Homework 6
Statistics 201B
Due 8:00am Dec 8

1. Suppose we observe an *iid* sample $X_1, \ldots, X_n$, with PDF

$$f(x|\lambda) = \lambda e^{-\lambda x}$$

for $x, \lambda > 0$. Suppose furthermore that we use the prior distribution $\lambda \sim Gamma(a, b)$. (You may need the following facts: Suppose $Y \sim Gamma(a, b)$. The pdf is $f(y; a, b) = \frac{y^{a-1}\exp\{-y/b\}}{\Gamma(a)b^a}$ for $a, b, y > 0$. $E[Y] = ab$, $V[Y] = ab^2$, and the mode of the distribution is $(a-1)b$ when $a \geq 1$.)

   (a) What is the posterior distribution for λ?

   (b) What is the Bayes estimator for λ under squared error loss?

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. Use `load("berkhousing.RData")` to load the data frame `berkhousing` into `R`.

   (a) Using K-nearest neighbor method to fit a non-parametric model for predicting house price by square footage. Use cross validation to find optimal k.

   (b) Using Kernel method (Gaussian kernel) to fit non-parametric model for predicting house price by square footage. Use cross validation to find the optimal bandwidth h.

   (c) Compute the leave-one-out cross-validation risk estimator (under integrated squared error loss) for fitted models in part (a) and part (b). Which model is preferred by this criterion?

4. (Optional; for extra credit) 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. (Optional; for extra credit) 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.