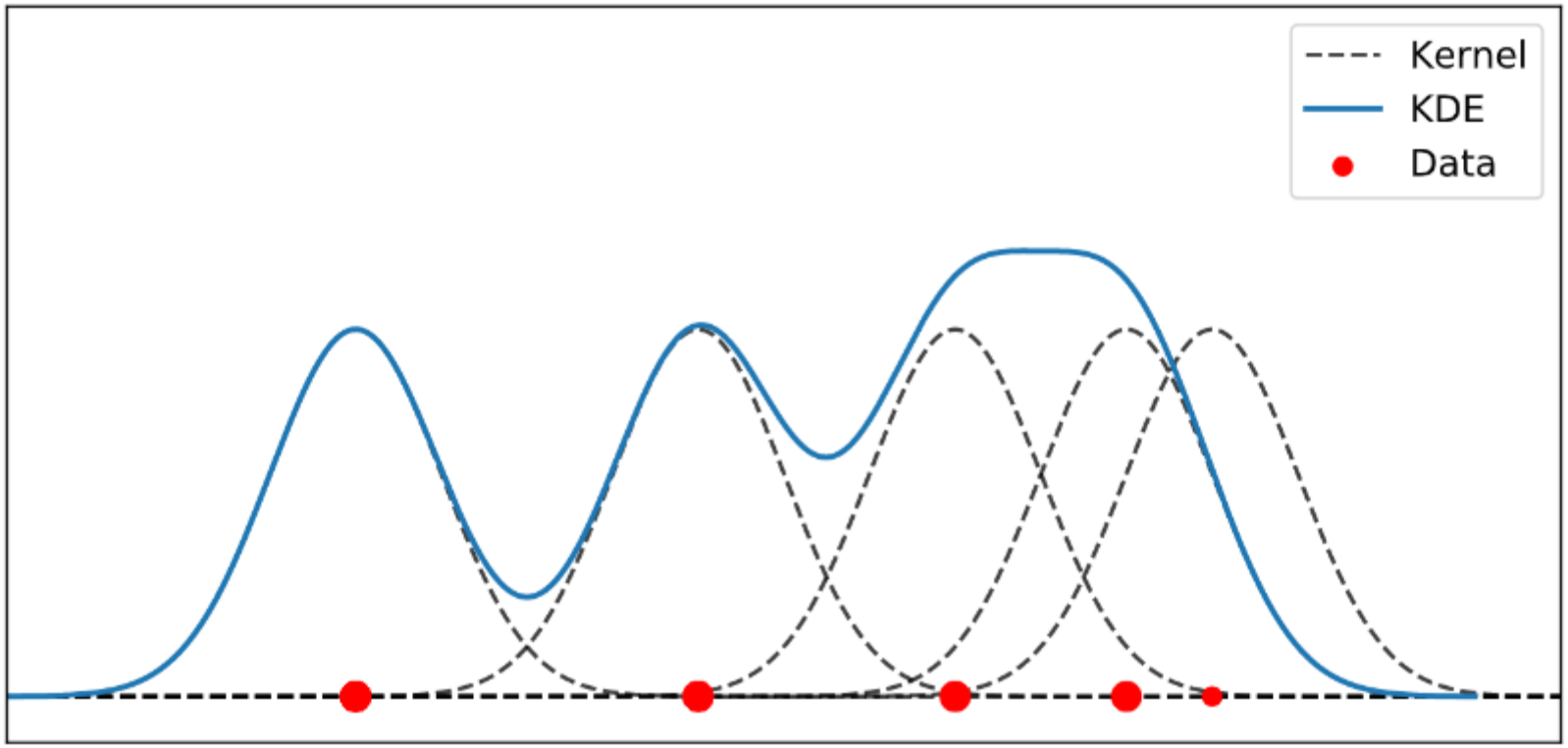
Estimating PDF from data: Kernel Density Estimate



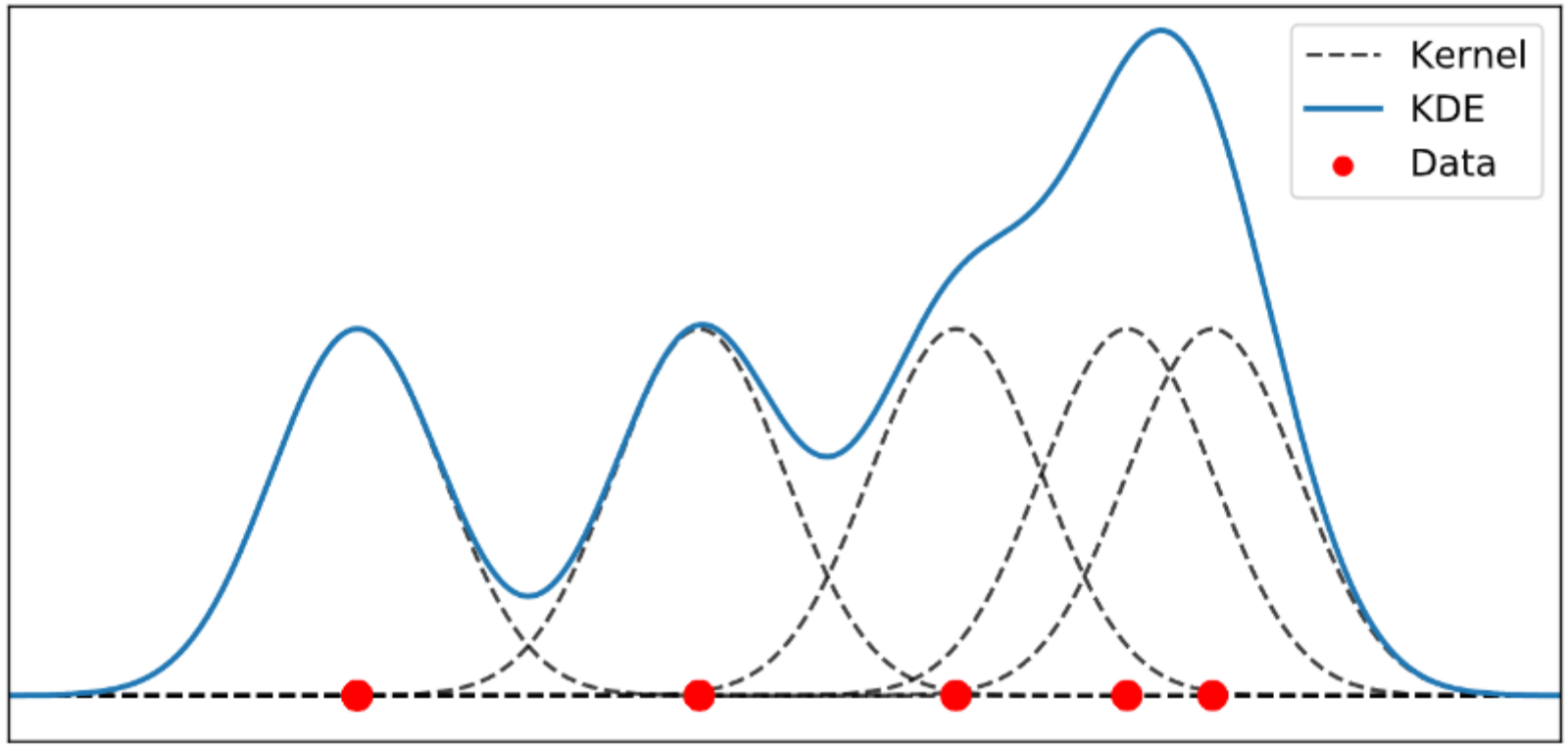
Estimating PDF from data: Kernel Density Estimate



Estimating PDF from data: Kernel Density Estimate



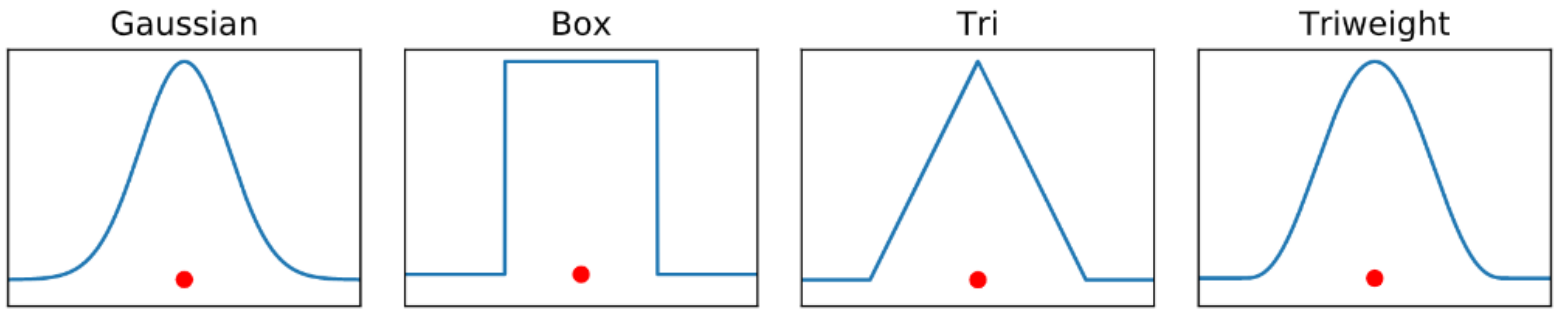
Estimating PDF from data: Kernel Density Estimate



Estimating PDF from data: Kernel Density Estimate

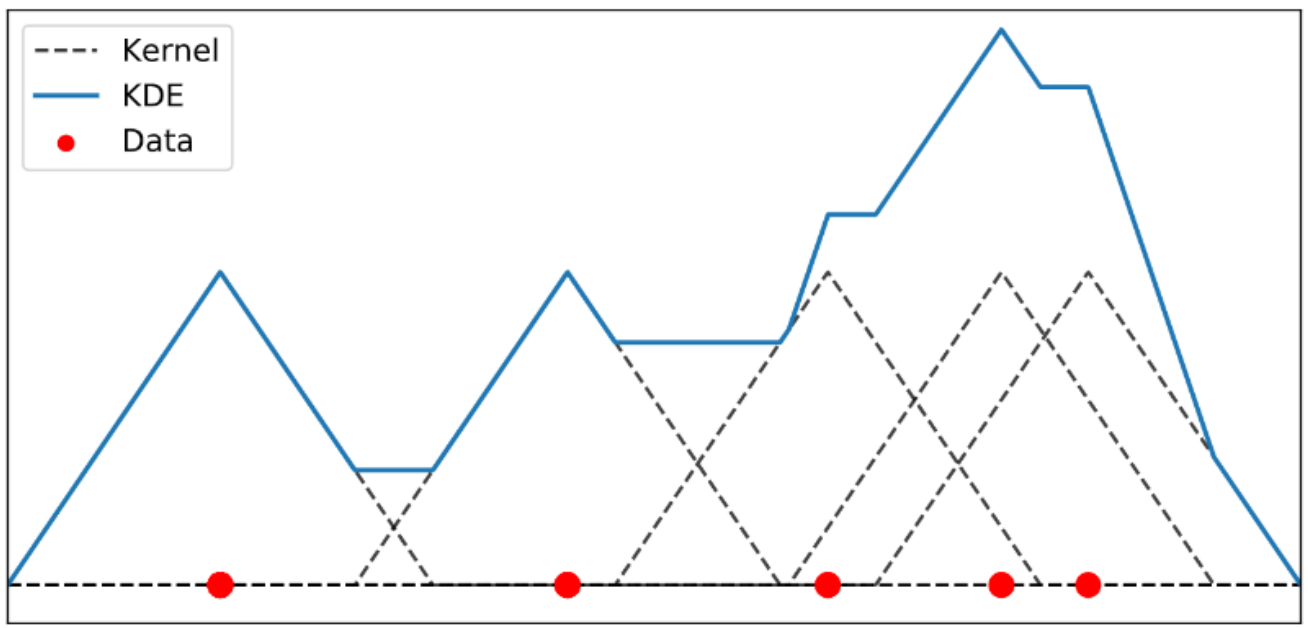
The kernel function K is typically

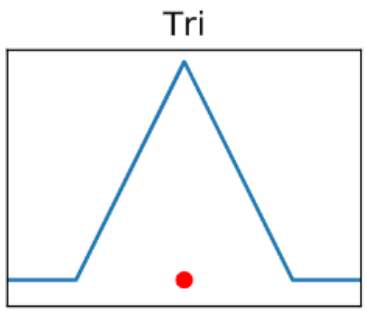
- everywhere non-negative: $K(x) \geq 0$ for every x
- symmetric: $K(x) = K(-x)$ for every x
- decreasing: $K'(x) \leq 0$ for every $x > 0$.



The *triangular* kernel (or *linear* kernel) is given by

$$f(x) \propto \max(1 - |x|, 0).$$

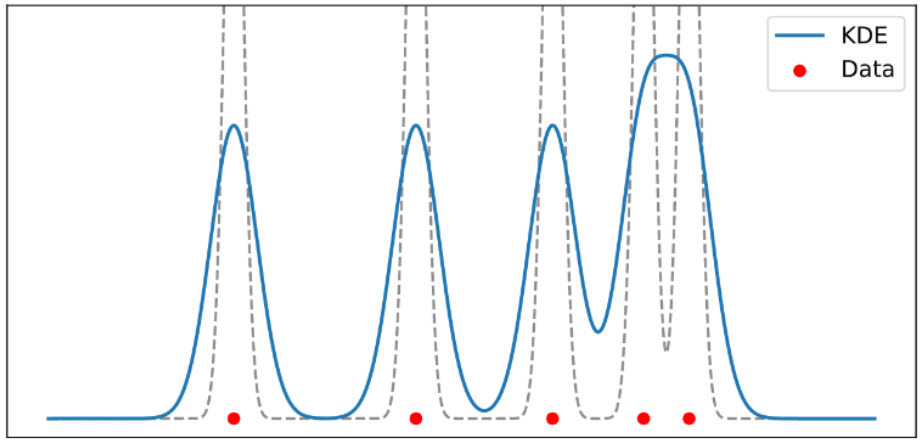




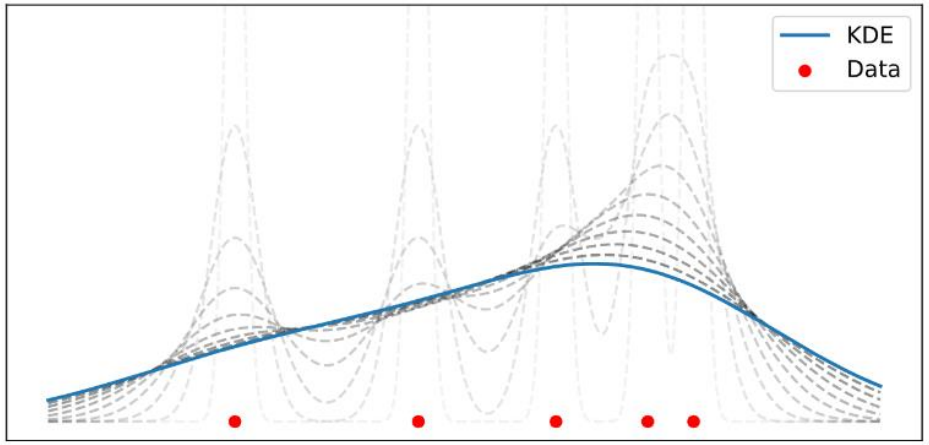
Estimating PDF from data: Kernel Density Estimate

Choice of bandwidth

Too narrow



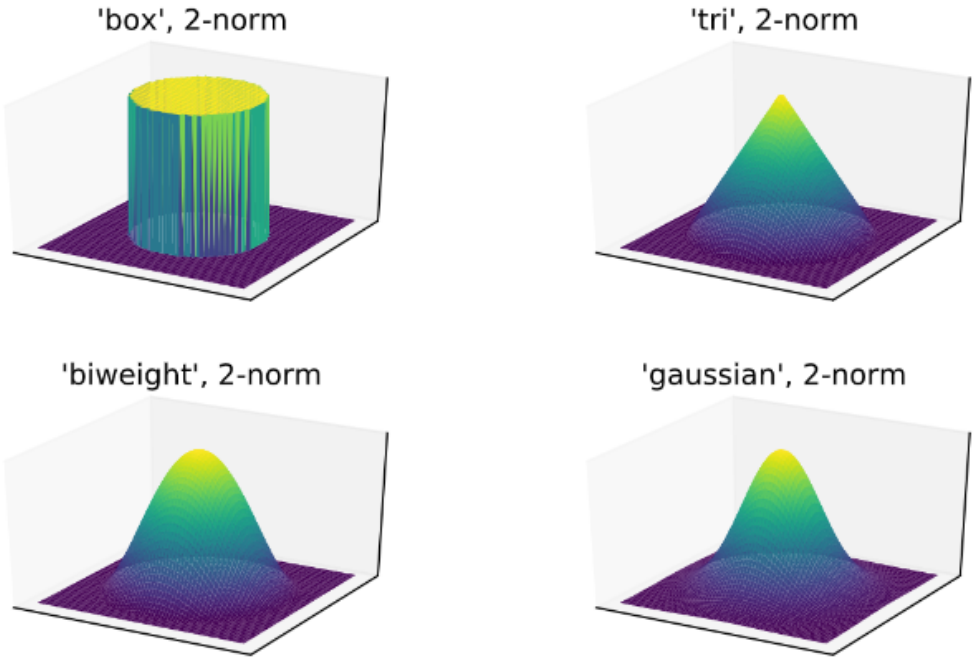
Wide



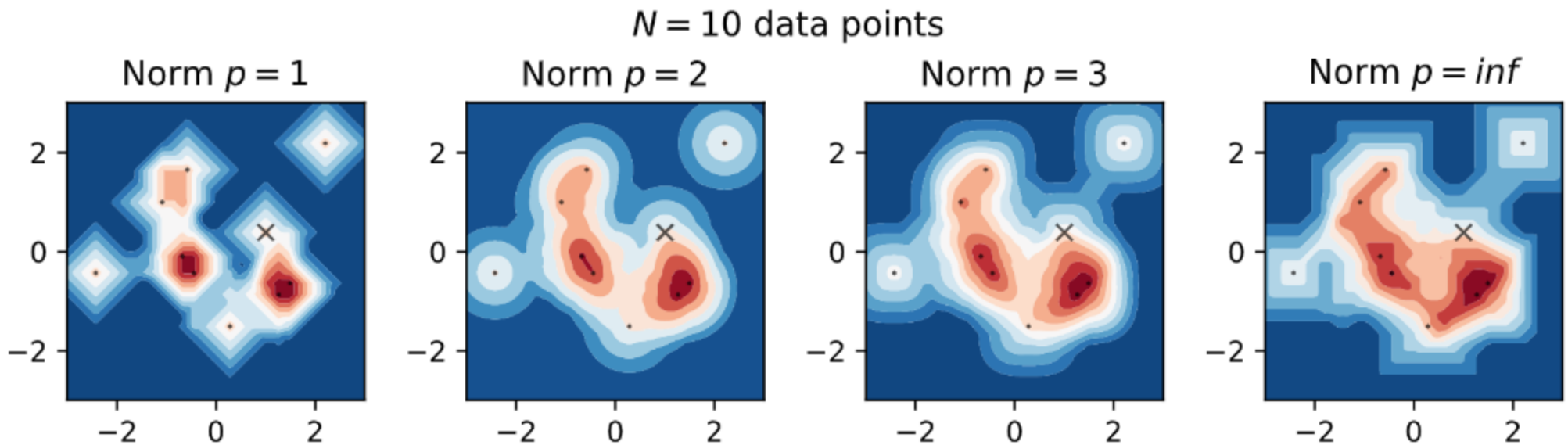
d-dimensional case

An approach to d-dimensional estimates is to write

$$\hat{f}(x) = \frac{1}{h^d} \sum_{i=1}^{N} w_i K\left(\frac{\|x - x_i\|_p}{h}\right), \text{ where } \sum_{i=1}^{N} w_i = 1.$$



As the number of samples grow, the choice of both kernel K and norm p becomes unimportant. The bandwidth H is still important.



What to take from this lesson

- Probability density function (PDF) is the right way to describe the joint probability distribution of continuous numerical features

Good news:

- Knowing PDF gives us all necessary information about the data
- There are ways to estimate PDF directly from data in non-parameteric way (KDE)

Bad news:

- In data spaces with high intrinsic dimension (not equivalent to the number of features!), PDF can not be computed from data in any reasonable form