ECE 901
Lecture 10: Expected Error Bounds for Complexity Regularization

R. Nowak

5/17/2009

1 Recap: PAC Bounds for Countably Infinite Classes of Models

In the previous class we have shown a PAC bound that applies to countable classes of models and bounded loss functions. In particular let $\mathcal{X}$ be the feature space and $\mathcal{Y}$ be the label space, and assume $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$. Let $\{X_i, Y_i\}_{i=1}^n$ be an i.i.d. sample from unknown P_{XY}, and $f : \mathcal{X} \to \mathcal{Y}$ be an arbitrary prediction rule. Define

$$\hat{R}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i), Y_i) \ ,$$

and

$$R(f) = E[\ell(f(X), Y)] \ ,$$

where $(X, Y) \sim P_{XY}$.

Recall our basic step in derivation of the basic PAC bounds: For a fixed model $f \in \mathcal{F}$ and for any $\delta(f) > 0$, with probability at least $1 - \delta(f)$

$$R(f) \leq \widehat{R}_n(f) + \sqrt{\frac{\log(1/\delta(f))}{2n}} \ .$$

Now if the class $\mathcal{F}$ is finite we can take $\delta(f) = \delta \frac{1}{|\mathcal{F}|}$, which yields $\log(1/\delta(f)) = \log|\mathcal{F}| + \log(1/\delta)$, and applying the union of events bound yields that for all $\delta > 0$, with probability at least $1 - \delta$

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

The main idea is that we have a "budget" δ to make errors, and now we can distribute it in any way we want over the model space, so that $\sum_{f \in \mathcal{F}} \delta(f) \leq \delta$. If the class of models is finite we can allocate an equal part of the budget to each model, but that is not the only way to do it. In fact we only require

$$\sum_{f \in \mathcal{F}} \delta(f) \quad \leq \quad \delta \ .$$

So, if $p(f)$ are positive numbers satisfying $\sum_{f \in \mathcal{F}} p(f) = 1$, then we can take $\delta_f = p(f)\delta$. This provides two advantages:

1. By choosing $p(f)$ larger for certain f, we can preferentially treat those candidates
2. We do not need $\mathcal{F}$ to be finite and we only require $\sum_{f \in \mathcal{F}} p(f) = 1$

Prefix codes are one way to construct $p(f)$. If we assign a binary prefix code of length $c(f)$ to each $f \in \mathcal{F}$, then the values $p(f) = 2^{-c(f)}$ satisfy $\sum_{f \in \mathcal{F}} p(f) \leq 1$ according to the Kraft inequality.

2 Complexity Regularization Bounds

The main point of this lecture is to examine how PAC bounds of the form w.p. $\geq 1-\delta$

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

can be used to select a model that comes close to achieving the best possible performance of any model in the class, that is

$$\inf_{f\in\mathcal{F}} R(f)$$

Let $\widehat{f}_n$ be the model selected from $\mathcal{F}$ using the training data $\{X_i, Y_i\}_{i=1}^n$. We will specify this model in a moment, but keep in mind that it is not necessarily the model with minimum empirical risk as before. We would like to have

$$E[R(\widehat{f}_n)] - \inf_{f\in\mathcal{F}} R(f)$$

as small as possible. First, for any $\delta > 0$, define

$$\widehat{f}_n^{\delta} \quad = \quad \arg\min_{f\in\mathcal{F}} \left\{ \widehat{R}_n(f) + C(f,n,\delta) \right\}$$

where

$$C(f,n,\delta) \equiv \sqrt{\frac{c(f)\log 2 + \log(1/\delta)}{2n}}$$

Then w.p. $\geq 1-\delta$

$$R(f) \leq \widehat{R}_n(f) + C(f,n,\delta) \quad , \quad \forall f \in \mathcal{F}$$

and in particular,

$$R(\widehat{f}_n^{\delta}) \leq \widehat{R}_n(\widehat{f}_n^{\delta}) + C(\widehat{f}_n^{\delta}, n, \delta) \ ,$$

so, by the definition of $\widehat{f}_n^{\delta}$, $\forall f \in \mathcal{F}$

$$R(\widehat{f}_n^{\delta}) \leq \widehat{R}_n(f) + C(f,n,\delta) \ .$$

We will make use of the inequality above in a moment. First note that for any fixed model $f \in \mathcal{F}$ we have $E[\widehat{R}_n(f)] = R(f)$. Therefore

$$E[R(\widehat{f}_n^{\delta})] - R(f) = E[R(\widehat{f}_n^{\delta}) - \widehat{R}_n(f)] \ .$$

Let Ω be the event on which

$$R(\widehat{f}_n^{\delta}) \leq \widehat{R}_n(f) + C(f,n,\delta), \ \forall \ f \in \mathcal{F} \ .$$

From the bound above, we know that $P(\Omega) \geq 1-\delta$. Thus,

$$\begin{aligned}
E[R(\widehat{f}_n^{\delta}) - \widehat{R}_n(f)] &= E[R(\widehat{f}_n^{\delta}) - \widehat{R}_n(f)|\Omega]P(\Omega) + E[R(\widehat{f}_n^{\delta}) - \widehat{R}_n(f)|\Omega^c](1-P(\Omega)) \\
&\leq C(f,n,\delta) + \delta \quad (\text{since } 0 \leq R(f), \widehat{R}_n(f) \leq 1, \ P(\Omega) \leq 1 \text{ and } 1-P(\Omega) \leq \delta) \\
&= \sqrt{\frac{c(f)\log 2 + \log(1/\delta)}{2n}} + \delta \\
&= \sqrt{\frac{c(f)\log 2 + \frac{1}{2}\log n}{2n}} + \frac{1}{\sqrt{n}} \quad (\text{by setting } \delta = \frac{1}{\sqrt{n}}) \ .
\end{aligned}$$

We can summarize our analysis with the following theorem.

Theorem 1 ***(Complexity Regularized Model Selection)*** *Let $\mathcal{F}$ be a countable collection of models, and assign a positive number $c(f)$ to each $f \in \mathcal{F}$ such that $\sum_{f\in\mathcal{F}} 2^{-c(f)} \leq 1$. Define the minimum complexity regularized risk model*

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

Then,

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

This shows that

$$\widehat{R}_n(f) + \sqrt{\frac{c(f)\log 2 + \frac{1}{2}\log n}{2n}}$$

is a reasonable surrogate for

$$R(f) + \sqrt{\frac{c(f)\log 2 + \frac{1}{2}\log n}{2n}}$$

Example 1 ***(Histogram Classifiers)*** *Let $\mathcal{X} = [0,1]^d$ be the input space and $\mathcal{Y} = \{0,1\}$ be the output space. Let $\mathcal{F}_k$, k = 1, 2, ... denote the collection of histogram classification rules with k hypercubical equal volume bins. One choice of prefix code for this example is: $k = 1 \Rightarrow code = 0, k = 2 \Rightarrow code = 10, k = 3 \Rightarrow code = 110$ and so on. Then, if first code is corresponding to $k \Rightarrow f \in \mathcal{F}_k$, followed by $k = \log_2 |\mathcal{F}_k|$ bits to indicate which of the 2^k histogram rules in $\mathcal{F}_k$ is under consideration, we have*

$$f \in \mathcal{F}_k \Rightarrow c(f) = 2k \ bits$$

Let $\widehat{f}_n$ be the model that solves the minimization i.e.,

$$\min_{k\geq 1} \left\{ \min_{f\in\mathcal{F}_k} \widehat{R}_n(f) + \sqrt{\frac{2k\log 2 + \frac{1}{2}\log n}{2n}} \right\}$$

That is, for each k, let

$$\widehat{f}_n^{(k)} \quad = \quad \arg\min_{f\in\mathcal{F}_k} \widehat{R}_n(f)$$

Then select the best k according to

$$\widehat{k} \quad = \quad \arg\min_{k\geq 1} \left\{ \widehat{R}_n(\widehat{f}_n^{(k)}) + \sqrt{\frac{2k\log 2 + \frac{1}{2}\log n}{2n}} \right\}$$

and set

$$\widehat{f}_n \quad = \quad \widehat{f}_n^{(\widehat{k})}$$

Then,

$$E[R(\widehat{f}_n)] \;\leq\; \inf_{k\geq 1}\left\{\min_{f\in\mathcal{F}_k} R(f) + \sqrt{\frac{2k\log 2 + \frac{1}{2}\log n}{2n}} + \frac{1}{\sqrt{n}}\right\}$$

It is a simple exercise (see homework) to show that if $d = 2$ and the Bayes decision boundary is a 1-d curve, then by setting $k = \sqrt{n}$ and selecting the best f from $\mathcal{F}_{\sqrt{n}}$ we have

$$E[R(\widehat{f}_n)] \;=\; O(n^{-1/4})$$

Note that the complexity regularized classifier $\widehat{f}_n$ adaptively achieves this rate, without user intervention.