Homework 3
Statistical Machine Learning
10/36-702
Due Friday Feb 24

1. Let $Y_i = m(X_i) + \epsilon_i$ for $i = 1, \ldots, n$. Assume that $|Y_i| \leq M$ and $X_i \in [0,1]^d$. Assume that X_i has a uniform distribution. Let

$$\widehat{m}_h(x) = \frac{\sum_{i=1}^n Y_i K\left(\frac{||x-X_i||}{h}\right)}{\sum_{i=1}^n K\left(\frac{||x-X_i||}{h}\right)}$$

be the kernel estimator. Assume that $m(x) = \mathbb{E}(Y|X=x)$ satisfies

$$|m(x_2) - m(x_1)| \leq L||x_2 - x_1||$$

for all x_1, x_2. Show that

$$\Big|\mathbb{E}(\widehat{m}_h(x)) - m(x)\Big| \leq c_1 h$$

and

$$\mathsf{Var}(\widehat{m}_h(x)) \leq \frac{c_2}{nh^d}.$$

For simplicity, you mau assume that $K(||x||) = I(||x|| \leq 1)$.

2. Generate data as follows. Let $n = 100$ Let $X_i \sim \text{Uniform}[0,1]^d$ where $d = 30$. Let

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

where $\epsilon_i \sim N(0,1)$. Take

$$m(x) = \sum_{j=1}^{30} m_j(x_j)$$

where

$$m_1(x) = 3x, \quad m_2(x) = \cos(5x), \quad m_3(x) = e^x, \quad m_j(x) = 0, j = 4, \ldots, 30.$$

(a) Fit a kernel regression estimator separately to each covariate. Use cross-validation to choose the bandwidth. Plot the data, the estimated functions and the residuals.

(b) Fit a SpAM model. Use the same bandwidth for each covariate. Summarize your results.

(c) Explain why, in this particular case, the marginal regression estimators from part (a), are consistent estimators of the m_j's. Why is it not true in general?

3. In this question, you will derive a generalization bound based on the VC dimension for Adaboost. Let $\mathcal{H} = \{h_1, \ldots, h_N\}$. Let $\mathcal{G}$ denote all functions of form $\text{sign}(\sum_{t=1}^{T} \alpha_t h_t(x))\}$ where $\alpha_t \in \mathbb{R}$ and $h_t \in \mathcal{H}$.

   (a) Note that the Adaboost final classifier is a hyperplane classifier with coordinates $h_1, h_2, \ldots, h_T$. Argue that the number of ways that n data points can be partitioned by $\mathcal{G}$ is bounded as $(en/T)^T$.

   (b) Now consider how many choices of $h_1, h_2, \ldots, h_T$ are possible. Use this to derive a bound on the growth function $s_n(\mathcal{G}, n)$, and a generalization error bound of the form: With probability $> 1 - \delta$, for all $H \in \mathcal{G}$

$$R(H) \leq \widehat{R}(H) + O\left(\sqrt{\frac{T\log(N) + T\ln(en/T) + \ln(1/\delta)}{n}}\right)$$

4. Here is a classifier based on coverings. Let $\mathcal{M}$ be a class of functions $m : [0,1]^d \to [0,1]$. For any $m \in \mathcal{M}$ define the classifier

$$h_m(x) = \begin{cases} 1 & \text{if } m(x) > 1/2 \\ 0 & \text{if } m(x) \leq 1/2. \end{cases}$$

   Let $\mathcal{N}(\epsilon)$ be the smallest number of balls of size ϵ in the metric $||f-g||_\infty = \sup_x |f(x) - g(x)|$ needed to cover $\mathcal{M}$. Assume that $\mathcal{N}(\epsilon) < \infty$ for every $\epsilon > 0$ and that the true regression function $m(x) = \mathbb{E}(Y|X = x)$ is contained in $\mathcal{M}$. Let ϵ_n satisfy

$$\log \mathcal{N}(\epsilon_n) \asymp n\epsilon_n^2.$$

   Let $\mathcal{M}_n$ be an ϵ_n net of $\mathcal{F}$. Finally let m_n minimize

$$\widehat{R}(m) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \neq h_m(X_i))$$

   for $m \in \mathcal{M}_n$. Show that, for large anough C_1,

$$\mathbb{P}(R(m_n) - R(m) > C_1\epsilon_n) \leq C_2 e^{-C_3 n\epsilon_n^2}$$

   where $R(m) = \mathbb{P}(Y \neq h_m(X))$ and $C_1, C_2, C_3 > 0$.