ECE901 Spring 2004 Statistical Regularization and Learning Theory **Lecture: 10**

Complexity Regularization

Lecturer: Rob Nowak *Scribe: Waheed Bajwa*

1 Review: PAC Bounds

Consider a finite collection of candidate functions $\mathcal{F}$, and recall the basic PAC bound: $\forall f \in \mathcal{F}$ *and* $\forall \delta > 0,\ w.p. \geq 1-\delta$;

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

where

$$\begin{aligned} \widehat{R}_n(f) &= \frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i), Y_i) \\ R(f) &= E\left[\ell(f(X), Y)\right] \end{aligned}$$

Note that we can write the inequality above as:

$$R(f) \leq \widehat{R}_n(f) + \sqrt{\frac{\log\left(\frac{|\mathcal{F}|}{\delta}\right)}{2n}}$$

Letting $\delta_f = \frac{\delta}{|\mathcal{F}|}$, we have:

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

This is precisely the form of Hoeffding's inequality, with δ_f in place of the usual δ. In effect, in order to have Hoeffding's inequality hold with probability $1-\delta$ for all $f \in \mathcal{F}$, we must distribute "δ-budget" or "confidence-budget" over all $f \in \mathcal{F}$ (in this case, evenly distributed):

$$\begin{aligned} \sum_{f\in\mathcal{F}} \delta_f &= \sum_{f\in\mathcal{F}} \frac{\delta}{|\mathcal{F}|} \\ &= \delta \end{aligned}$$

However, to apply the union bound, we do not need to distribute δ evenly among the candidate functions. We only require:

$$\sum_{f\in\mathcal{F}} \delta_f = \delta$$

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

The main point of this lecture is to examine how PAC bounds of the form $\forall f \in \mathcal{F}\ and\ \forall \delta > 0,\ w.p. \geq 1-\delta$;

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

can be used to select a model that comes close to achieving the best possible performace

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

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

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

as small as possible. First define,

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

where

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

According to the PAC bound, $\forall f \in \mathcal{F}\ and\ \forall \delta > 0,\ w.p. \geq 1-\delta$;

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

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 $\forall f\ in \mathcal{F}$

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

The second term is exactly 0, since $E[\widehat{R}_n(f)] = R(f)$.

Now consider the first term $E[R(\widehat{f}_n^{\delta}) - \widehat{R}_n(f)]$. Let Ω be the set of events on which

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

From our PAC bound, we know that $P(\Omega) \geq 1-\delta$. Thus;

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

We can summarize our analysis with the following theorem.

Theorem 1 ***(Complexity Regularized Model Selection)*** *Let $\mathcal{F}$ be a collection of functions, 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 penalized function*

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

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, ... denotes the collection of histogram classification rules with k equal volume bins. One choice of prefix code for this example is: $k = 1 \Rightarrow code = 0, k = 3 \Rightarrow code = 10, k = 3 \Rightarrow code = 110$ 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*

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

Let $\widehat{f}_n$ be the function 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,

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

From the third homework, we know 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

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

It is a simple exercise to show that the complexity regularized classifier will perform just as well ***automatically****. That is, the proper k is selected automatically, without user intervention.*