Homework 7: Fun with Kernel Density Estimation

36-402, Advanced Data Analysis

Due at 10:30 am on Tuesday, 27 March 2012

1. *A little theory* (25 points total)

   (a) You are given a kernel K which satisfies $K(u) \geq 0$, $\int K(u)du = 1$, $\int uK(u)du = 0$, $\int u^2 K(u)du = \sigma^2_K < \infty$. You are also given a bandwidth $h > 0$, and a collection of n univariate observations $x_1, x_2, \ldots x_n$. Assume that the data are independent samples from some unknown density f.

      i. (1 point) Give the formula for $\hat{f}_h$, the kernel density estimate corresponding to these data, this bandwidth, and this kernel.

      ii. (10 points) Find the expectation of a random variable whose density is $\hat{f}_h$, in terms of the sample moments, h, and the properties of the kernel function.

      iii. (10 points) Find the variance of a random variable whose density is $\hat{f}_h$, in terms of the sample moments, h, and the properties of the kernel function.

      iv. (4 points) How must h change as n grows to ensure that the expectation and variance of $\hat{f}_h$ will converge on the expectation and variance of f?

2. *The transformation trick* (40 points total) Many variables have natural range restrictions, like being non-negative, or being forced to be between 0 and 1. Kernel density estimators do not generally obey these restrictions, so they can give positive probability density to impossible values. One way around this is the *transformation method* or *the transformation trick*: use an invertible function q to map the limited range of X to the whole real line, find the density of the transformed variable, and then undo the transformation.

   In what follows, X is a random variable with pdf f, Y is a random variable with pdf g, and $Y = q(X)$, for a known function q. You may assume that q is continuous, differentiable and monotonically increasing, inverse q^{-1} exists, and is also continuous, differentiable and monotonically increasing.

   (a) (4 points) Find $g(y)$ in terms of f and q.

(b) (3 points) Find $f(x)$ in terms of g and q.

(c) (3 points) Suppose X is confined to the unit interval $[0, 1]$ and $q(x) = \log \frac{x}{1-x}$. Find $f(x)$ in terms of g and this particular q.

(d) (5 points) The Beta distribution is confined to $[0, 1]$. Draw 1000 random values from the Beta distribution with both shape parameters equal to $1/2$. Call this sample `x`, and plot its histogram. (Hint: `?rbeta`.)

(e) (8 points) Fit a Gaussian kernel density estimate to `x` , using `density`, `npudens`, or any other existing one-dimensional density estimator you like.

(f) (7 points) Find a Gaussian kernel density estimate for `logit(x)`.

(g) (10 points) Using your previous results, convert the KDE for `logit(x)` into a density estimate for `x` .

(h) (5 points) Make a plot showing (i) the true Beta density, (ii) the "raw" kernel density estimate from 2e, and (iii) the transformed KDE from 2g. Make sure that the plotting region shows all three curves adequately, and that the three curves are visually distinct.

3. (35 points total) The data set n90_pol.csv contains information on 90 university students who participated in a psychological experiment designed to look for relationships between the size of different regions of the brain and political views. The variables `amygdala` and `acc` indicate the volume of two particular brain regions knwon to be involved in emotions and decision-making, the amygdala and the anterior cingulate cortex; more exactly, these are residuals from the predicted volume, after adjusting for height, sex, and similar body-type variables. The variable `orientation` gives the students' locations on a five-point scale from 1 (very conservative) to 5 (very liberal). `orientation` is an ordinal but not a metric variable, so scores of 1 and 2 are not necessarily as far apart as scores of 2 and 3.

   (a) (5 points) Ignoring the fact that `orientation` is an ordinal variable, what is the correlation between it and the volume of the amygdala? Between `orientation` and the volume of the ACC?

   (b) (10 points) Using case resampling, give 95% bootstrap confidence intervals for these correlations.

   (c) (15 points) Using `npcends`, plot the condition distribution of the volume of the amygdala as a function of political orientation. Do the same for the volume of the ACC. Make sure that in both cases you are treating `orientation` as an ordinal variable. You will be graded on how easy your plots are to read.

   (d) (5 points) What (if anything) can you conclude about the differences in brains between more and less conservative university students? Justify your answer by referring to your earlier work.