Ben Aldous

Sarah Krstyen

MDCRC 6521

April 17, 2014

densityestimation 0.1.1

**Problem**

Given a sample, it is sometimes easy to assume the data is independent and identically distributed without actually knowing the distribution from which the data comes. Density estimation techniques attempt to predict the distribution of a random variable. This is currently a field of high interest in statistics, and much investigational work is yet to be done.

**Solution**

The package densityestimation 0.1.0 contains three functions designed to assist in visualizing the unknown density from which a given sample might come. Given a one-dimensional numeric numpy array (or something the function can coerce into such), the functions use the techniques described below to estimate the density and plot the result using matplotlib.

* **hist** employs a histogram density estimator. Given an origin *o* and a bandwidth *h*, for each element of the sample, the function identifies the integer *j* such that the element falls within the range *[o+jh,o+(j+1)h)*. It then predicts the placement of future elements based on the proportion of the sample in each such “bin.”
* **kde** uses kernel density estimation. This technique attempts to smooth the histogram density estimator by means of a “kernel,” a function with certain desirable properties. In general, kernel density estimators are more efficient and more accurate than histogram density estimators. The function allows the user to choose a bandwidth and a kernel, allowing for wide flexibility in application.
* **nnde** is based on the theory of nearest-neighbor density estimation. Rather than smooth the estimated density via a kernel, it averages distances using the *k*th neighbor, where *k* is chosen by the user. In particular, the estimation will be smoother for a larger *k*.

**Future Objectives**

There are a variety of improvements that could be made to these functions. Some examples include “real-time” selection of tuning parameters, suggestion of optimal values for these parameters, and expansion to multidimensional data.