Kernel Density Estimation

Aishwarya Mandyam

Dec 27th, 2022

Kernel density estimation is a nonparametric way to estimate the probability density function of a random variable. For example, suppose we have a collection of random variable samples and are curious about the shape of the underlying distribution, while only having access to a finite and possibly limited sample size. One way to visualize this distribution is to create a histogram. However, a histogram discretizes the samples into fixed bin sizes; it can be more useful to learn a continuous distribution.

One way to estimate this distribution is by estimating the probability density function of the given random variables. A kernel density estimator estimates this probability density function by first placing a kernel around each of the samples. The estimated function is then the average over all of these kernels; the result is a smoothed kernel that estimates the density of all the samples together [5, 4]. The density of given a sample x is:

$$\widehat{p}(x) = \frac{1}{m} \sum_{j=1}^{m} K\left(\frac{x - x_j}{h}\right), \tag{1}$$

where K is a kernel function, m is the number of random variable samples, and h is a bandwidth hyperparameter. The kernel function determines how highly to weight samples based on the distance they are to the given point. There are several choices available for the kernel K, and choosing this can be task dependent; the kernel choice can significantly affect the distribution shape, so this is an important decision to make. Example kernel functions include normal, uniform, triangular, and quartic. A kernel allows us to predict density at points we have not seen based on points that we have.

The bandwidth hyperparameter h can be chosen depending on the samples, but there are rule-of-thumb estimation strategies as well [6]. The bandwidth controls the width of the kernel; in short, it is a way to dictate the "smoothness" of the distribution. h does not need to be fixed though; adaptive bandwidth kernel density estimation can be powerful when the samples are multi-dimensional [7].

Let's visualize an example of the smoothing that a KDE can do. In Figure 1, we select 100 samples from $N(0,1)$ and plot the KDE fit from these samples. A priori, we know that these samples were generated from $N(0,1)$, so we can compare the KDE to the known distribution. As we can see, the KDE is a close approximation, but not perfect. The KDE is in particular, a function of the samples that we generated; with a larger number of samples from $N(0,1)$, it will more closely match the known distribution.

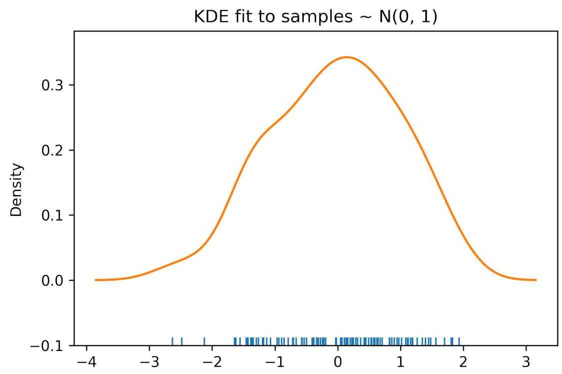

Figure 1: 100 random samples generated from $N(0, 1)$ are shown on the x-axis in a rug-plot. We fit a KDE to this set of samples, and see that it roughly resembles a Normal distribution centered at 0.

Now, suppose instead of estimating a probability density function, we want to estimate a conditional probability density function. That is, instead of estimating $p(x)$, we want to estimate $p(x|y)$ where y is another random variable. The density function for conditional probabilities is a simple extension to Equation 1:

$$\widehat{p}(x|y) = \frac{\widehat{p}(x,y)}{\widehat{p}(x)} = \sum_{j=1}^{m} \frac{K\left(\frac{x-x_j}{h}\right) K'\left(\frac{y-y_j}{h'}\right)}{\sum_{\ell=1}^{m} K\left(\frac{y-y_\ell}{h}\right)}. \tag{2}$$

where we use two kernels to estimate the joint density $p(x, y)$ and one kernel to estimate the marginal density $p(y)$. Conditional kernel density estimation has a wide range of applications including timeseries data, nonparametric Bayesian inference, and visualization on large datasets [1]. The CKDE has several favorable theoretical guarantees [8], but it notably does not scale as the size of the parameter space increases [2]. Several attempts have been made to speed them up the CKDE [3], resulting in applications to datasets with high-dimensional samples.

References

[1] M. P. Holmes, A. G. Gray, and C. L. Isbell, "Fast nonparametric conditional density estimation," 2012. [Online]. Available: https://arxiv.org/abs/1206.5278

[2] R. Izbicki and A. B. Lee, "Converting high-dimensional regression to high-dimensional conditional density estimation," 2017. [Online]. Available: https://arxiv.org/abs/1704.08095

[3] ——, "Nonparametric conditional density estimation in a high-dimensional regression setting," *Journal of Computational and Graphical Statistics*, vol. 25, no. 4, pp. 1297–1316, 2016. [Online]. Available: https://doi.org/10.1080/10618600.2015.1094393

[4] E. Parzen, "On Estimation of a Probability Density Function and Mode," *The Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 1065 – 1076, 1962. [Online]. Available: https://doi.org/10.1214/aoms/1177704472

[5] M. Rosenblatt, "Remarks on Some Nonparametric Estimates of a Density Function," *The Annals of Mathematical Statistics*, vol. 27, no. 3, pp. 832 – 837, 1956. [Online]. Available: https://doi.org/10.1214/aoms/1177728190

[6] S. J. Sheather and M. C. Jones, "A reliable data-based bandwidth selection method for kernel density estimation," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 53, no. 3, pp. 683–690, 1991. [Online]. Available: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1991.tb01857.x

[7] G. R. Terrell and D. W. Scott, "Variable Kernel Density Estimation," *The Annals of Statistics*, vol. 20, no. 3, pp. 1236 – 1265, 1992. [Online]. Available: https://doi.org/10.1214/aos/1176348768

[8] A. van der Vaart, *Asymptotic Statistics*. Cambridge University Press, 2000, vol. 3.