ELEN6887
Homework 3
Due February 26th, 2010

In lecture 4 we consider the estimation of a smooth function using a deterministic design (i.e. the sample locations $x_i$ were deterministic). In this homework we will put together some of the ideas of lectures 4 and 5 to extend to results to a random design setting.

Let $\mathcal{X} = [0, 1]$ and $\mathcal{Y} = [-r, r]$, where $r > 0$ is known. Let $P_{XY}$ be a joint probability distribution over $\mathcal{X} \times \mathcal{Y}$. Suppose you have $n$ i.i.d. samples for $P_{XY}$, $D_n = \{X_i, Y_i\}_{i=1}^n$. We will use $D_n$ to construct a consistent prediction rule $\hat{f}_n$ such that the expected excess risk with respect to the quadratic loss has a fast decay rate (with $n$). That is

$$E[R(\hat{f}_n)] - R^* = E[(\hat{f}_n(X) - Y)^2] - R^* \rightarrow 0,$$

at a fast pace as $n \rightarrow \infty$.

Let $f^*(x) = E[Y|X = x]$ be the regression function (the “best” prediction rule possible). Recall that $R(f^*) = R^*$ and

$$E[(\hat{f}_n(X) - Y)^2] - R^* = E[(\hat{f}_n(X) - f^*(X))^2].$$

Assume $f^*$ is a Lipschitz function with Lipschitz constant $L > 0$ (i.e. $|f^*(x) - f^*(y)| \leq L|x - y|$ $\forall x, y \in [0, 1]$). Finally define the estimator

$$\hat{f}_n(x) = \sum_{j=1}^m \hat{c}_j 1\{x \in I_j\},$$

where $I_j = [\frac{j-1}{m}, \frac{j}{m})$, and

$$\hat{c}_j = \left\{ \begin{array}{ll}
\frac{\sum_{i=1}^n Y_i 1\{X_i \in I_j\}}{\sum_{i=1}^n 1\{X_i \in I_j\}} & \text{if } \sum_{i=1}^n 1\{X_i \in I_j\} > 0 \\
0 & \text{otherwise}
\end{array} \right.$$

We will proceed by carefully decomposing the excess risk $E[(\hat{f}_n(X) - f^*(X))^2]$ into and estimation and approximation error (and also a cross-term). Let $f$ be an arbitrary prediction rule. It is easy to show that

$$E[(\hat{f}_n(X) - f^*(X))^2] \leq E[(\hat{f}_n(X) - \hat{f}(X))^2] + E[(\hat{f}(X) - f^*(X))^2]$$

$$+ 2 \sqrt{E[(\hat{f}_n(X) - \hat{f}(X))^2]} E[(\hat{f}(X) - f^*(X))^2].$$
where this result follows from the application of Cauchy-Schwarz’s inequality. The “best” approximating function \( \bar{f} \) we will use in this case is simply

\[
\bar{f}(x) = \sum_{j=1}^{m} \bar{c}_j 1\{x \in I_j\},
\]

where

\[
\bar{c}_j = \begin{cases} \frac{\int_{I_j} f^*(x) dP_X(x)}{\int_{I_j} dP_X(x)} & \text{if } \int_{I_j} dP_X(x) > 0 \\ 0 & \text{otherwise} \end{cases}
\]

a) Give an upper bound on the approximation error \( E[(\bar{f}(X) - f^*(X))^2] \). (\textbf{Hint:} this is almost analogous to what we did in lecture 4).

b) Give an upper bound on the estimation error \( E[(\hat{f}_n(X) - \bar{f}(X))^2] \). For this you will need to use a similar approach as used in lecture 5, by conditioning on the number of sample points that fall inside a bin. Start by examining \( E[(\hat{f}_n(x) - \bar{f}(x))^2] \) for an arbitrary \( x \in [0,1] \), and then proceed with the bound on \( E[(\hat{f}_n(X) - \bar{f}(X))^2] \).

(\textbf{Hint:} you will find the following fact quite useful - for a Binomial random variable \( N \sim \text{Binomial}(n,p) \) we have \( E\left[\frac{1}{N+1}\right] \leq \frac{1}{(n+1)p} \). This implies that \( E\left[\frac{1}{N+1} 1\{N > 0\}\right] \leq \frac{2}{(n+1)p} \).

c) Given your answers to the previous questions what is the proper choice of \( m \) as a function of \( n \)? What is a bound on the rate of excess risk decay of the procedure provided \( m \) is chosen appropriately? How does this compare with the results of lecture 4?

\textbf{Possible extensions: } You can get essentially the same results without assuming \( \mathcal{Y} \) is bounded, and instead assuming \( E[(Y - f^*(x))^2|X = x] \leq \sigma^2 < \infty \). This allows us to consider unbounded observation noise (for example Gaussian noise).