Assignment 2
10/36-702
Due Friday Feb 19 3:00 pm

1. In this question we will study k-nearest neighbors regression. Consider data $(x_1, Y_1), \ldots (x_n, Y_n) \in \mathbb{R}^d \times \mathbb{R}$. To make things simpler, we will assume that $x_1, \ldots x_n$ are fixed (non-random). Further, we assume that

$$Y_i = m(x_i) + \epsilon_i, \;\; \mathbb{E}(\epsilon_i) = 0, \;\; \mathrm{Var}(\epsilon_i) = \sigma^2, \;\; i = 1, \ldots n.$$

The knn regression estimator is

$$\widehat{m}(x) = \frac{1}{k} \sum_{i \in \mathcal{N}_k(x)} Y_i$$

where $\mathcal{N}_k(x)$ contains the indices of the k x_i's closest to x.

(a) Show that, at any fixed $x \in \mathbb{R}^d$,

$$\mathbb{E}\big(\widehat{m}(x) - m(x)\big)^2 = \Big(\frac{1}{k} \sum_{i \in \mathcal{N}_k(x)} \big(m(x_i) - m(x)\big)\Big)^2 + \frac{\sigma^2}{k}.$$

(b) Assume that $x_1, \ldots x_n \in [0,1]^d$ and are distributed on a uniform grid, over this d-dimensional unit cube. Show that for any $i \in \mathcal{N}_k(x)$,

$$\|x_i - x\|_2 \leq C\Big(\frac{k}{n}\Big)^{1/d},$$

for some constant C that depends only on d but not on n or k. Assume further that $n^{1/d}$ and $k^{1/d}$ are both integers. Suppose that m is Lipschitz with constant L:

$$|m(y) - m(x)| \leq L \, ||y - x||$$

for all x, y. Show that

$$\mathbb{E}\big(\widehat{m}(x) - m(x)\big)^2 \leq (CL)^2\Big(\frac{k}{n}\Big)^{2/d} + \frac{\sigma^2}{k}.$$

(c) Choose k by optimizing the upper bound in part (b). Then plug in this value of k to the upper bound to derive a mean squared error bound for $\widehat{m}(x)$.

2. In this question you will verify the leave-one-out cross-validation formula for kernel regression. (It holds for other linear smoothers too.) Let $S_{ij} = K(X_i, X_j)/\sum_{k=1}^n K(X_i, X_k)$. (For simplicity I am suppressing the bandwidth.) Let $\widehat{m}_{(-i)}$ be the kernel regression estimator obtained by leaving out (X_i, Y_i).

(a) Show that $\widehat{m}_{(-i)}(X_i) = \sum_{j=1}^n S_{ij} Z_j$ where

$$Z_j = \begin{cases} Y_j & j \neq i \\ \widehat{m}_{(-i)}(X_i) & j = i. \end{cases}$$

(b) Now show that

$$\widehat{m}(X_i) - \widehat{m}_{(-i)}(X_i) = S_{ii}(Y_i - \widehat{m}_{(-i)}(X_i)).$$

Hence,

$$Y_i - \widehat{m}_{(-i)}(X_i) = \frac{Y_i - \widehat{m}(X_i)}{1 - S_{ii}}.$$

Conclude that

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

3. Let $\mathbb{Y} = (Y_1, \ldots, Y_n) \in \mathbb{R}^n$ and let $\mathbb{X}$ be then $n \times d$ design matrix so that $\mathbb{X}(i,j) = X_i(j)$. Suppose that $\mathbb{X}$ has orthonormal columns so that $\mathbb{X}^T\mathbb{X} = I$ where I is the $d \times d$ identity matrix.

   (a) Show that the least squares solution is $\widehat{\beta} = \mathbb{X}^T\mathbb{Y}$.

   (b) Let $\widehat{\beta}$ minimize

   $$\frac{1}{2}\|\mathbb{Y} - \mathbb{X}\beta\|_2^2 + \lambda||\beta||_0.$$

   Let v_j be the j^{th} column of $\mathbb{X}$. Show that the solution is

   $$\widehat{\beta}_j = \begin{cases} v_j^T\mathbb{Y} & \text{if } v_j^T\mathbb{Y} > \sqrt{2\lambda} \\ 0 & \text{if } v_j^T\mathbb{Y} \in [-\sqrt{2\lambda}, \sqrt{2\lambda}], \quad j = 1, \ldots d. \\ v_j^T Y & \text{if } v_j^T\mathbb{Y} < -\sqrt{2\lambda} \end{cases}$$

   This is called the hard thresholding estimator.

   Remark: If we had used the ℓ_1 loss, the solution would have been the soft-thresholding estimator $\widehat{\beta}_j = S_\lambda(v_T\mathbb{Y})$ where

   $$S_\lambda(u) = \begin{cases} u - \lambda & \text{if } u > \lambda \\ 0 & \text{if } u \in [-\lambda, \lambda], \quad j = 1, \ldots d. \\ u + \lambda & \text{if } u < -\lambda \end{cases}$$

   Proving this requires knowledge of basic convex analysis, especially subgradients.

4. Consider regression data $(X_1,Y_1),\ldots,(X_n,Y_n)$ where $X_i \in [0,1]^d$. Assume also that $|Y_i| \leq B < \infty$ for some B. Assume that X_i has a density p and that $\inf_{x\in[0,1]^d} p(x) \geq c > 0$. Let k be an integer. Divide $[0,1]^d$ into cubes $C_1,\ldots,C_N$ with length of size $h = 1/k$. Here, $N = k^d$. For $x \in C_j$ define

$$\widehat{m}(x) = \frac{\sum_i Y_i I(X_i \in C_j)}{\sum_i I(X_i \in C_j)}.$$

Assume that m is Lipschitz.

(a) Show that, if $nh^d \to \infty$ then the probability that there exists a cube with no data in it, tends to 0 as $n \to \infty$.

(b) Bound the MSE of $\widehat{m}(x)$. (You may assume there are no empty cubes for this part.)

5. Consider data $(X_1,Y_1),\ldots,(X_n,Y_n)$ where $X_i \in [0,1]$ and $Y_i \in \mathbb{R}$. Suppose that

$$Y_i = m(X_i) + \epsilon_i$$

where $\mathbb{E}[\epsilon_i|X_i] = 0$, $\mathbb{E}[\epsilon_i^2] < \infty$, and $m \in L_2[0,1]$. Let $\psi_1,\psi_2,\ldots$ be an orthonormal basis. Assume that $\max_j \sup_x |\psi_j(x)| \leq C < \infty$. Hence,

$$m(x) = \sum_j \beta_j \psi_j(x)$$

where $\beta_j = \int_0^1 \psi_j(x) m(x) dx$. Suppose that X_i has a density p and that $\inf_{x\in[0,1]} p(x) > 0$.

(a) Suppose the density p is known. Define

$$\widehat{m}(x) = \sum_{j=1}^{k} \widehat{\beta}_j \psi_j(x)$$

where

$$\widehat{\beta}_j = \frac{1}{n}\sum_i \frac{Y_i \psi_j(X_i)}{p(X_i)}.$$

Find an upper bound on $\mathbb{E}\int_0^1 (\widehat{m}(x) - m(x))^2 dx$ when $\sum_j \beta_j^2 j^{2q} \leq C^2$. Find k_n to minimize your upper bound.

(b) Now suppose that p is not known. In this case we can use the same estimate except that

$$\widehat{\beta}_j = \frac{1}{n}\sum_i \frac{Y_i \psi_j(X_i)}{\widehat{p}(X_i)}$$

where $\widehat{p}$ is an estimate of the density. Assume that $||\widehat{p} - p||_\infty = O_P(r_n)$ where $r_n = o(1)$. Show that $\widehat{m}(x)$ is consistent as long as $r_n k_n = o(1)$.