ECE 901
Lecture 12: Complexity Regularization and the Squared Loss

R. Nowak

5/17/2009

In the previous lectures we made use of the Chernoff/Hoeffding bounds for our analysis of classifier errors. Hoeffding's inequality states that for a sum of independent random variables $0 \leq L_i \leq 1$, $i = 1, \ldots, n$

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

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

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

When considering a countable collection $\mathcal{F}$ of candidate 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

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

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 bound is valid for *any* bounded loss function.

The above 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 typically 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 $Y_i = \theta + W_i$, $i = 1, \ldots, n$, where θ is a fixed unknown scalar parameter and the W_i are independent, zero-mean, unit variance random variables. Let $\hat{\theta} = 1/n \sum_{i=1}^n Y_i$. Then we have*

$$\begin{aligned} E[|\hat{\theta} - \theta|^2] &= E\left[\left(\frac{1}{n}\sum_{i=1}^{n} W_i\right)^2\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n} E[W_i^2] \;=\; n^{-1} \end{aligned}$$

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 *for regression, since it is the rate at which the MSE decays in simple parametric inference. A similar conclusion can be arrived at through large deviation analysis. According to the Central Limit Theorem*

$$\sqrt{n}(\hat{\theta} - \theta) \stackrel{dist}{\rightarrow} N(0,1) \ ,$$

as $n \rightarrow \infty$. A simple tail-bound on the Gaussian distribution gives us

$$P(\sqrt{n}(\hat{\theta} - \theta) > t) \leq \frac{1}{2} e^{-t^2/2} \ ,$$

for large n, which implies that

$$P((\hat{\theta} - \theta)^2 > \epsilon) \leq \frac{1}{2} e^{-n\epsilon/2} \ .$$

This is a bound on the deviations of the squared error $|\hat{\theta} - \theta|^2$. The squared error concentration inequality implies that $E[|\hat{\theta} - \theta|^2] = O(\frac{1}{n})$ (just write $E[(\hat{\theta} - \theta)^2] = \int_0^\infty P((\hat{\theta} - \theta)^2 > t) dt$). Note that the main difference between Hoeffding's inequality and the above concentration bound is the dependence in ϵ, linear in the latter and quadratic in the former, therefore much weaker in the former case.

1 Risk Bounds for Squared Error Loss

Let $\mathcal{X}$ be the feature space (e.g., $\mathcal{X} = \mathbf{R}^d$) and $\mathcal{Y} = [-b/2, b/2]$, where $b > 0$ is known. In other words, assume the label space is bounded. Consider the squared error loss $\ell(y_1, y_2) = (y_1 - y_2)^2$. Take $\mathcal{F}$ such that $f \in \mathcal{F}$ is a map $f : \mathcal{X} \rightarrow \mathcal{Y}$. We have training data $\{X_i, Y_i\}_{i=1}^n \stackrel{i.i.d.}{\sim} P_{XY}$. 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 therefore 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 (also known as the regression function):

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

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

$$E[R(\widehat{f}_n)] - R^* \geq 0$$

is small.

Like we did in lecture 9 we will take advantage of the fact that $\widehat{R}(f)$ concentrates (in probability) around $R(f)$, but to take advantage of the particular aspects of the squared error loss it is convenient to look at "relative" versions of the risk, namely the excess risk and its empirical counterpart

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

The first thing to note is that, as shown in lecture 2

$$\mathcal{E}(f) = E[(f(X) - f^*(X))^2] \ .$$

Furthermore note that $E[\widehat{\mathcal{E}}(f)] = \mathcal{E}(f)$, and 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, $\mathcal{E}(f) - \widehat{\mathcal{E}}(f) = \frac{1}{n}\sum_{i=1}^{n}(U_i - E[U_i])$. Clearly the strong law of large numbers tells us that for a fixed prediction rule f

$$\widehat{\mathcal{E}}(f) \to \mathcal{E}(f) \ ,$$

as $n \to \infty$. All we need to know if to determine the "speed" of 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 Barron[1]. 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 independent and 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:

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

for some $h > 0$ and all $k \geq 2$.

The moment condition can be difficult to verify in general, but it does hold, for example, for bounded random variables. In that case the Craig-Bernstein (CB) inequality (Craig 1933) states that, for independent r.v.'s U_i satisfying (1):

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

for $0 < \epsilon h \leq c < 1$ 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

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

Recall our assumption that $\mathcal{Y}$ is bounded, in particular that $\mathcal{Y}$ is contained in a interval of length b (without loss of generality we can assume $\mathcal{Y} = [-b/2, b/2]$).

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

Proposition 1. *The moment condition* (1)*) holds with* $h = \frac{2b^2}{3}$.

Proof. Left as an exercise. ■

Proposition 2. *The variance of* U_i *satisfies*

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

Proof. We can write U_i as

$$\begin{aligned} U_i &= 2Y_i f(X_i) - 2Y_i f^*(X_i) + f^*(X_i)^2 - f(X_i)^2 \\ &= 2Y_i f(X_i) - 2Y_i f^*(X_i) + 2f^*(X_i)^2 - f^*(X_i)^2 - f(X_i)^2 + 2f(X_i)f^*(X_i) - 2f(X_i)f^*(X_i) \\ &= \underbrace{2\left(Y_i - f^*(X_i)\right)\left(f(X_i) - f^*(X_i)\right)}_{T_1} - \underbrace{\left(f(X_i) - f^*(X_i)\right)^2}_{T_2} . \end{aligned}$$

Note that the variance of U_i is upper-bounded by its second moment, that is

$$var(U_i) \leq E[U_i^2] = E[(T_1 - T_2)^2] = E[T_1^2] + E[T_2^2] - 2E[T_1T_2] .$$

Also note that the covariance of T_1 and T_2 is zero:

$$\begin{aligned} E[2\left(Y_i - f^*(X_i)\right)\left(f(X_i) - f^*(X_i)\right)\left(f(X_i) - f^*(X_i)\right)^2] &= E[T_1T_2] \\ &= E[E[T_1T_2]|X_i]] \\ &= E[T_2E[T_1]|X_i]] \\ &= E[2T_2(f(X_i) - f^*(X_i))E[Y_i - f^*(X_i)|X_i]] \\ &= 0 . \end{aligned}$$

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

$$\mathcal{E}(f) = E[\left(f(X) - f^*(X)\right)^2] .$$

. Now

$$\begin{aligned} E[T_1^2] &= 4E[((Y_i - f^*(X_i))(f(X_i) - f^*(X_i)))^2] \\ &= 4E[(Y_i - f^*(X_i))^2(f(X_i) - f^*(X_i))^2] \\ &\leq 4E[b^2(f(X_i) - f^*(X_i))^2] \\ &= 4b^2\mathcal{E}(f) , \\ E[T_2^2] &= E[(f(X_i) - f^*(X_i))^4] \\ &= E[\left(f(X_i) - f^*(X_i)\right)^2\left(f(X_i) - f^*(X_i)\right)^2] \\ &\leq E[b^2\left(f(X_i) - f^*(X_i)\right)^2] \\ &= b^2\mathcal{E}(f) . \end{aligned}$$

So $var(U_i) \leq 5b^2E[\left(f(X_i) - f^*(X_i)\right)^2]$. ■

Thus, $n\ var(\frac{1}{n}\sum_{i=1}^n U_i) \leq 5b^2\mathcal{E}(f)$. Using the CB inequality (for properly chosen values of ϵ and c, to be discussed later) we have that, with probability at least $1 - e^{-t}$,

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

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

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

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

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

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 define

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

Taking $\epsilon < \frac{6}{19b^2}$ guarantees that $\alpha < 1$. Using this fact we have

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

with probability at least $1 - \delta$.

Since we want to find $f \in \mathcal{F}$ that minimizes $\mathcal{E}(f)$ it is a good bet to minimize the right-hand-side of the above bound. Recall that $\widehat{\mathcal{E}}(f) = \widehat{R}(f) - \widehat{R}(f^*)$, and so define

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

so that $\widehat{f}_n$ minimizes the upper bound. Thus, with probability at least $1 - \delta$,

$$\begin{aligned}(1-\alpha)\mathcal{E}(\widehat{f}_n) \;&\leq\; \widehat{\mathcal{E}}(\widehat{f}_n) + \frac{c(\widehat{f}_n)\log 2 + \log\frac{1}{\delta}}{\epsilon\; n} \\ &\leq\; \widehat{\mathcal{E}}(\tilde{f}) + \frac{c(\tilde{f})\log 2 + \log\frac{1}{\delta}}{\epsilon\; n}\end{aligned} \tag{5}$$

where $\tilde{f} \in \mathcal{F}$ is arbitrary (but not a function of the training data).

Now we use the Craig-Bernstein inequality to bound the difference between $\widehat{\mathcal{E}}(\tilde{f})$ and $\mathcal{E}(\tilde{f})$. In order to get the correct direction in the bound we will apply CB to $-U_i$ instead (very similar derivation as before). With probability at least $1 - \delta$,

$$\widehat{\mathcal{E}}(\tilde{f}) \;\leq\; \mathcal{E}(\tilde{f}) + \alpha\;\mathcal{E}(\tilde{f}) + \frac{\log(\frac{1}{\delta})}{n\epsilon}. \tag{6}$$

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

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

At this point we have shown the following PAC bound.

Theorem 1. *Consider the squared error loss. Let $\mathcal{X}$ be the feature space and $\mathcal{Y} = [-b/2, b/2]$ be the label space. Let $\{X_i, Y_i\}_{i=1}^n$ be i.i.d. according to P_{XY}, unknown. Let $\mathcal{F}$ be a collection of predictors (i.e. $f \in \mathcal{F}$ are functions $f : \mathcal{X} \to \mathcal{Y}$) such that there are numbers $c(f)$ satisfying $\sum_{f \in \mathcal{F}} 2^{-c(f)} \leq 1$.*

Select a function $\widehat{f}_n$ *according to*

$$\widehat{f}_n = \arg\min_{f\in\mathcal{F}} \left\{ \widehat{R}_n(f) + \frac{1}{\epsilon}\frac{c(f)\log 2}{n} \right\} ,$$

with $0 < \epsilon < \frac{6}{19b^2}$. *Then, for any* $\delta > 0$ *with probability at least* $1 - 2\delta$

$$\mathcal{E}(\widehat{f}_n) \;\le\; \frac{1+\alpha}{1-\alpha}\mathcal{E}(f) + \frac{1}{\epsilon}\frac{c(f)\log 2 + 2\log 1/\delta}{(1-\alpha)n} \quad \forall f \in F ,$$

where $\alpha = \frac{15b^2\epsilon}{6-2b^2\epsilon}$, *and*

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

Finally we can use this result to get a bound on the expected excess risk. Although this result is just a corollary of the above theorem we will state is as a theorem due to its importance.

Theorem 2. *Under the conditions of the above theorem*

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

Proof. Let

$$\tilde{f} = \inf_{f\in\mathcal{F}} \left\{ \frac{1+\alpha}{1-\alpha}\mathcal{E}(f) + \frac{1}{\epsilon}\frac{c(f)\log 2}{(1-\alpha)n} \right\} ,$$

and define

$$\Phi(\widehat{f}_n) = \mathcal{E}(\widehat{f}_n) - \frac{1+\alpha}{1-\alpha}\mathcal{E}(\tilde{f}) - \frac{1}{\epsilon}\frac{c(\tilde{f})\log 2}{(1-\alpha)n} .$$

The previous theorem implies that

$$\Pr\left(\Phi(\widehat{f}_n) > \frac{2}{\epsilon}\frac{\log(1/\delta)}{(1-\alpha)n} \right) \le 2\delta .$$

Take $\delta = e^{\frac{-(1-\alpha)n\epsilon\, t}{2}}$. Then

$$\begin{aligned} E\left[\Phi(\widehat{f}_n)\right] &\le \int_0^\infty P(\Phi(\widehat{f}_n) \ge t)\, dt \\ &\le \int_0^\infty 2e^{\frac{-(1-\alpha)n\epsilon t}{2}} \\ &= \frac{4}{(1-\alpha)n\epsilon} , \end{aligned}$$

concluding the proof. ∎

As a final remark notice that the above bound can be much better than the one derived for general losses, in particular 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!