

# ELEN6887

## Lecture 8: General Bounded Losses - Error Bounds

R. Castro

2/15/2010

### 1 Recap: Hoeffding's inequality

In the last lecture we introduced Hoeffding's inequality, that quantifies how the sample average of a number of random variable converges to the mean. This was motivated by our goal of constructing PAC bounds, relating the empirical risk  $\hat{R}_n(f)$  and the true risk  $R(f)$ .

### 2 Bounded Loss Functions

In this lecture we will consider loss functions that are bounded. Without loss of generality let  $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow [0, 1]$ . The zero-one loss is a particularly useful example of such a loss function (e.g.  $\ell(y_1, y_2) = \mathbf{1}\{y_1 \neq y_2\}$ ).

Given a set of training data  $\{X_i, Y_i\}_{i=1}^n$  and a finite collection of candidate functions  $\mathcal{F}$ , we can 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

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

where  $Z_i = \ell(f(X_i), Y_i)$ . Note that  $E[\hat{R}_n(f)] = R(f)$ . We can now apply Hoeffding's inequality to the collection  $Z_i$ 's. Let  $\epsilon > 0$ , then

$$\begin{aligned} P(|\hat{R}_n(f) - R(f)| \geq \epsilon) &= P\left(\left|\frac{1}{n} \sum_{i=1}^n Z_i - E[Z_i]\right| \geq \epsilon\right) \\ &= P\left(\left|\sum_{i=1}^n Z_i - E[Z_i]\right| \geq n\epsilon\right) \\ &\leq 2e^{-\frac{2(n\epsilon)^2}{n}} = 2e^{-2n\epsilon^2} . \end{aligned}$$

Note that this bound applies to a *single* model  $f \in \mathcal{F}$ .

Since our selection process involves deciding among all  $f \in \mathcal{F}$ , we would like to gauge how close the empirical risks are to their expected values. We can do this by studying the probability that one or more of the empirical risks deviates significantly from its expected value. This is captured by the probability

$$P\left(\exists f \in \mathcal{F} : |\hat{R}_n(f) - R(f)| \geq \epsilon\right) = P\left(\max_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)| \geq \epsilon\right) .$$

Note that the event

$$\max_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)| \geq \epsilon$$

is equivalent to union of the events

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

Therefore, we can use Bonferroni's bound (aka the "union of events" or "union" bound) to obtain

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

where  $|\mathcal{F}|$  is the number of classifiers in  $\mathcal{F}$ . 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

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

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  $\epsilon$ . Thus with a high probability (greater than  $1 - \delta$ ), 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 for Empirical Risk Minimization

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

$$\tilde{f} \equiv \arg \min_{f \in \mathcal{F}} R(f)$$

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,

$$R(\hat{f}_n) \leq \hat{R}_n(\tilde{f}) + C(\mathcal{F}, n, \delta), \quad (1)$$

where  $\tilde{f} = \arg \min_{f \in \mathcal{F}} R(f)$ .

Let  $\Omega$  denote the event that the inequality (1) holds. Our application of the Hoeffding's inequality characterizes that probability, that is  $P(\Omega) \geq 1 - \delta$ . We can now bound  $E[R(\hat{f}_n)] - R(\tilde{f})$  as follows

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

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

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

since  $P(\Omega) \leq 1$ ,  $1 - P(\Omega) \leq \delta$  and  $R(\hat{f}_n) - \hat{R}_n(\tilde{f}) \leq 1$ . Now

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

and so we have

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

In other words

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

Remember that  $\delta > 0$  is arbitrary, therefore we should choose  $\delta$  so that the bound is as tight as it can. Although that is possible it gives rise to complicated expressions. Noticing that the first term has to decay slower than  $1/\sqrt{n}$  a reasonable choice is  $\delta = \sqrt{1/n}$ , yielding

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

## 4 Application: Histogram Classifier

Let  $\mathcal{F}$  be the collection of all classifiers with  $M$  equal volume cells. 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

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

Without any further assumptions on the distribution  $P_{XY}$  this suggests the choice  $M = \log_2 n$  (balancing  $M \log 2$  with  $\log n$ ), resulting in

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

This choice of  $M$  guarantees we have  $O(n/\log(n))$  samples per bin, which is almost all the samples. This is clearly very conservative most of the times. If we had more assumptions on the distribution  $P_{XY}$  we could take this into account to assess the approximation error  $R(\tilde{f}) - R^*$ , and balance this with the estimation error provided by the analysis above.