Homework 8
Statistics 201B

This homework is optional. If you choose to turn it in, it can replace your lowest HW grade. You only need to turn in the parts marked "Graded." We will not be grading the other parts, even for completion, because we will be giving out the homework solutions for the other portions before this homework is due (and thus you can check you did other portions correctly for the computational parts of this question).

1. Show the Bayes rule h^* is optimal, that is, if h is any other classification rule, then $R(h^*) \leq R(h)$. (Theorem 22.5 in Wasserman)

2. Prove Theorem 22.7 in Wasserman.

3. `berkhousing.csv` on bcourses gives data on houses sold in Berkeley.

   Using K-nearest neighbor method to fit a non-parametric model for predicting house price (in 100K) by square footage. Use cross validation to find optimal k.

4. Let λ represent the average time (in units of days) between earthquakes in the Berkeley area. To make this more precise, let's consider only earthquakes with magnitude 3 or greater on the Richter scale, and whose epicenter is within a 10 mile radius of downtown Berkeley, whose coordinates I have as $37°52'18''N$ and $122°16'22''W$.

   Consider a Bayesian model in which, conditional on unknown parameter λ, $X_1, \ldots, X_n$ are iid with exponential PDF

   $$f(x|\lambda) = \frac{1}{\lambda} e^{-x/\lambda}$$

   for $x > 0$, and the prior distribution is $InverseGamma(a, b)$, with PDF

   $$f(\lambda) = \frac{b^a}{\Gamma(a)} \lambda^{-a-1} e^{-b/\lambda}$$

   for $\lambda > 0$.

   (a) Find the posterior distribution for λ, conditioning on $X_1, \ldots, X_n$. It is fine to write the family and specify its parameters; you do not need to write out the CDF or PDF.

   (b) Show that the posterior mean can be written as a weighted average of the prior mean and the MLE for λ. You may use the fact that the mean of an Inverse Gamma distribution with parameters a and b is $\frac{b}{a-1}$. What happens as $n \to \infty$?

(c) Consider using an Inverse Gamma prior for λ. We need to choose the parameters a and b. You may have some prior knowledge about λ, but it may be difficult to translate this into a choice of a and b. To facilitate this, write expressions for a and b in terms of the prior mean m and the prior variance v, using that $m = \frac{b}{a-1}$ and $v = \frac{b^2}{(a-1)^2(a-2)}$ when $a > 2$.

(d) (**Graded**) Based on your current knowledge, choose parameters a and b, and make a plot of the prior PDF. (There is a `dinvgamma` function in the `R` package `MCMCpack`, which you can install and load using `install.packages("MCMCpack")` and then `library(MCMCpack)`, or you can just code the mathematical form of the prior PDF directly.) Turn in a sentence of explanation with your plot regarding how your prior knowledge (or lack of it) informed your choice of prior distribution.

(e) (**Graded**) The file `BerkeleyEarthquakes.csv` on bCourses contains a data frame called with information about earthquakes within a 10 mile radius of Berkeley, from 1969-2008. The variable `Lag` contains the time between that earthquake and the preceeding earthquake (the earthquakes are in order); hence the first element has `NA`, so you should not use that element.

Using the results you found in problem 2, calculate the posterior distribution for λ, conditional on the observed waiting times. Make a plot comparing your posterior PDF to your prior PDF. Turn in a sentence of explanation with your plot regarding any changes in your knowledge about λ after seeing the data.

(f) (**Graded**) Use rejection sampling to create an i.i.d draw from the posterior and compare how accurate your estimate of the mean of the posterior is with your analytical calculations above. You should use the prior as your sampling distribution, like in the lecture notes. (There is a `rinvgamma` function in the `R` package `MCMCpack` or `invgamma` to draw i.i.d samples from an Inverse Gamma).

(g) (**Graded**) Use importance sampling to estimate the mean of the posterior and compare how accurate your estimate is with your analytical calculations above. Again, use the prior as your sampled data as suggested in the lecture notes.