1. Progressive Validation Technique

Progressive validation is a technique which allows you to use almost $\frac{1}{2}$ of the data in a holdout set for training purposes while still providing the same guarantee as the holdout bound. It first appeared in [**?**] and is discussed in a more refined and detailed form here.

Suppose that you have a training set of size m_{train} and test set of size m_{pv}. Progressive validation starts by first learning a hypothesis on the training set and then testing on the first example of the test set. Then, we train on training set plus the first example of the test set and test on the second example of the test set. The process continues m_{test} iterations. Let m abbreviate m_{pv}. Then, we have m hypotheses, $h_1, ..., h_m$ and m error observations, $\hat{e}_1, ..., \hat{e}_m$. The hypothesis output by progressive validation is the randomized hypothesis which chooses uniformly from $h_1, ..., h_m$ and evaluates to get an estimated output.

Since we are randomizing over hypotheses trained on m_{train} to $m_{\text{train}} + m_{\text{pv}} - 1$ examples, the expected number of examples used by any hypothesis is $m_{\text{train}} + \frac{m_{\text{pv}}-1}{2}$. Given that training can exhibit phase transitions, the extra few examples can greatly improve the accuracy of the trained example.

The true error rate of this randomized hypothesis will be:

$$e_{\text{pv}} = \frac{1}{m_{\text{pv}}} \sum_{i=1}^{m_{\text{pv}}} e(h_i)$$

where $e(h_i) = \Pr_D(h_i(x) \neq y)$ and the empirical error estimate of this randomized hypothesis will be:

$$\hat{e}_{\text{pv}} = \frac{1}{m_{\text{pv}}} \sum_{i=1}^{m_{\text{pv}}} \hat{e}_i$$

2. Variance Analysis

Bounding the deviation of $\hat{e}_{\text{pv}}$ is more difficult than bounding the deviation of a holdout error. To understand this, we can think of two games, the holdout game and the progressive validation game.

In the holdout game, your opponent chooses a bias and then nature flips m coins with that bias. If the deviation of the average number of heads is larger than ϵ, then you lose. Otherwise, you win.

In the progressive validation game, the opponent chooses the bias of *each* coin just before it is flipped. The goal of the opponent remains the same, and the opponent wins if a large deviation is observed.

The progressive validation opponent is stronger than the holdout opponent since the progressive validation opponent could choose the same bias for every coin. Nonetheless, we will see that the progressive validation opponent is *not* much stronger.

There are two ways in which we can show that the progressive validation opponent is not much stronger. The first technique will show that the variance of the progressive validation estimate is smaller than might be expected.

Theorem 2.1. *Suppose we test the progressive validation hypothesis on $m_{test} = m_{pv}$ additional examples. Let $\hat{e}_{test}$ be the empirical error on these examples. Then, we have:*

$$E_{(x,y)^{m_{pv}+m_{test}} \sim D^{m_{pv}+m_{test}}} (\hat{e}_{test} - e_{pv})^2$$

$$\geq E_{(x,y)^{m_{pv}} \sim D^{m_{pv}}} (\hat{e}_{pv} - e_{pv})^2$$

Proof. Every example on the left hand side can be though of as a coin with bias e_{pv}. The variance of the LHS is then $m * e_{\mathrm{pv}}(1 - e_{\mathrm{pv}})$. The right hand side is:

$$E_{(x,y)^m \sim D^m} (\hat{e}_{\mathrm{pv}} - e_{\mathrm{pv}})^2$$

$$= E_{(x,y)^m \sim D^m} \left[\sum_{i=1}^{m} \hat{e}_i - e_i \right]^2$$

where $e_i = e(h_i)$

$$= E_{(x,y)^m \sim D^m} \left[\sum_{i \neq j} (\hat{e}_i - e_i)(\hat{e}_j - e_j) + \sum_{i=1}^{m} (\hat{e}_i - e_i)^2 \right]$$

$$= \sum_{i \neq j} E_{(x,y)^m \sim D^m} (\hat{e}_i - e_i)(\hat{e}_j - e_j) + \sum_{i=1}^{m} E_{(x,y)^m \sim D^m} (\hat{e}_i - e_i)^2$$

The cross product term is:

$$E_{(x,y)^m \sim D^m} (\hat{e}_i - e_i)(\hat{e}_j - e_j)$$

Without loss of generality, assume that $i < j$. What we wish to prove is that the expected value of this quantity is 0 conditional on the values of all random variables other than ith or jth example. Let S_{ij} be the set of examples minus the ith and jth example. Also, let z_i and z_j be the ith and jth labeled examples. If we can show that:

$$\forall S_{ij} : \; E_{z_i, z_j | S_{ij}} (\hat{e}_i - e_i)(\hat{e}_j - e_j) = 0$$

then linearity of expectation will imply that:

$$E_{(x,y)^m \sim D^m} (\hat{e}_i - e_i)(\hat{e}_j - e_j) = 0$$

The value of e_i is fixed after conditioning on S_{ij} (and assuming a deterministic learning algorithm) while the value of e_j is not fixed: it is dependent on the random variable z_i. Let $e_j(z_i)$ be the derived random variable. Then, the expectation is:

$$E_{z_i, z_j | S_{ij}} (\hat{e}_i - e_i)(\hat{e}_j - e_j)$$

(implicitly conditioning on S_{ij})

$$= \sum_{z_i, z_j} \Pr_D(z_i) \Pr_D(z_j) (\hat{e}_i(z_i) - e_i)(\hat{e}_j(z_j, z_i) - e_j(z_i))$$

$$= \sum_{z_i} \Pr_D(z_i)(\hat{e}_i(z_i) - e_i) \sum_{z_j} \Pr_D(z_j)(\hat{e}_j(z_j, z_i) - e_j(z_i))$$

For any fixed z_i, we want to show that $\sum_{z_j} \Pr_D(z_j)(\hat{e}_j(z_j, z_i) - e_j(z_i)) = 0$.With z_i fixed, the true error rate of the jth hypothesis is $e_j(z_i)$. Therefore, the probability of observing an error ($\hat{e}_j(z_j, z_i) = 1$) is $e_j(z_i)$, and the probability of observing no error ($\hat{e}_j(z_j, z_i) = 0$) is $1 - e_j(z_i)$. This implies:

$$\sum_{z_j} \Pr_D(z_j)(\hat{e}_j(z_j, z_i) - e_j(z_i))$$

$$= \sum_{z_j : \hat{e}_j(z_j, z_i) = 0} \Pr_D(z_j)(-e_j(z_i)) + \sum_{z_j : \hat{e}_j(z_j, z_i) = 1} \Pr_D(z_j)(1 - e_j(z_i))$$

$$= (1 - e_j(z_i))(-e_j(z_i)) + e_j(z_i)(1 - e_j(z_i))$$

$$= 0$$

Putting this together, we get:

$$\text{cross term} = \sum_{z_i} \Pr_D(z_i)(\hat{e}_i(z_i) - e_i) * 0$$

$$= 0$$

Since this sum is zero regardless of the values of all other random variables, the cross product terms all have expectation zero. Note that we can extend this proof to randomized algorithms by conditioning on the random bits of the algorithm in the above argument. Consequently:

$$= \sum_{i=1}^{m} E_{(x,y)^m \sim D^m} \left[(\hat{e}_i - e_i)^2\right]$$

$$= \sum_{i=1}^{m} e_i(1 - e_i)$$

So, all that we must show is:

$$m e_{\text{pv}}(1 - e_{\text{pv}}) \geq \sum_{i=1}^{m} e_i(1 - e_i)$$

which follows from Jensen's inequality and the convexity of $x(1-x)$ on the interval $[0, 1]$. □

3. Deviation Analysis

The second way in which we can show that the progressive validation opponent is not so strong is by bounding the deviation directly. Surprisingly, we can prove exactly the same bound as for Hoeffding inequality.

Theorem 3.1.

$$\Pr_{(x,y)^m \sim D^m} (e_{pv} \geq \hat{e}_{pv} + \epsilon) \leq e^{-2m\epsilon^2}$$

Proof. (A variant of Chernoff)

$$\Pr_{(x,y)^m \sim D^m} (e_{\text{pv}} \geq \hat{e}_{\text{pv}} + \epsilon)$$

$$\Leftrightarrow \forall \lambda \quad \Pr_{(x,y)^m \sim D^m} (e^{\lambda m e_{\text{pv}}} \geq e^{\lambda m (\hat{e}_{\text{pv}} + \epsilon)})$$

$$\Leftrightarrow \Pr_{(x,y)^m \sim D^m} (e^{\lambda m (e_{\text{pv}} - \hat{e}_{\text{pv}} - \epsilon)} \geq 1)$$

$$\leq E_{(x,y)^m \sim D^m} e^{\lambda m (e_{\text{pv}} - \hat{e}_{\text{pv}} - \epsilon)}$$

$$= E_{(x,y)^m \sim D^m} e^{\lambda (\sum_{i=1}^m e_i - \hat{e}_i - \epsilon)}$$

$$= \prod_{i=1}^{m} E_{(x,y) \sim D} \left[e^{\lambda (e_i - \hat{e}_i - \epsilon)} | \hat{e}_1, ..., \hat{e}_{i-1}\right]$$

The value of $e^{\lambda(e_i - \hat{e}_i - \epsilon)}$ is only dependent on $\hat{e}_1, ..., \hat{e}_{i-1}$ through the value of e_i. For all possible hypotheses, h_i, we know that $\hat{e}_i$ is a Bernoulli random variable with bias $e_i = \Pr_D(h_i(x) \neq y)$. This is true regardless of how we choose h_i. Consequently, we have:

$$= \prod_{i=1}^{m} E_{(x,y)\sim D}\left[e^{\lambda(e_i - \hat{e}_i - \epsilon)}\right] \tag{3.1}$$

Since this is true for all λ, we can choose the optimal λ. In general, this is an optimization problem which can be worst-case simplified with the inequality:

$$E_{(x,y)\sim D} e^{\lambda(e_i - \hat{e}_i)} \leq e^{\frac{\lambda^2}{8}}$$

which implies:

$$\forall\lambda \ \leq \prod_{i=1}^{m} e^{\frac{\lambda^2}{8} - \lambda\epsilon)}$$

This is optimized for $\frac{2\lambda}{8} = \epsilon$ which implies $\lambda = 4\epsilon$ giving:

$$\leq \prod_{i=1}^{m} e^{-2\epsilon^2}$$

$$= e^{-2m\epsilon^2}$$

$\square$

Open Problem: Starting with 3.1, derive a bound similar to the relative entropy Chernoff bound which holds for progressive validation.