

Cross Validated is a question and answer site for people interested in statistics, machine learning, data analysis, data mining, and data visualization. Join them; it only takes a minute:

[Sign up](#)

Here's how it works:

Anybody can ask a question

Anybody can answer

The best answers are voted up and rise to the top

Can you explain Parzen window (kernel) density estimation in layman's terms?

More jobs means more choice



Get started

Parzen window density estimation is described as

$$p(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h^2} \phi\left(\frac{x_i - x}{h}\right)$$

where n is number of elements in the vector, x is a vector, $p(x)$ is a probability density of x , h is dimension of the Parzen Window, and ϕ is a window function.

My questions are:

1. What is the basic difference between a Parzen Window Function and other density functions like Gaussian Function and so on?
2. What is the role of the Window Function (ϕ) in finding the density of x ?
3. Why can we plug other density functions in place of the Window Function?
4. What is the role of h in finding the density of x ?

[pdf](#) | [kernel-smoothing](#) | [intuition](#) | [density-estimation](#)

edited Aug 31 '17 at 15:28



kjetil b halvorsen
21.5k 9 63 151

asked Nov 3 '16 at 14:30



why
200 4 19

2 Answers

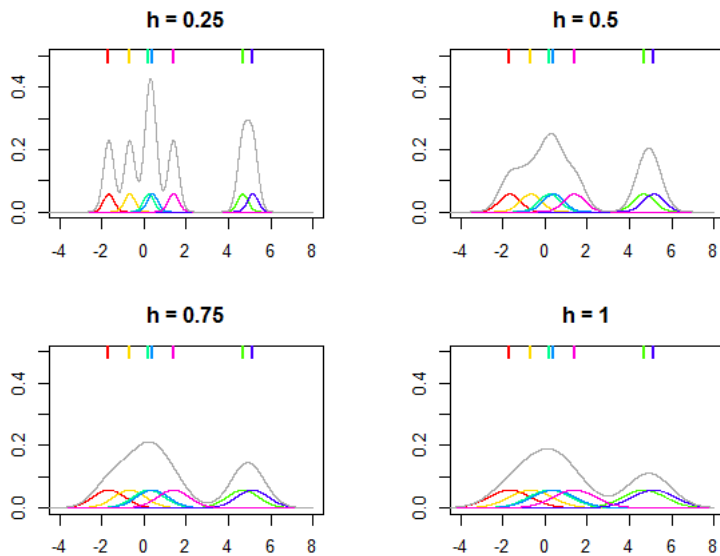
Parzen window density estimation is another name for *kernel density estimation*. It is a nonparametric method for estimating continuous density function from the data.

Imagine that you have some datapoints x_1, \dots, x_n that come from common unknown, presumably continuous, distribution f . You are interested in estimating the distribution given your data. One thing that you could do is simply to look at the empirical distribution and treat it as a sample equivalent of the true distribution. However if your data is continuous, then most probably you would see each x_i point appear only once in the dataset, so based on this, you would conclude that your data comes from an uniform distribution since each of the values have equal probability. Hopefully, you can do better then this: you can pack your data in some number of equally-spaced intervals and count the values that fall into each interval. This method would be based on estimating the **histogram**. Unfortunately, with histogram you end up with some number of bins, rather than with continuous distribution, so it's only a rough approximation.

Kernel density estimation is the third alternative. The main idea is that you approximate f by a **mixture** of continuous distributions K (using your notation ϕ), called *kernels*, that are centered at x_i datapoints and have scale (*bandwidth*) equal to h :

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right)$$

This is illustrated on the picture below, where normal distribution is used as kernel K and different values for bandwidth h are used to estimate distribution given the seven datapoints (marked by the colorful lines on the top of the plots). The colorful densities on the plots are kernels centered at x_i points. Notice that h is a *relative* parameter, it's value is always chosen depending on your data and the same value of h may not give similar results for different datasets.



Kernel K is a probability density function, so it needs to integrate to unity. It also needs to be symmetric so that $K(x) = K(-x)$ and, what follows, centered at zero. [Wikipedia article on kernels](#) lists many popular kernels, like Gaussian (normal distribution), Epanechnikov, rectangular (uniform distribution), etc. Basically any distribution meeting those requirements can be used as a kernel.

Obviously, the final estimate will depend on your choice of kernel (but not that much) and on the bandwidth parameter h . The following thread [How to interpret the bandwidth value in a kernel density estimation?](#) describes the usage of bandwidth parameters in greater detail.

Saying this in plain English, what you assume in here is that the observed points x_i are just a sample and follow some distribution f to be estimated. Since the distribution is continuous, we assume that there is some unknown but nonzero density around the near neighborhood of x_i points (the neighborhood is defined by parameter h) and we use kernels K to account for it. The more points are in some neighborhood, the more density is accumulated around this region and so, the higher the overall density of \hat{f}_h . The resulting function \hat{f}_h can be now evaluated for any point x (without subscript) to obtain density estimate for it, this is how we obtained function $\hat{f}_h(x)$ that is an approximation of unknown density function $f(x)$.

The nice thing about kernel densities is that, not like histograms, they are continuous functions and that they are themselves valid probability densities since they are a mixture of valid probability densities. In many cases this is as close as you can get to approximating f .

The difference between kernel density and other densities, as normal distribution, is that "usual" densities are mathematical functions, while kernel density is an approximation of the true density estimated using your data, so they are not "standalone" distributions.

I would recommend you the two nice introductory books on this subject by Silverman (1986) and Wand and Jones (1995).

Silverman, B.W. (1986). Density estimation for statistics and data analysis. CRC/Chapman & Hall.

Wand, M.P and Jones, M.C. (1995). Kernel Smoothing. London: Chapman & Hall/CRC.

edited Apr 13 '17 at 12:44

answered Nov 3 '16 at 15:31



Community ♦
1




Tim ♦
47.5k 7 105 188


What is x here? – why Nov 4 '16 at 13:38

@anonymous x_i are your datapoints, x is the point at which you evaluate the density function. – Tim ♦ Nov 4 '16 at 13:42

@anonymous I added edit referring your question in comment in the end of "Saying this in plain English..." paragraph. – Tim ♦ Nov 4 '16 at 13:48



Love remote work?
Find it on a new kind of career site



stackoverflow
JOBS

Get started

1) My understanding is that users have a choice of functions to use for ϕ , and that the Gaussian function is a very common choice.

2) The density at x is the mean of the different values of $\phi_h(x_i - x)$ at x . For example, you might have $x_1 = 1, x_2 = 2$, and a Gaussian distribution with $\sigma = 1$ for ϕ_h . In this case, the density at x would be

$$\frac{\hat{\mathcal{N}}_{1,1}(x) + \hat{\mathcal{N}}_{2,1}(x)}{2}.$$

3) You can plug in any density function you like as your window function.

4) h determines the width of your chosen window function.

answered Nov 3 '16 at 15:12



David J. Harris

8,389 1 20 45
