Homework 3
10-702/36-702 Statistical Machine Learning

Due: Friday Feb 18 3:00

Hand in to: Michelle Martin GHC 8001.

1 Nonparametric Regression

1.1 Theory

Consider the following nonparametric regression model:

$$Y_i = m(x_i) + \epsilon_i, \quad i = 1, \ldots, n$$

where $x_i = i/n$ an $\epsilon_i \sim N(0, \sigma^2)$. Define $\psi_0, \psi_1, \psi_2, \ldots,$ on $[0, 1]$ by $\psi_0(x) = 1$ and

$$\psi_j(x) = \sqrt{2}\cos(\pi j x)$$

for $j \geq 1$. Note that $\psi_0, \psi_1, \psi_2, \ldots,$ are orthonormal: $\int_0^1 \psi_j^2(x)dx = 1$ for each j and $\int_0^1 \psi_j(x)\psi_k(x)dx = 0$ for each $j \neq k$. Assume that m can be expanded in this basis. Hence,

$$m(x) = \sum_{j=1}^{\infty} \theta_j \psi_j(x)$$

where $\theta_j = \int_0^1 m(x)\psi_j(x)dx$. Let $\beta \geq 1$. Assume that $\theta = (\theta_0, \theta_1, \ldots,) \in \Theta$ where

$$\Theta = \left\{ \theta : \sum_{j=0}^{\infty} \theta_j^2 j^{2\beta} \leq C \right\}$$

where $0 \leq C < \infty$. Let

$$\widehat{\theta}_j = \frac{1}{n} \sum_{i=1}^{n} Y_i \psi_j(x_i)$$

and

$$\widehat{m}(x) = \sum_{j=0}^{J} \widehat{\theta}_j \psi_j(x).$$

1. Show that

$$\mathbb{E}(\widehat{\theta}_j) = \theta_j + O\left(\frac{1}{\sqrt{n}}\right).$$

For the rest of the question, you can ignore the second term and assume that $\mathbb{E}(\widehat{\theta}_j) = \theta_j$.

2. Show that
$$\mathrm{Var}(\widehat{\theta}_j) = \frac{\sigma^2}{n}.$$

3. Show that
$$\mathbb{E}\left(\int_0^1 (\widehat{m}(x) - m(x))^2 dx\right) = \frac{J\sigma^2}{n} + \sum_{j=J+1}^{\infty} \theta_j^2.$$

4. Let $J = n^{\frac{1}{2\beta+1}}$. Show that
$$\sup_{\theta\in\Theta} \mathbb{E}\left(\int_0^1 (\widehat{m}(x) - m(x))^2 dx\right) \le C n^{-\frac{2\beta}{2\beta+1}}.$$

2 Nonparametric Classification

1. In the Adaboost algorithm, the choice of a weak classifier h_t and its weight α_t is specified as follows: h_t is chosen to minimize the error ϵ_t on the weighted training data, and $\alpha_t = \frac{1}{2}\ln(\frac{1-\epsilon_t}{\epsilon_t})$. In this problem, we will show that these choices correspond to greedily minimizing the exponential loss at each iteration.

   (a) Show that the exponential loss
   $$\frac{1}{n}\sum_{i=1}^{n} e^{-Y_i f(X_i)} = \Pi_{t=1}^{T} Z_t,$$
   where $Z_t = \sum_{i=1}^{n} w_t(i) e^{-\alpha_t Y_i h_t(X_i)}$ is the normalizing factor for the data weights at iteration t and $f(X_i) = \sum_{t=1}^{T} \alpha_t h_t(X_i)$. (Hint: Express the data weights at each iteration in terms of the initial data weights and then use the fact that the weights at iteration $T+1$ sum to 1.)

   (b) Show that choosing α_t and h_t greedily to minimize Z_t at each iteration, leads to the choices used in Adaboost. (Hint: Express Z_t in terms of ϵ_t and then minimize Z_t with respect to α_t and then for h_t).

2. In this part, we will derive a generalization bound based on the VC dimension for Adaboost, when the weak hypothesis are chosen from a finite class $\mathcal{H}$. Let $\mathcal{G} = \{\text{all functions of form } sign(\sum_{t=1}^{T} \alpha_t h_t(x))\}$.

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

(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 $S(\mathcal{G}, n)$, and a generalization error bound of the form: With probability $> 1 - \delta$, for all $H \in \mathcal{G}$

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

3 Nonparametric Bayes

Let $X_1, \ldots, X_n \sim F$ where $X_i \in \mathbb{R}$. Let the prior π for F be $\mathrm{DP}(\alpha, F_0)$.

(a) Let $\overline{F}_n(x)$ be the posterior Bayes estimator of $F(x)$. Here x is some arbitrary, fixed value. Find the bias and variance of $\overline{F}_n(x)$. Let $F_n(x) = \frac{1}{n}\sum_{i=1}^n I(X_i \leq x)$ be the empirical measure. When is the the mean squared error of $\overline{F}_n(x)$ smaller than the mean squared error of $F_n(x)$?

(b) Use Hoeffding's inequality to get bounds on

$$\mathbb{P}(F_n(x) - F(x) > \epsilon)$$

and

$$\mathbb{P}(\overline{F}_n(x) - F(x) > \epsilon).$$