ECE901 Spring 2004 Statistical Regularization and Learning Theory **Lecture: 8**

Estimation Error Bounds for Classifiers

Lecturer: Rob Nowak *Scribe: Tee Sivanadyan*

1 Recap: Classifier design

Given a set of training data $\{X_i, Y_i\}_{i=1}^n$ and a finite collection of candidate functions $\mathcal{F}$, select $\hat{f}_n \in \mathcal{F}$ that (hopefully) is a good predictor for future cases. That is

$$\hat{f}_n = \arg\min_{f \in \mathcal{F}} \hat{R}_n(f)$$

where $\hat{R}_n(f)$ is the empirical risk. For any particular $f \in \mathcal{F}$, the corresponding empirical risk is defined as

$$\hat{R}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{f(X_i) \neq Y_i\}}.$$

2 Hoeffding's inequality

Hoeffding's inequality (Chernoff's bound in this case) allows us to gauge how close $\hat{R}_n(f)$ is to the true risk of f, $R(f)$, in probability

$$P(|\hat{R}_n(f) - R(f)| \geq \epsilon) \leq 2e^{-2n\epsilon^2}$$

Since our selection process involves deciding among all $f \in \mathcal{F}$, we would like to gauge how close all the resulting empirical risks are to their expected values. That is, we want to bound

$$P\left(\bigcup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)| \geq \epsilon \right)$$

the probability that $|\hat{R}_n(f) - R(f)| \geq \epsilon$ for any $f \in \mathcal{F}$. We use the *union bound* for this:

$$\begin{aligned} P(\bigcup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)| \geq \epsilon) &\leq \sum_{f \in \mathcal{F}} P(|\hat{R}_n(f) - R(f)| \geq \epsilon) \\ &\leq \sum_{f \in \mathcal{F}} 2e^{-2n\epsilon^2} \\ &= 2|\mathcal{F}|e^{-2n\epsilon^2} \end{aligned}$$

In the proof of Hoeffding's inequality we also obtained a one-sided inequality that implied

$$P(R(f) - \hat{R}_n(f) \geq \epsilon) \leq e^{-2n\epsilon^2}$$

and hence

$$P(\bigcup_{f \in \mathcal{F}} R(f) - \hat{R}_n(f) \geq \epsilon) \leq |\mathcal{F}|e^{-2n\epsilon^2}$$

We can restate the inequality above as follows, *For all $f \in \mathcal{F}$ and for all $\delta > 0$ with probability at least $1 - \delta$*

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

This follows by setting $\delta = |\mathcal{F}|e^{-2n\epsilon^2}$ and solving for ϵ. Thus with a high probability $(1-\delta)$, the true risk for all $f \in \mathcal{F}$ is bounded by the empirical risk of f plus a constant that depends on $\delta > 0$, the number of training samples n, and the size $\mathcal{F}$. Most importantly the bound does not depend on the unknown distribution P_{XY}. Therefore, we can call this a *distribution-free* bound.

3 Error Bounds

We can use the *distribution-free* bound above to obtain a bound on the expected performance of the minimum empirical risk classifier

$$\hat{f}_n = \arg\min_{f \in \mathcal{F}} \hat{R}_n(f)$$

We are interested in bounding

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

the expected risk of $\hat{f}_n$ minus the minimum risk for all $f \in \mathcal{F}$. Note that this difference is always non-negative since $\hat{f}_n$ is at best as good as

$$f^* = \arg\min_{f \in \mathcal{F}} R(f)$$

Recall that $\forall f \in \mathcal{F}$ and $\forall \delta > 0$, with probability at least $1 - \delta$

$$R(f) \leq \hat{R}_n(f) + C(\mathcal{F}, n, \delta)$$

where

$$C(\mathcal{F}, n, \delta) = \sqrt{\frac{\log|\mathcal{F}| + \log(1/\delta)}{2n}}$$

In particular, since this holds for all $f \in \mathcal{F}$ including $\hat{f}_n$,

$$R(\hat{f}_n) \leq \hat{R}_n(\hat{f}_n) + C(\mathcal{F}, n, \delta)$$

and for any other $f \in \mathcal{F}$

$$R(\hat{f}_n) \leq \hat{R}_n(f) + C(\mathcal{F}, n, \delta)$$

since $\hat{R}_n(\hat{f}_n) \leq \hat{R}_n(f)\ \forall f \in \mathcal{F}$. In particular,

$$R(\hat{f}_n) \leq \hat{R}_n(f^*) + C(\mathcal{F}, n, \delta)$$

where $f^* = \arg\min_{f \in \mathcal{F}} R(f)$

Let Ω denote the set of events on which the above inequality holds. Then by definition

$$P(\Omega) \geq 1 - \delta$$

We can now bound $E[R(\hat{f}_n)] - R(f^*)$ as follows

$$\begin{aligned} E[R(\hat{f}_n)] - R(f^*) &= E[R(\hat{f}_n) - \hat{R}_n(f^*) + \hat{R}_n(f^*) - R(f^*)] \\ &= E[R(\hat{f}_n) - \hat{R}_n(f^*)] \end{aligned}$$

since $E[\hat{R}_n(f^*)] = R(f^*)$. The quantity above is bounded as follows.

$$\begin{aligned} E[R(\hat{f}_n) - \hat{R}_n(f^*)] &= E[R(\hat{f}_n) - \hat{R}_n(f^*)|\Omega]\, P(\Omega) + E[R(\hat{f}_n) - \hat{R}_n(f^*)|\Omega]\, P(\bar{\Omega}) \\ &\leq E[R(\hat{f}_n) - \hat{R}_n(f^*)|\Omega] + \delta \end{aligned}$$

since $P(\Omega) \leq 1$, $1 - P(\Omega) \leq \delta$ and $R(\hat{f}_n) - \hat{R}_n(f^*) \leq 1$

$$\begin{aligned} E[R(\hat{f}_n) - \hat{R}_n(f^*)|\Omega] &\leq E[R(\hat{f}_n) - \hat{R}_n(\hat{f}_n)|\Omega] \\ &\leq C(\mathcal{F}, n, \delta) \end{aligned}$$

Thus

$$E[R(\hat{f}_n) - \hat{R}_n(f^*)] \leq C(\mathcal{F}, n, \delta) + \delta$$

So we have

$$E[R(\hat{f}_n)] - \min_{f \in \mathcal{F}} R(f) \leq \sqrt{\frac{\log |\mathcal{F}| + \log(1/\delta)}{2n}} + \delta, \ \forall \delta > 0$$

In particular, for $\delta = \sqrt{1/n}$, we have

$$\begin{aligned} E[R(\hat{f}_n)] - \min_{f \in \mathcal{F}} R(f) &\leq \sqrt{\frac{\log |\mathcal{F}| + \log n}{2n}} + \frac{1}{\sqrt{n}} \\ &\leq \sqrt{\frac{\log |\mathcal{F}| + \log n + 2}{n}}, \ \text{ since } \ \sqrt{x} + \sqrt{y} \leq \sqrt{2}\sqrt{x+y}, \ \forall \ x, y > 0 \end{aligned}$$

4 Application: Histogram Classifier

Let $\mathcal{F}$ be the collection of all classifiers with M equal volume bins. Then $|\mathcal{F}| = 2^M$, and the histogram classification rule

$$\hat{f}_n = \arg\min_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^{n} 1_{\{f(X_i) \neq Y_i\}} \right)$$

satisfies

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

which suggests the choice $M = \log_2 n$ (balancing $M \log 2$ with $\log n$), resulting in

$$E[R(\hat{f}_n)] - \min_{f \in \mathcal{F}} R(f) = O\left(\sqrt{\frac{\log n}{n}} \right)$$