10/36-702 Homework 2
Due: Friday 3/1/2013

Instructions: Hand in your homework to Michelle Martin (GHC 8001) before 3:00pm on Friday 3/1/2013.

1. Let $X_1, \ldots, X_n \sim P$ where P has density p and $0 \leq X_i \leq 1$. Find the asymptotic bias of $\widehat{p}_h(0)$ where $\widehat{p}_h$ is the kernel density estimator. Assume that $p \in \Sigma(2, L)$ and assume a Gaussian kernel. Note: because 0 is a boundary point, the bias is different than the bias we computed in class.

2. Let $X_1, \ldots, X_n \sim P$ where P has density p and $X_i \in \mathbb{R}$. Let

$$\widehat{p}_h(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right)$$

be the usual kernel density estimator. Assume that $h = h_n$ is such that $h_n \to 0$ and $nh_n \to \infty$ as $n \to \infty$.

(a) Assuming only that $p(x)$ is continuous in a neighborhood of x, show that

$$\widehat{p}_h(x) \xrightarrow{P} p(x).$$

(b) Let $p_h(x) = \mathbb{E}(\widehat{p}_h(x))$. It can be shown that

$$\frac{\widehat{p}_h(x) - p_h(x)}{\mathrm{se}_n(x)} \rightsquigarrow N(0, 1)$$

where $\mathrm{se}_n(x) = \sqrt{\mathrm{Var}(\widehat{p}_h(x))}$. Note that $\mathrm{se}_n(x)$ is the standard error of the mean as $\widehat{p}_h(x) = \frac{1}{n}\sum_{i=1}^{n} Z_i$ where

$$Z_i = \frac{1}{h} K\left(\frac{x - X_i}{h}\right).$$

Let s^2 be the sample variance of $Z_1, \ldots, Z_n$:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(Z_i - \overline{Z}_n)^2.$$

Let $\widehat{\mathrm{se}}_n(x) = s/\sqrt{n}$. Show that

$$\frac{\widehat{\mathrm{se}}_n(x)}{\mathrm{se}_n(x)} \xrightarrow{P} 1$$

and that

$$\frac{\widehat{p}_h(x) - p_h(x)}{\widehat{\mathrm{se}}_n(x)} \rightsquigarrow N(0, 1).$$

Note: keep in mind that $h = h_n$ is changing with n. You may make extra assumptions if needed.

(c) Suppose that h_n takes the form of the optimal bandwidth, e.g., $h_n = Cn^{-1/5}$ for $p \in \Sigma(2, L)$. Show that

$$\frac{\widehat{p}_h(x) - p(x)}{\mathrm{se}_n(x)} \rightsquigarrow N(b(x), 1)$$

for some function $b(x)$.

(d) Continuing from part (c), suppose we use the confidence interval

$$C_n = [\widehat{p}_h(x) - z_{\alpha/2}\,\widehat{\mathrm{se}}_n(x),\ \widehat{p}_h(x) + z_{\alpha/2}\,\widehat{\mathrm{se}}_n(x)].$$

Explain why $P(p(x) \in C_n)$ does not converge to $1 - \alpha$ as $n \to \infty$. (In fact, no one really knows how to compute a confidence interval for a density estimator.)

3. Let $\widehat{m}$ be a linear smoother so that $\widehat{m}(x) = \sum_i \ell_i(x) Y_i$ and $\widehat{Y} = LY$. Prove the leave-one-out identity:

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{m}_{(-i)}(X_i))^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \widehat{m}(X_i)}{1 - L_{ii}}\right)^2.$$

4. Let $\mathcal{H}$ be a Hilbert space of functions. Suppose that the evaluation functionals $\delta_x f = f(x)$ are continuous. Show that $\mathcal{H}$ is a reproducing kernel Hilbert space and find the kernel.