ECE901 Spring 2007 Statistical Learning Theory

Instructor: R. Nowak

Lecture 12: Complexity Regularization for Squared Error Loss

1 Complexity Regularization in Regression

The Chernoff/Hoeffding bounds were central to our analysis of classifier errors. Hoeffding's inequality states that for a sum of i.i.d. random variables $0 \le L_i \le 1, i = 1, ..., n$

$$P\left(\frac{1}{n}\sum_{i=1}^{n}E[L_i] - L_i > \epsilon\right) \le e^{-2n\epsilon^2}$$

If $L_i = \ell(f(X_i), Y_i)$, the loss of f in the prediction of Y_i from X_i , then we have

$$P\left(R(f) - \widehat{R}(f) > \epsilon\right) \le e^{-2n\epsilon^2}$$

When considering collection of candidate predictors, the union bound is used to obtain the following: with probability at least $1 - \delta$

$$R(f) \leq \widehat{R}(f) + \sqrt{\frac{\log |\mathcal{F}| + \log(1/\delta)}{2n}} , \quad \forall f \in \mathcal{F}$$

Taking \hat{f}_n to be the minimizer of the upper bound above, with $\delta = 1/\sqrt{n}$, leads to the following bound on the expected excess risk of \hat{f}_n :

$$E[R(\widehat{f}_n)] - \min_{f \in \mathcal{F}} R(f) \leq \sqrt{\frac{\log |\mathcal{F}| + \log n + 2}{n}}$$

More generally, if we have a countable collection of predictors and penalties c(f) assigned to each $f \in \mathcal{F}$ that satisfy the summability condition $\sum_{f \in \mathcal{F}} 2^{-c(f)} \leq 1$, then we showed that

$$E[R(\hat{f}_n)] - R^* \le \min_{f \in \mathcal{F}} \left\{ R(f) - R^* + \sqrt{\frac{c(f)\log 2 + \frac{1}{2}\log n}{2n}} + \frac{1}{\sqrt{n}} \right\}.$$

Consider the two terms in this upper bound: $R(f) - R^*$ is a bound on the approximation error of a model f, and remainder is a bound on the estimation error associated with f. Thus, we see that complexity regularization automatically optimizes a balance between approximation and estimation errors.

Note that the upper bound is at least $n^{-1/2}$. This is the best one can expect, in general, when considering the 0/1 or ℓ_1 (absolute error) loss functions, but in regression we are often interested in the squared error or ℓ_2^2 loss (corresponding to the mean square error risk). The squared error decays faster than the 0/1 or absolute error (since squaring small numbers makes them smaller yet). Unfortunately, the Chernoff/Hoeffding bounds are not capable of handling such cases, and more sophisticated techniques are required. Before delving into those methods, consider the following simple example.

Example 1 To illustrate the distinction between classification and regression, consider a simple, scalar signal plus noise problem. Consider $Y_i = \theta + W_i$, i = 1, ..., n, where θ is a fixed unknown scalar parameter

and the W_i are independent, zero-mean, unit variance random variables. Let $\bar{Y} = 1/n \sum_{i=1}^{n} Y_i$. Then we have

$$E[|\bar{Y} - \theta|^2] = E\left[\left(\frac{1}{n}\sum_{i=1}^n W_i\right)^2\right]$$
$$= \frac{1}{n^2}\sum_{i=1}^n E[W_i^2] = n^{-1}$$

Thus, the mean square error decays like n^{-1} , notably faster than $n^{-1/2}$. The convergence rate of n^{-1} is called the parametric rate, since it is the rate at which the MSE decays in simple parametric inference. A similar conclusion can be arrived at through a large deviation analysis. According to the Central Limit Theorem, \bar{Y} is distributed approximately $N(\theta, 1/n)$. A simple tail-bound on the Gaussian distribution gives us

$$P(\bar{Y} - \theta > \epsilon) = P(W > \epsilon) \le \frac{1}{2}e^{-n\epsilon^2/2}$$

which implies that

$$P(|\bar{Y} - \theta|^2 > \epsilon) \leq e^{-n\epsilon/2}$$

This is a bound on the deviations of the squared error $|\bar{Y} - \theta|^2$. The squared error concentration inequality implies that $E[|\bar{Y} - \theta|^2] = O(\frac{1}{n})$ (just write $E[(\bar{Y} - \theta)^2] = \int_0^\infty P((\bar{Y} - \theta)^2 > t) dt$).

1.1 Risk Bounds for Squared Error Loss

Based on the example above, we hope to achieve a risk bound for squared error loss of the form

$$E[R(\widehat{f}_n)] - R^* \leq C \min_{f \in \mathcal{F}} \left\{ R(f) - R^* + \frac{c(f)\log 2 + \frac{1}{2}\log n}{2n} \right\},$$

where C > 0 is a constant. That is, the bound on the estimation error should be $O(c(f)n^{-1})$, rather than $O(\sqrt{c(f)n^{-1}})$. To begin our investigation into regression and function estimation, let us consider the following. Let $\mathcal{X} = \mathbf{R}^d$ and $\mathcal{Y} = \mathbf{R}$. Take \mathcal{F} such that $f \in \mathcal{F}$ is a map $f : \mathbf{R}^d \to \mathbf{R}$. We have training data $\{X_i, Y_i\}_{i=1}^n \overset{i.i.d.}{\sim} P_{XY}$. As our loss function, we take the squared error

$$l(f(X_i), Y_i) = (f(X_i) - Y_i)^2$$

The empirical risk function is simply the sum of squared prediction errors

$$\widehat{R}(f) = \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - Y_i)^2$$

The risk is then the MSE

$$R(f) = E[(f(X) - Y)^2].$$

We know that the function f^* that minimizes the MSE is just the conditional expectation of Y given X:

$$f^* = E[Y|X = x].$$

Now let $R^* = R(f^*)$. We would like to select an $\hat{f}_n \in \mathcal{F}$ using the training data $\{X_i, Y_i\}_{i=1}^n$ such that the excess risk

$$E[R(\widehat{f_n})] - R^* \ge 0$$

is small. Let's consider the difference between the empirical risks:

$$\widehat{R}(f) - \widehat{R}(f^*) = \frac{1}{n} \sum_{i=1}^n (f(X_i) - Y_i)^2 - \frac{1}{n} \sum_{i=1}^n (f^*(X_i) - Y_i)^2.$$

Lecture 12: Complexity Regularization for Squared Error Loss

Note that $E[\widehat{R}(f) - \widehat{R}(f^*)] = R(f) - R(f^*)$. Hence, by the SLLN, we know that

$$\widehat{R}(f) - \widehat{R}(f^*) \to R(f) - R(f^*)$$

as $n \to \infty$. But how fast is this convergence?

We will derive a bound for the difference $[R(f) - R(f^*)] - [\widehat{R}(f) - \widehat{R}(f^*)]$. The following derivation is due to Andrew Barron¹. The excess risk and it empirical counterpart will be denoted by

$$\begin{aligned} \mathcal{E}(f) &:= R(f) - R(f^*) \\ \widehat{\mathcal{E}}(f) &:= \widehat{R}(f) - \widehat{R}(f^*) \end{aligned}$$

Note that $\widehat{\mathcal{E}}(f)$ is the sum of independent random variables:

$$\widehat{\mathcal{E}}(f) = -\frac{1}{n} \sum_{i=1}^{n} U_i$$

where $U_i = -(Y_i - f(X_i))^2 + (Y_i - f^*(X_i))^2$. Therefore, $\mathcal{E}(f) - \widehat{\mathcal{E}}(f) = \frac{1}{n} \sum_{i=1}^n (U_i - E[U_i])$. We are looking for a bound of the form

$$P(\mathcal{E}(f) - \widehat{\mathcal{E}}(f) > \epsilon) < \delta.$$

If the variables U_i are bounded, then we can apply Hoeffding's inequality. However, a more useful bound for our regression problem can be derived if the the variables U_i satisfy the following moment condition:

$$E[|U_i - E[U_i]|^k] \le \frac{var(U_i)}{2} \ k! \ h^{k-2}$$
(1)

for some h > 0.

The moment condition can be difficult to verify in general, but it does hold, for example, for bounded random variables. If (1) holds, then the Craig-Bernstein (CB) inequality (Craig 1933) states:

$$P\left(\frac{1}{n}\sum_{i=1}^{n}(U_i - E[U_i]) \ge \frac{t}{n\epsilon} + \frac{n\epsilon \ var(\frac{1}{n}\sum U_i)}{2(1-c)}\right) \le e^{-t},$$

for $0 < \epsilon h \le c < 1$ and t > 0. This shows that the tail decays exponentially in t, rather than exponentially in t^2 . Recall Hoeffding's inequality:

$$P\left(\frac{1}{n}\sum_{i=1}^{n}(Z_i-E[Z_i])\geq \frac{t}{n}\right)\leq e^{\frac{-2t^2}{n}}.$$

If $\frac{t}{n} \ll 1$, then $\frac{t^2}{n} \ll t$, which implies $e^{\frac{-2t^2}{n}} \gg e^{-t}$. This indicates that the CB inequality may be much tighter than Hoeffding's, when the variance term $\frac{n\epsilon \ var(\frac{1}{n} \sum U_i)}{2(1-c)}$ is small. To use the CB inequality, we need to bound the variance of $\frac{1}{n} \sum_{i=1}^{n} U_i$. Note that

$$var(U_i) = var(-(Y_i - f(X_i))^2 + (Y_i - f^*(X_i))^2).$$

Assumption 1 The support of Y and the range f(X) is in a known interval of length b.

Proposition 1 With the above assumption, (1) holds with $h = \frac{2b^2}{3}$.

¹A. R. Barron, "Complexity regularization with application to artificial neural networks," in *Nonparametric Functional Estimation and Related Topics*. Kluwer Academic Publishers, 1991, pp. 561-576.

Lecture 12: Complexity Regularization for Squared Error Loss

Proposition 2 Again, with the above assumption, it may be shown that

$$var(U_i) \le 5b^2 \mathcal{E}(f) \tag{2}$$

Proof 1 You can write U_i as

$$U_{i} = 2Y_{i}f(X_{i}) - 2Y_{i}f^{*}(X_{i}) + f^{*}(X_{i})^{2} - f(X_{i})^{2}$$

= $2Y_{i}f(X_{i}) - 2Y_{i}f^{*}(X_{i}) + 2f^{*}(X_{i})^{2} - f^{*}(X_{i})^{2} - f(X_{i})^{2} + 2f(X_{i})f^{*}(X_{i}) - 2f(X_{i})f^{*}(X_{i})$
= $2(Y_{i} - f^{*}(X_{i}))(f(X_{i}) - f^{*}(X_{i})) - (f(X_{i}) - f^{*}(X_{i}))^{2}$

Note that the variance of U_i is upper-bounded by its second moment. Also note that the covariance of the two terms above is zero:

$$E[2(Y_i - f^*(X_i))(f(X_i) - f^*(X_i))(f(X_i) - f^*(X_i))^2] = E[T_1T_2]$$

= $E_X[E_{Y|X}[T_1T_2]]$
= $E_X[T_2E_{Y|X}[T_1]]$
= $E_X[T_2 + 0]$
= 0

This is evident when you recall that $f^*(X_i) = E[Y|X = X_i]$. Now we can bound the second moments of T_1 and T_2 :

$$E[T_1] = 4E[((Y_i - f^*(X_i))(f(X_i) - f^*(X_i)))^2]$$

= $4E[(Y_i - f^*(X_i))^2(f(X_i) - f^*(X_i))^2]$
 $\leq 4E[b^2(f(X_i) - f^*(X_i))^2]$
 $E[T_2] = E[(f(X_i) - f^*(X_i))^4]$
= $E[(f(X_i) - f^*(X_i))^2(f(X_i) - f^*(X_i))^2]$
 $\leq E[b^2(f(X_i) - f^*(X_i))^2]$

So $var(U_i) \leq 5b^2 E[(f(X_i) - f^*(X_i))^2]$. The final step is to see that

$$\mathcal{E}(f) = E[U_i] = E_X[E_{Y|X}[U_i]] = E[(f(X_i) - f^*(X_i))^2].$$

Thus, $n \, var(\frac{1}{n} \sum_{i=1}^{n} U_i) \leq 5b^2 \mathcal{E}(f)$. And therefore, we can say that, with probability at least $1 - e^{-t}$,

$$\mathcal{E}(f) - \widehat{\mathcal{E}}(f) \le \frac{t}{n \epsilon} + \frac{5\epsilon b^2 \mathcal{E}(f)}{2(1-c)}.$$

In other words, with probability at least $1 - \delta$ (where $\delta = e^{-t}$),

$$\mathcal{E}(f) - \widehat{\mathcal{E}}(f) \le \frac{\log \frac{1}{\delta}}{n \ \epsilon} + \frac{5\epsilon \ b^2 \ \mathcal{E}(f)}{2(1-c)}.$$
(3)

Now, suppose we have assigned positive numbers c(f) to each $f \in \mathcal{F}$ satisfying the Kraft inequality:

$$\sum_{f \in \mathcal{F}} 2^{-c(f)} \le 1.$$

Note that (3) holds $\forall \delta > 0$. In particular, we let δ be a function of f:

$$\delta(f) = 2^{-c(f)}\delta.$$

So we can use this δ along with the procedure introduced in Lecture 9 (i.e., the union bound followed by the Kraft inequality) to obtain the following. For any $\delta > 0$, with probability at least $1 - \delta$

$$\mathcal{E}(f) - \widehat{\mathcal{E}}(f) \leq \frac{c(f)\log 2 + \log \frac{1}{\delta}}{n \epsilon} + \frac{5\epsilon b^2 \mathcal{E}(f)}{2(1-c)} \quad , \quad \forall f \in \mathcal{F}$$

$$\tag{4}$$

Now set $c = \epsilon \ h = \frac{2b^2 \ \epsilon}{3}$ and assume $\epsilon < \frac{6}{19b^2}$. Then define

$$\alpha = \frac{5\epsilon \ b^2}{2(1-c)} < 1.$$

Now, after using α and rearranging terms, we have

$$(1-\alpha)\mathcal{E}(f) \leq \widehat{\mathcal{E}}(f) + \frac{c(f)\log 2 + \log \frac{1}{\delta}}{\epsilon n}$$

Let us choose f to minimize this upper bound. Recall that $\widehat{\mathcal{E}}(f) = \widehat{R}(f) - \widehat{R}(f^*)$, and so

$$\widehat{f}_n = \arg\min_{f\in\mathcal{F}} \left\{ \widehat{R}(f) + \frac{c(f)\log 2}{n\epsilon} \right\}$$

minimizes the upper bound. Thus, with probability at least $1 - \delta$,

$$(1-\alpha)\mathcal{E}(\widehat{f}_n) \leq \widehat{\mathcal{E}}(\widehat{f}_n) + \frac{c(\widehat{f}_n)\log 2 + \log\frac{1}{\delta}}{\epsilon n} \\ \leq \widehat{\mathcal{E}}(f_n^*) + \frac{c(\widehat{f}_n^*)\log 2 + \log\frac{1}{\delta}}{\epsilon n}$$
(5)

where $f_n^* = \arg \min_{f \in \mathcal{F}} \left\{ R(f) + \frac{c(f) \log 2}{n\epsilon} \right\}$.

Now we use the Craig-Bernstein inequality to bound the difference between $\widehat{\mathcal{E}}(f_n^*)$ and $\mathcal{E}(f_n^*)$. With probability at least $1 - \delta$,

$$\widehat{\mathcal{E}}(f_n^*) \leq \mathcal{E}(f_n^*) + \alpha \, \mathcal{E}(f_n^*) + \frac{\log(\frac{1}{\delta})}{n\epsilon}.$$
(6)

Now we can again use the union bound to combine (5) and (6). For any $\delta > 0$, with probability at least $1-2\delta$,

$$\mathcal{E}(\widehat{f}_n) \leq \frac{1+\alpha}{1-\alpha} \mathcal{E}(f_n^*) + \frac{c(f_n^*)\log 2 + 2\log 1/\delta}{n\epsilon}.$$

Now set $\delta = e^{\frac{-n\epsilon t}{2}}$, then we have

$$P\left(\mathcal{E}(\widehat{f}_n) - \frac{1+\alpha}{1-\alpha}\mathcal{E}(f_n^*) + \frac{c(f_n^*)\log 2}{n\epsilon} \ge t\right) \le 2e^{\frac{-n\epsilon t}{2}}.$$

Integrating, we get

$$E\left[\mathcal{E}(\widehat{f}_n) - \frac{1+\alpha}{1-\alpha}\mathcal{E}(f_n^*) + \frac{c(f_n^*)\log 2}{n\epsilon}\right] \leq \int_0^\infty P(" \geq t) dt$$
$$\leq \int_0^\infty 2e^{\frac{-n\epsilon t}{2}}$$
$$= \frac{4}{n\epsilon}$$

To sum up, we have shown that for $\epsilon < \frac{6}{19b^2}$ we have $\alpha < 1$ and

$$E[\mathcal{E}(\widehat{f}_n)] \leq \left(\frac{1+\alpha}{1-\alpha}\right) \mathcal{E}(f_n^*) + \frac{c(f_n^*)\log 2 + 4}{n\epsilon} \\ = \left(\frac{1+\alpha}{1-\alpha}\right) \min_{f \in \mathcal{F}} \left\{ \mathcal{E}(f) + \frac{c(f)\log 2}{n\epsilon} \right\} + \frac{4}{n\epsilon}$$

Or, in expanded form:

$$E[R(\widehat{f}_n)] - R(f^*) \leq \left(\frac{1+\alpha}{1-\alpha}\right) \min_{f \in \mathcal{F}} \left\{ R(f) - R(f^*) + \frac{c(f)\log 2}{n\epsilon} \right\} + \frac{4}{n\epsilon}$$

Notice that if $f^* \in \mathcal{F}$ and if $c(f^*)$ is not too large (e.g., $c(f^*) \approx \log n$), then we have $E[R(\widehat{f_n})] - R(f^*) = O(n^{-1} \log n)$, within a logarithmic factor of the parametric rate of convergence!