STAT 201B HW6

Hanze Yao
3033093286

1. (a) $f(\lambda \mid X_1 \cdots X_n) \propto f(X_1 \cdots X_n \mid \lambda)\, f(\lambda)$

$= \lambda^n e^{-\lambda(X_1 + \cdots + X_n)} \lambda^{a-1} e^{-\frac{\lambda}{b}}$

$= \lambda^{n+a-1} e^{-\lambda \sum_{i=1}^{n} X_i - \frac{\lambda}{b}} \sim \text{Gamma}\left(n+a, \frac{b}{b\sum_{i=1}^{n} X_i + 1}\right)$

(b) $r(\hat\lambda \mid x^n) = E_{\lambda|x^n}(\lambda - \hat\lambda)^2 = E_{\lambda|x^n}(\lambda^2 - 2\lambda\hat\lambda + \hat\lambda^2)$

$= \frac{(n+a)\, b^2}{(b\sum_{i=1}^{n} X_i + 1)^2} + \frac{(n+a)^2 b^2}{(b\sum_{i=1}^{n} X_i+1)^2} - 2\hat\lambda \frac{(n+a)\, b}{b\sum_{i=1}^{n} X_i + 1} + \hat\lambda^2$ to minimize this $\frac{\partial}{\partial\hat\lambda} r(\hat\lambda \mid x^n) = 0$

$\frac{-2(n+a)\, b}{b\sum_{i=1}^{n} X_i + 1} + 2\hat\lambda = 0 \quad \hat\lambda = \frac{(n+a)\, b}{b\sum_{i=1}^{n} X_i + 1}$ which is the posterior mean this also minimizes Bayes risk, so it is Bayes estimator

2. $E[\hat f_n(x)] = E\left(\frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right)\right)$

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

$\because X_1 \cdots X_n$ are iid $\therefore = \frac{1}{nh} \times n \times \int K\left(\frac{x-X_1}{h}\right) f(X_1)\, dX_1$

$= \frac{1}{h}\int K\left(\frac{x-y}{h}\right) f(y)\, dy$

$V[\hat f_n(x)] = V\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h} K\left(\frac{x - X_i}{h}\right)\right)$

$\because X_1 \cdots X_n$ are iid

$\therefore = \frac{1}{n^2 h^2} \times n \times V\left(K\left(\frac{x-X_1}{h}\right)\right)$

$= \frac{1}{nh^2}\left\{ E\left[K^2\left(\frac{x-X_1}{h}\right)\right] - \left[E\left(K\left(\frac{x-X_1}{h}\right)\right)\right]^2 \right\}$

$= \frac{1}{nh^2}\left[\int K^2\left(\frac{x-y}{h}\right) f(y)\, dy - \left(\int K\left(\frac{x-y}{h}\right) f(y)\, dy\right)^2\right]$

3. (a) The row 63 in berkhousing. RData has a "NA", so first I need to remove this row and then proceed the analysis. The Rcode is shown on the code sheet. The risk is calculated using the equation $\hat J(h) = \sum_{i=1}^{n} (Y_i - \hat r_{-i}(x_i))^2$

Using k-neares neighbor estimator $\hat r(x) = \sum_{i=1}^{n} w_i(x) Y_i$ where

$$w_i = \begin{cases} \frac{1}{k} & x_i \in N_k(x) \\ 0 & x_i \notin N_k(x) \end{cases}$$

After runing the code, the k value with the smallest risk is $k = 29$, the risk is 869219.

(b) The kernel method assigns weight for all Y_i based on kernel function.

$\hat J(h) = \sum_{i=1}^{n} (Y_i - \hat r_{-i}(x_i))^2 = \frac{\sum_{i=1}^{n} (Y_i - \hat r(x_i))^2}{\left(1 - \frac{K(0)}{\sum_{j=1}^{n} K\left(\frac{x_i - x_j}{h}\right)}\right)^2}$

$\hat r_{-i}(x_i) = \frac{\sum_{j=1, j\neq i}^{n} K\left(\frac{x_i - x_j}{h}\right) Y_j}{\sum_{j=1, j\neq i}^{n} K\left(\frac{x_i - x_j}{h}\right)} = \frac{\sum_{j=1, j\neq i}^{n} K\left(\frac{x_i - x_j}{h}\right) Y_j}{\sum_{j=1}^{n} K\left(\frac{x_i-x_j}{h}\right) - K(0)}$

Guassion kernel $= K\left(\frac{x-x_i}{h}\right) = \frac{1}{\sqrt{2\pi}(h/\sqrt{2})} \exp\left\{-\left(\frac{x-x_i}{h}\right)^2\right\}$

The R code is shown on the code sheet. After running this code, I get the optimized h is 124.7581, the minimum risk is 893382.8

(c) The cross-validation risk estimator in (a) is 869219, in (b) is 893382.8. Because the risk of (a) is lower, so the k nearest neighbor method in (a) is prefered.

R code Sheet

```
objects()
load("~/Documents/STAT_201_B/Homework/HW6/berkhousing.RData")
objects()

head(berkhousing)
dim(berkhousing)

berkhousing = berkhousing[-63,]
dim(berkhousing)

## Question 3(a)

install.packages("fields")
k_n_n <- function(x, y, xseq, k){
   require(fields)
   dmat <- rdist(x, xseq)
   indices <- order(dmat)[1:k]
   return(mean(y[indices]))
}

kseq = 1:(dim(berkhousing)[1]-1)

k_n_n.risk = sapply(kseq, FUN=function(k){
                  sum((berkhousing$price-
                           sapply(1:dim(berkhousing)[1],
                                      FUN=function(i){
                                         k_n_n(x=berkhousing$sqft[-i],
                                                 y=berkhousing$price[-i],
                                                 xseq=berkhousing$sqft[i],
                                                 k=k)
                                      }))^2)
   }
)
kseq[which(k_n_n.risk==min(k_n_n.risk))]

## Question 3(b)

n_w_k.risk <- function(h, x, y){
   require(fields)
   dmat <- rdist(x)
```

```
  K <- dnorm(dmat/h)
  rhat <- sapply(1:length(x), function(j){
    sum(K[,j]/sum(K[,j])*y)
  })
  sum((y-rhat)^2 / (1-dnorm(0)/apply(K, 1, sum))^2)
}

h_opt <- optimize(n_w_k.risk, lower=0.00001,
                  upper=diff(range(berkhousing$sqft)), x=berkhousing$sqft,
                  y=berkhousing$price)$min

n_w_k.risk(h=h_opt, x=berkhousing$sqft, y=berkhousing$price)
```

```
> kseq[which(k_n_n.risk==min(k_n_n.risk))]
[1] 29
```

```
> min(k_n_n.risk)
[1] 869219
```

```
> n_w_k.risk(h=h_opt, x=berkhousing$sqft, y=berkhousing$price)
[1] 893382.8
```

```
> h_opt
[1] 124.7581
```