COS 511: Foundations of Machine Learning

Rob Schapire — Lecture #7
Scribe: Forrest Collman — February 25, 2003

1 Bounding Error on Mean

Given a set of examples $x_i = [0,1]$ $i = \{1...m\}$ drawn from a distribution D we would like to calculate the observed mean $\hat{p}$

$$\hat{p} = \frac{1}{m}\sum_{i=1}^{m} x_i \tag{1}$$

and compare that to the real mean of D, call it p. It is useful to define the quantity

$$q \equiv p + \epsilon. \tag{2}$$

We could weakly bound the probability of $\hat{p}$ being greater than q using Markov's inequality:

$$\Pr[x \geq kEX] \leq 1/k \tag{3}$$
$$\Pr[\hat{p} \geq s] \leq p/s \tag{4}$$
$$\Pr[\hat{p} \geq q] \leq p/q < 1. \tag{5}$$

However, we can do better than that, in fact
THEOREM

$$\Pr[\hat{p} \geq q] \leq \exp(-\mathrm{RE}(q||p)m) \tag{6}$$

where

$$\mathrm{RE}(q||p) = p\,\ln\left(\frac{p}{q}\right) + (1-p)\,\ln\left(\frac{1-p}{1-q}\right) \tag{7}$$

PROOF

$$\Pr[\hat{p} \geq q] = \Pr[e^{\lambda\hat{p}m} \geq e^{\lambda qm}] \leq e^{-\lambda qm} E\left[e^{\lambda\hat{p}m}\right] \tag{8}$$
$$= e^{-\lambda qm} E\left[e^{\lambda\sum x_i}\right] = e^{-\lambda qm} E\left[\prod_{i=1}^{m} e^{\lambda x_i}\right] \tag{9}$$

now because each x_i is an independant measurement

$$e^{-\lambda qm}\prod_{i=1}^{m}\mathrm{E}[e^{\lambda x_i}]. \tag{10}$$

Now note that

$$e^{\lambda x} \leq 1 - x + e^{\lambda}x \;\forall x \in [0,1] \tag{11}$$

which we can use to show that

$$e^{-\lambda qm}\prod_{i=1}^{m}\mathrm{E}[e^{\lambda x_i}] \leq e^{-\lambda qm}\prod_{i=1}^{m}\mathrm{E}[1 - x_i + e^{\lambda}x_i] \tag{12}$$
$$= e^{-\lambda qm}\prod_{i=1}^{m}\left(1 - p + e^{\lambda}p\right) = e^{-\lambda qm}\left(1 - p + e^{\lambda}p\right)^m \tag{13}$$
$$= \left(e^{-\lambda q}\left(1 - p + e^{\lambda}p\right)\right)^m. \tag{14}$$

Now if we minimize this probability with respect to lambda you get

$$\lambda_{min} = \ln\left(\frac{q(1-p)}{(1-q)p}\right). \tag{15}$$

Plugging that back into the probability you get the desired bound

$$\Pr[\hat{p} \geq q] \leq \exp(-\mathrm{RE}(q||p)m) \tag{16}$$

Note that by defining $y_i = 1 - x_i$ we can prove the symmetric result that

$$\Pr[\hat{p} \geq p - \epsilon] \leq \exp(-\mathrm{RE}(p-\epsilon||p)m) \tag{17}$$

This theorom can also be used to prove corollaries

$$\Pr[\hat{p} \geq p + \alpha p] \leq e^{-mp\alpha^2/3} \tag{18}$$

$$\Pr[\hat{p} \leq p - \alpha p] \leq e^{-mp\alpha^2/2} \tag{19}$$

This is done by plugging in $\epsilon = \alpha p$ and then bounding $\mathrm{RE}(q||p)$.

Thinking back to the double-sample proof: $m(h)$ was the number of mistakes on S'. Thus $\hat{p} = m(h)/m$ and $p = \epsilon$ and so

$$\Pr[m(h) < m\epsilon/2] = \Pr[\hat{p} < p - p/2] \leq e^{-mp/8}. \tag{20}$$

If $m \geq 8/\epsilon$ then

$$\Pr[m(h) < m\epsilon/2] \leq e^{-1} < 1/2. \tag{21}$$

2 McDiarmid's Inequality

Let

$$f(x_1, \ldots, x_m) \tag{22}$$

be any function such that for all $x_1, \ldots, x_m; x_i'$,

$$\left|f(x_1, \ldots, x_i, \ldots, x_m) - f(x_1, \ldots, x_i', \ldots, x_m)\right| \leq c_i. \tag{23}$$

In other words, changing x_i can never change f by more than c_i. Let $X_1 ... X_m$ be independant but not necessarily identically distributed.

THEOREM

$$\Pr\left[f(x_1 ... x_m) \geq \mathrm{E}[f(x_1 ... x_m)] + \epsilon\right] \leq \exp\left(\frac{-2\epsilon^2}{\sum c_i^2}\right) \tag{24}$$

3 Hoeffding

Hoeffding's inequality

$$\Pr[\hat{p} \geq p + \epsilon] \leq e^{-2\epsilon^2 m} \tag{25}$$

is a special case of McDiarmid. Let

$$f(x_1 ... x_m) = \frac{1}{m}\sum_{i=1}^{m} x_i. \tag{26}$$

$\mathrm{E}[f] = p$, $c_i = 1/m$ (for $0 \leq x \leq 1$). So making use of McDiarmid's Inequality,

$$\Pr[\hat{p} \geq p + \epsilon] \leq \exp\left(\frac{-2\epsilon^2}{\sum 1/m^2}\right) = \exp(-2\epsilon^2 m) \tag{27}$$

4 General Strategy

Let

$$err(h) = Pr_{x,y \sim D}[y \neq h(x)] \tag{28}$$

$$\hat{err}(h) = \frac{1}{m}|\{i : y_i \neq h(x_i)\}| \tag{29}$$

Minimize $\hat{err}(h)$. Show close to true error, thus effectively minimizing the true error.

5 Finite Example

THEOREM Assume $|H|$ finite. Given m random examples, with probability $1 - \delta$

$$\forall\, h \epsilon H \;:\; |err(h) - \hat{err}(h)| < \epsilon \tag{30}$$

$$\text{if } m \geq O\left(\frac{\ln|H| + \ln(1/\delta)}{\epsilon^2}\right) \tag{31}$$

PROOF
Fix an h.

$$X_i = \{1 \text{ if } h(x_i) \neq y_i, 0 \text{ else}\}. \tag{32}$$

$$\mathrm{E}[X_i] = err(h) \tag{33}$$

$$\frac{1}{m}\sum_{i=1}^{m} X_i = \hat{err}(h) \tag{34}$$

$$\Pr[\hat{p} \geq p + \epsilon] \leq e^{-2\epsilon^2 m} \tag{35}$$

$$\Pr[\hat{err}(h) \geq err(h) + \epsilon] \leq e^{-2\epsilon^2 m} \tag{36}$$

$$\Pr[|\hat{err}(h) - err(h)| \geq \epsilon] \leq 2e^{-2\epsilon^2 m} \tag{37}$$

Now using union bound we can bound the probability for all h.

$$\Pr[\exists h \epsilon H : |err(h) - \hat{err}(h)| \geq \epsilon] \leq 2|H|e^{-2\epsilon^2 m} \tag{38}$$

So by increasing m we can bound the probability. So in order to make the probability less than δ we need

$$m \geq \frac{\ln(2|H|) + \ln(1/\delta)}{\epsilon^2} \tag{39}$$

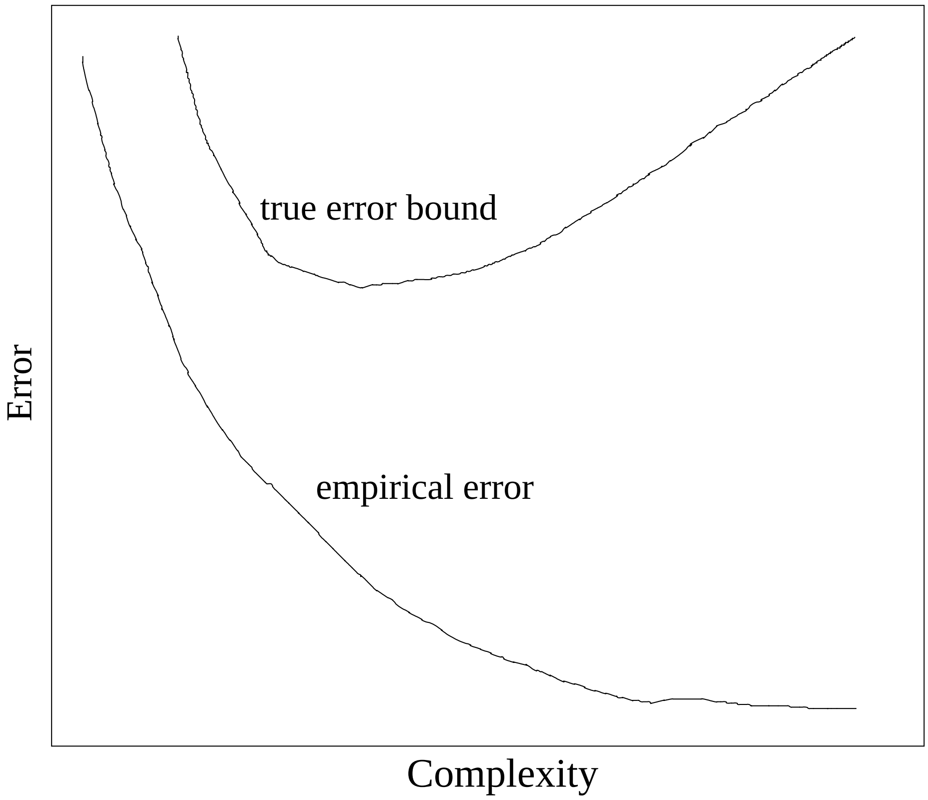
Note that this bound is weaker than before, it requires more examples for smaller errors (grows like $1/\epsilon^2$ vs $1/\epsilon$).

6 Overfitting

$$err(h) \leq \hat{err}(h) + O\left(\sqrt{\frac{\ln(|H|) + \ln(1/\delta)}{m}}\right) \tag{40}$$

$$err(h) \leq \hat{err}(h) + O\left(\sqrt{\frac{|h| + \ln(1/\delta)}{m}}\right) \tag{41}$$

As h gets more complex, the training error $\hat{err}(h)$ tends to go down while the complexity $|h|$ increases, so the true error $err(h)$ may at first go down but then increase, as in the figure. This is called overfitting.



Solutions: (1) Cross Validation, which means, hold out some of the training data and use it to determine when to stop training. (2) Treat bound as real and optimize in $|h|$. This is called structural risk minimization. (3) Build algorithms that are resistant to overfitting.