STA 621: Nonparametric Statistics
Breheny

Assignment 4
Due: Thursday, November 1

Mathematical concepts and derivations

1. Show that, for the integrated squared error loss,

$$\mathbb{E}L(f, \hat{f}) = \int b(x)^2 dx + \int v(x) dx,$$

where $b(x) = \mathbb{E}\hat{f}(x) - f(x)$ is the bias of $\hat{f}(x)$ and $v(x) = \mathbb{V}\hat{f}(x)$ is the variance of $\hat{f}(x)$.

2. We showed in class that

$$\mathbb{E}\{\hat{f}(x) - f(x)\}^2 \approx \frac{1}{4}\sigma_K^4 h^4 f''(x)^2 + \frac{f(x)\int (K^2(x)dx}{nh},$$

where $\sigma_K^2 = \int x^2 K(x) dx$. Note that one of these terms comes from the bias of $\hat{f}(x)$ and the other comes from the variance of $\hat{f}(x)$.

   (a) How does the variance of $\hat{f}(x)$ change as a function of x?
   (b) How does the variance of $\hat{f}(x)/f(x)$ change as a function of x?
   (c) It is sometimes claimed that methods using an adaptive bandwidth (in which h changes as function of x) correct for the tendency of fixed-bandwidth estimators to have high variance in regions with little data. Are such claims referring to the variance of the density itself or the relative accuracy of the density?

Simulation

3. When we discussed the bootstrap, we noted that drawing random samples from $\hat{F}$, the empirical CDF, is equivalent to resampling the original data with replacement. Suppose that instead, we wish to draw random samples from $\hat{f}$, a kernel density estimate. Write a function called `rdensity` that produces random samples from an estimated density. The function should take three arguments: `n`, the desired number of draws from $\hat{f}$, `d`, a fitted density object as returned by the `density` function, and `x`, the original data (strictly speaking, you do not *need* the original data, but it is convenient to be able to access it. If you would like to write a function that does not require `x`, feel free to do so). Your function only needs to work for the case of Gaussian kernels, but it does need to work for arbitrary bandwidths.

Application

4. The course website contains a data set from the National Health and Nutrition Examination Survey (NHANES) that lists the triglyceride levels of 3,026 adult women.

   (a) Obtain a kernel density estimate for the distribution of triglyceride levels in adult women and plot it. You are free to decide on whatever kernel and bandwidth you like, but describe which ones you used.

   (b) Obtain a parametric density estimate assuming that triglyceride levels follow a normal distribution and overlay this density estimate with your estimate from (a).

5. Try to obtain a kernel density estimate for the nerve pulse data on the course website, with bandwidth chosen by (unbiased) cross-validation. You will receive a warning message, and your estimate will appear clearly incorrect.

   (a) What's going on? What is causing this problem?

   (b) Fix the problem and obtain a reasonable-looking estimate of the density of waiting times between nerve pulses.

6. The course website contains a data set with measurements of the rainfall from 26 clouds (for the purposes of this assignment, ignore the "Seeded" column). Most clouds gave off very little precipitation, so the density near 0 is high. Standard kernel approaches produce an estimate of the density $\hat{f}$ that is high near zero and – this is the problem – below zero. Obviously, rainfall cannot be negative, so this estimate is unappealing. For (a)-(c) below, estimate the density according to the described approach, and plot the resulting estimate (you may overlay all your answers into a single plot or keep them separate; either way is fine).

   (a) Estimate the density using the standard approach, ignoring the boundary problem.

   (b) Estimate the density by taking a log transformation of the data, fitting the density on this scale, then transforming the estimated density back to the original scale.

   (c) Estimate the density by taking the standard approach and reflecting the estimated density that lies in $(-\infty, 0)$ about 0.