Homework 11
Statistics 200B
Due May 2, 2019

1. Prove Theorem 20.7 from Wasserman: that the following identity holds for the cross-validation estimator of risk for a histogram:

$$\int \left(\hat{f}_n(x)\right)^2 dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{(-i)}(X_i) = \frac{2}{(n-1)h} - \frac{n+1}{(n-1)h}\sum_{j=1}^{m} \hat{p}_j^2$$

2. Calculate $E[\hat{f}_n(x)]$ and $V[\hat{f}_n(x)]$ when $X_1, \ldots, X_n$ are *iid* random variables with PDF f and

$$\hat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right)$$

for some kernel function K. Express the variance using the form $V[Z] = E[Z^2] - (E[Z])^2$, rather than $V[Z] = E(Z - E[Z])^2$.

3. Let $b(x) = E[\hat{f}_n(x)] - f(x)$, using the same setup as in Problem 1. Show that

$$b(x) \approx \frac{1}{2} h^2 f''(x) \sigma_K^2$$

where $\sigma_K^2 = \int x^2 K(x) dx$. *Hint: Your expression for $E[\hat{f}_n(x)]$ will contain a term like $\frac{x-y}{h}$. Make the change of variable $y = x - ht$, then use the Taylor expansion*

$$f(x - ht) \approx f(x) - h(t) f'(x) + \frac{1}{2} h^2 t^2 f''(x).$$

4. Read the file `glass.dat` from bCourse into `R`. Estimate the density of the first variable (refractive index) using a histogram and using a kernel density estimator with Normal kernel. Use cross-validation to choose the amount of smoothing in each case. For both the histogram and kernel density estimator, turn in plots showing (a) the estimated risk for different choices of smoothing, and (b) the estimator using the optimal choice of smoothing. Also turn in your code.

5. Suppose we observe $X_1, \ldots, X_n$ *iid* bivariate random variables, with $X_i = (X_{i1}, X_{i2})$. Consider the two-dimensional kernel density estimator of the form

$$\hat{f}_n(x) = \frac{1}{nh_1h_2} \sum_{i=1}^{n} K\left(\frac{x_1 - X_{i1}}{h_1}\right) K\left(\frac{x_2 - X_{i2}}{h_2}\right)$$

where $x = (x_1, x_2)$ and K is the (univariate) Normal PDF with mean zero and variance one.

(a) Show this is equivalent to using a two-dimensional Normal kernel, with different standard deviations for each element and zero correlation between them.

(b) This estimator is implemented in `R` by the function `kde2d`, which is part of the `MASS` package. Use this function to estimate the joint density of the first and seventh variables (`RI` and `Ca`) in the glass dataset. (You may use the function's default bandwidth for this problem.) Experiment with plotting the results, using the functions `image`, `persp`, and `contour`, and turn in the one you like best. *Hint: the argument* `n` *to* `kde2d` *changes the resolution of the resulting image.*

(c) Comment on how the estimator is able to capture the correlation between the two variables, even though there is no correlation in the kernel itself.