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 \( \hat{p} \)

\[
\hat{p} = \frac{1}{m} \sum_{i=1}^{m} x_i
\]  

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

\[
q = p + \epsilon.
\]  

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

\[
\Pr[\hat{p} \geq \epsilon] \leq 1/k
\]  

\[
\Pr[\hat{p} \geq \epsilon] \leq \frac{p}{\epsilon}
\]  

\[
\Pr[\hat{p} \geq q] \leq \frac{p}{q} < 1.
\]  

However, we can do better than that, in fact

**THEOREM**

\[
\Pr[\hat{p} \geq q] \leq \exp(-\text{RE}(q|p)m)
\]  

where

\[
\text{RE}(q|p) = p \ln \left( \frac{q}{p} \right) + (1 - p) \ln \left( \frac{1 - p}{1 - q} \right)
\]  

**PROOF**

\[
\Pr[\hat{p} \geq q] = \Pr[e^{\lambda \hat{p}m} \geq e^{\lambda qm}] \leq e^{-\lambda qm} E[e^{\lambda \hat{p}m}]
\]  

\[
= e^{-\lambda qm} E[e^{\lambda \sum x_i}] = e^{-\lambda qm} E[\prod_{i=1}^{m} e^{\lambda x_i}]
\]  

now because each \( x_i \) is an independant measurement

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

Now note that

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

which we can use to show that

\[
e^{-\lambda qm} \prod_{i=1}^{m} E[e^{\lambda x_i}] \leq e^{-\lambda qm} \prod_{i=1}^{m} E[1 - x_i + e^\lambda x_i]
\]  

\[
= e^{-\lambda qm} \prod_{i=1}^{m} \left( 1 - p + e^\lambda p \right)^m = e^{-\lambda qm} \left( 1 - p + e^\lambda p \right)^m
\]  

\[
= \left( e^{-\lambda q} \left( 1 - p + e^\lambda p \right) \right)^m.
\]
Now if we minimize this probability with respect to lambda you get
\[ \lambda_{\text{min}} = \ln \left( \frac{q(1-p)}{(1-q)p} \right). \] (15)

Plugging that back into the probability you get the desired bound
\[ \Pr[\hat{\mu} \geq q] \leq \exp(-\text{RE}(q||p)m) \] (16)

Note that by defining \( y_i = 1 - x_i \) we can prove the symmetric result that
\[ \Pr[\hat{\mu} \geq p - \epsilon] \leq \exp(-\text{RE}(p - \epsilon||p)m) \] (17)

This theorem can also be used to prove corollaries
\[ \Pr[\hat{\mu} \geq p + \alpha p] \leq \exp\left(-\frac{2\epsilon^2}{m}\right) \] (18)
\[ \Pr[\hat{\mu} \leq p - \alpha p] \leq \exp\left(-\frac{2\epsilon^2}{2m}\right) \] (19)

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

Thinking back to the double-sample proof: \( m(h) \) was the number of mistakes on \( S' \). Thus \( \hat{\mu} = m(h)/m \) and \( p = \epsilon \) and so
\[ \Pr[m(h) < m\epsilon/2] = \Pr[\hat{\mu} < p - p/2] \leq \exp(-mp/8). \] (20)

If \( m \geq 8/\epsilon \) then
\[ \Pr[m(h) < m\epsilon/2] \leq \exp(-1) < 1/2. \] (21)

## 2 McDiarmid’s Inequality

Let
\[ f(x_1, \ldots, x_m) \] (22)
be any function such that for all \( x_1, \ldots, x_m, x_i' \),
\[ |f(x_1, \ldots, x_i, \ldots, x_m) - f(x_1, \ldots, x_i', \ldots, x_m)| \leq c_i. \] (23)

In other words, changing \( x_i \) can never change \( f \) by more than \( c_i \). Let \( X_1 \ldots X_m \) be independent but not necessarily identically distributed.

**THEOREM**
\[ \Pr[f(x_1 \ldots x_m) \geq E[f(x_1 \ldots x_m)] + \epsilon] \leq \exp\left(-\frac{2\epsilon^2}{\sum c_i^2}\right) \] (24)

## 3 Hoeffding

Hoeffding’s inequality
\[ \Pr[\hat{\mu} \geq p + \epsilon] \leq \exp(-2\epsilon^2 m) \] (25)
is a special case of McDiarmid. Let
\[ f(x_1 \ldots x_m) = \frac{1}{m} \sum_{i=1}^{m} x_i. \] (26)

\( E[f] = p, c_i = 1/m \) (for \( 0 \leq x \leq 1 \)). So making use of McDiarmid’s Inequality,
\[ \Pr[\hat{\mu} \geq p + \epsilon] \leq \exp\left(-\frac{2\epsilon^2}{\sum 1/m^2}\right) = \exp(-2\epsilon^2 m) \] (27)
4 General Strategy

Let

$$err(h) = Pr_{x,y \sim D}[y \neq h(x)]$$ \hspace{1cm} (28)

$$\hat{err}(h) = \frac{1}{m}|\{i : y_i \neq h(x_i)\}|$$ \hspace{1cm} (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 $1 - \delta$

$$\forall h \in H : |err(h) - \hat{err}(h)| < \epsilon$$ \hspace{1cm} (30)

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

**PROOF**

Fix an $h$.

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

$$E[X_i] = err(h)$$ \hspace{1cm} (33)

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

$$\Pr[\hat{p} \geq p + \epsilon] \leq e^{-2\epsilon^2}$$ \hspace{1cm} (35)

$$\Pr[\hat{err}(h) \geq err(h) + \epsilon] \leq e^{-2\epsilon^2}$$ \hspace{1cm} (36)

$$\Pr[|err(h) - \hat{err}(h)| \geq \epsilon] \leq 2e^{-2\epsilon^2}$$ \hspace{1cm} (37)

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

$$\Pr[\exists h \in H : |err(h) - \hat{err}(h)| \geq \epsilon] \leq 2|H|e^{-2\epsilon^2}$$ \hspace{1cm} (38)

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

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

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

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