Kernel Density Estimation

Create a smooth curve given a set of data

Visualization technique

Continuous replacement of histogram

Estimate probability density function of random variable

higher bandwidth → smoother curve, distant points can contribute

low bandwidth → consider points close to current position

Weight distances of all data points

Kernel function - how the point distances are weighted

$$\hat{p}_n(x) = \sum_{\text{observations}} K\left(\frac{x - \text{observation}}{\text{bandwidth}}\right)$$

K - kernel function

Probability density - relationship between observations & their probability

Histogram plots provide a fast & reliable way to visualize the probability density of a data sample

Parametric probability density estimation involves selecting a common distribution and estimating the parameters for the density function from a data sample.

Non parametric probability density estimation involves using a technique to fit a model to the arbitrary distribution of the data eg. kernel density estimation

KDE function

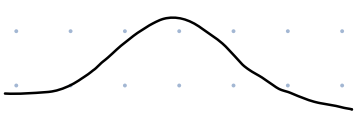
$$\hat{p}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)$$

$h>0$ - bandwidth (controls amount of smoothing)

$K(x)$ - kernel function, smooth, symmetric

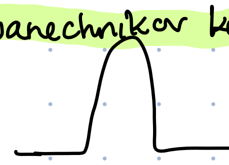
eg.

Gaussian Kernel



Uniform Kernel

Epanechnikov kernel



KDE smoothes each data point X_i into a small density bump & sums all the small bumps together to get a final density estimate

Gaussian kernels

In 1D

$$K(x;\sigma) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp(-x^2/2\sigma^2)$$

$$K(x,y;\sigma) = \frac{1}{2\pi\sigma^2} \exp\left(-(x^2+y^2)/2\sigma^2\right)$$

$$\boxed{K(x,y) = \frac{1}{N\pi w^2} \sum_{i=1}^{N} \exp\left(-\left(\left(\frac{x-x_i}{w}\right)^2 + \left(\frac{y-y_i}{w}\right)^2\right)\right)}$$

In Matlab