

**CS281B/Stat241B Homework Assignment 5 (new version: Wed Apr 16 2008; due April 22, 2008)**

1. **(Kernel density estimation)** Suppose  $X_1, \dots, X_n$  are i.i.d. according to some continuous probability distribution on  $\mathbb{R}$ , with density  $f$ . The kernel density estimate

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

is used to estimate the unknown  $f$ . Here,  $K$  is a nonnegative function that satisfies

$$\int K(x)dx = 1,$$

and  $h$  is a smoothing parameter. Suppose we use the  $L_1$  error to measure the performance of the density estimate  $\hat{f}$ ,

$$J(\hat{f}) = \int |f(x) - \hat{f}(x)| dx.$$

Use the bounded difference concentration inequality to show that  $J(\hat{f})$  is concentrated about its expectation.

2. **(Estimating Rademacher complexities of neural network classes)**

- (a) **(Size of parameters)** We can bound the Rademacher complexity of a neural network with Lipschitz nonlinearities in terms of the size of the parameters. Consider the following class  $\mathcal{F}_B$  of two-layer neural networks:

$$\mathcal{F}_B = \left\{ x \mapsto \sum_{i=1}^k w_i \sigma(v_i^T x) : w_i \geq 0, \|w\|_1 \leq B, \|v_i\|_1 \leq B, k \geq 1 \right\},$$

where  $B > 0$  and the nonlinear function  $\sigma : \mathbb{R} \rightarrow \mathbb{R}$  satisfies the Lipschitz condition,  $|\sigma(a) - \sigma(b)| \leq |a - b|$ , and  $\sigma(0) = 0$ . Suppose that the distribution is such that  $\|X\|_\infty \leq 1$  a.s. Show that

$$R_n(\mathcal{F}_B) \leq B^2 R_n(\mathcal{G} \cup -\mathcal{G}) \leq B^2 \sqrt{\frac{2 \log 2d}{n}},$$

where  $\mathcal{G} = \{(x_1, \dots, x_d) \mapsto x_j : 1 \leq j \leq d\}$  and  $d$  is the dimension of the input space,  $\mathcal{X} = \mathbb{R}^d$ . Notice the *two corrections*. Here's why they are necessary:

$$\begin{aligned} & \left\{ \sum_i w_i f_i : w_i \geq 0, \|w\|_1 \leq 1, f_i \in \mathcal{F} \right\} \\ &= \text{co}(\{0\} \cup \mathcal{F}) = \left\{ \sum_i w_i f_i : w_i \geq 0, \|w\|_1 = 1, f_i \in \{0\} \cup \mathcal{F} \right\}, \end{aligned}$$

and  $\sigma(0) = 0$  implies that  $\text{co}(\{0\} \cup \sigma(B \text{co}(\mathcal{G} \cup -\mathcal{G}))) = \text{co}(\sigma(B \text{co}(\mathcal{G} \cup -\mathcal{G})))$ .

- (b) **(Number of parameters)** We have seen that, for the class  $\mathcal{G}_d$  of linear threshold functions on  $\mathbb{R}^d$ , defined by

$$\mathcal{G}_d = \{x \mapsto \text{sign}(w^T x - \theta) : w \in \mathbb{R}^d, \theta \in \mathbb{R}\},$$

the growth function satisfies

$$\Pi_{\mathcal{G}_d}(n) = 2 \sum_{i=0}^d \binom{n-1}{i} \leq 2 \left( \frac{e(n-1)}{d} \right)^d,$$

provided that  $n \geq d+1$ . Give an upper bound for the growth function for the class  $\mathcal{F}_k$  of two-layer networks of linear threshold functions,

$$\mathcal{F}_k = \{x \mapsto f(f_1(x), \dots, f_k(x)) : f \in \mathcal{G}_k, f_1, \dots, f_k \in \mathcal{G}_d\}.$$

Hence give an upper bound on the Rademacher complexity of this class. Assume that  $k, d < n$ .

3. **(Universal consistency)** Let  $\mathcal{G}_d$  be the class of linear threshold functions on  $\mathbb{R}^d$ , defined above. Consider the following pattern classification method. For a sample  $(X_1, Y_1), \dots, (X_n, Y_n)$  from  $\mathbb{R}^d \times \{\pm 1\}$ , let  $f_n \in \text{span}(\mathcal{G}_d)$  be the minimizer of the regularized empirical risk functional

$$\hat{R}_\phi(f) + \lambda_n \|f\|,$$

where

$$\begin{aligned} \phi(\alpha) &= \ln(1 + e^{-\alpha}), & \hat{R}_\phi(f) &= \frac{1}{n} \sum_{i=1}^n \phi(Y_i f(X_i)), \\ \|f\| &= \inf\{r : f \in r \text{co}(\mathcal{G}_d)\}, \end{aligned}$$

and  $\lambda_n$  is a regularization coefficient that is decreasing in  $n$ . Define

$$\begin{aligned} R_\phi(f) &= \mathbb{E}\phi(Yf(X)), & R_\phi^* &= \inf_f R_\phi(f), \\ R(f) &= \Pr(Yf(X) \leq 0), & R^* &= \inf_f R(f), \end{aligned}$$

where, in both cases, the infimum is over all measurable functions. Suppose that the distribution of  $(X, Y)$  satisfies

$$\liminf_{\lambda \rightarrow 0} \{R_\phi(f) + \lambda \|f\| : f \in \text{span}(\mathcal{G}_d)\} = R_\phi^*.$$

- (a) Give an upper bound on the excess risk  $R(f) - R^*$  in terms of the excess  $\phi$ -risk,  $R_\phi(f) - R_\phi^*$ .  
(b) Let  $f_n^*$  be the minimizer of  $R_\phi(f) + \lambda_n \|f\|$ . Show that

$$\|f_n\| \leq \frac{\ln 2}{\lambda_n} \quad \text{and} \quad \|f_n^*\| \leq \frac{\ln 2}{\lambda_n}.$$

- (c) Show that, with probability at least  $1 - e^{-x}$ , for all  $n$  and all  $f$  in  $\text{span}(\mathcal{G}_d)$  satisfying  $\|f\| \leq \ln 2 / \lambda_n$ ,

$$\left| R_\phi(f) - \hat{R}_\phi(f) \right| \leq \epsilon(n, \lambda_n, x),$$

for some suitable  $\epsilon$ . (Notice that you need to prove that this inequality holds uniformly across these  $f$ s and simultaneously for all  $n$ .)

- (d) Hence show that, with probability at least  $1 - e^{-x}$ , for all  $n$ ,

$$R_\phi(f_n) + \lambda_n \|f_n\| \leq R_\phi(f_n^*) + \lambda_n \|f_n^*\| + 2\epsilon(n, \lambda_n, x).$$

- (e) Suggest a sequence  $\lambda_n$  for which  $\mathbb{E}R(f_n) \rightarrow R^*$ .

4. **(Online learning)** Regret bounds for online prediction often have the form

$$R_T = L_T - \min_u L_T(u) \leq \frac{a}{\eta} + b\eta T, \tag{1}$$

where  $L_T$  is the cumulative loss of our algorithm after  $T$  rounds of the game and  $L_T(u)$  is the cumulative loss of a fixed choice  $u$ . In the above bound,  $a$  and  $b$  are some constants that depend on the problem and the algorithm, and  $\eta$  is a parameter of the algorithm.

- (a) Assuming that  $T$ , the time horizon, is known in advance, what is the optimal choice for  $\eta$ ? What is the bound on the regret  $R_T$  with this choice of  $\eta$ ?  
(b) If  $T$  is not known, we proceed as follows. We run the algorithm for 2, then 4, then 8, etc. iterations, resetting  $\eta$  at the beginning of each of these intervals. Hence, whatever  $T$  turns out to be, we will eventually reach it with a certain value of  $\eta$ . Show that the regret of this procedure is at most  $\frac{\sqrt{2}}{\sqrt{2}-1}$  times the regret guarantee in 4a. We conclude that the price for not knowing the time horizon is a multiplicative factor of about 3.41 for any algorithm with a bound (1) on its regret.