Cross-Validation

compiled by Alvin Wan from Professor Benjamin Recht's lecture

1 Generalization

We note that *not* generalizing means *not doing well* on the new data. Generalizing means that we do *the same* on new data. We want to achieve the following, in any machine learning problem.

$$\text{minimize}_w R[w] = E_{(x,y)}[\text{loss}(w;(x,y))]$$

Let us consider a model for **empirical risk**, which is the error $R_T[w] = \frac{1}{n_T}\sum_{i=1}^{n_T} \text{loss}(w,(x_i,y_i))$ on our training set $(x_1,y_1),(x_2,y_2),\cdots(x_{n_T},y_{n_T})$.

$$R[w] = R[w] - R_T[w] + R_T[w]$$

We can see that this statement is trivially true. The term $R[w]-R_T[w]$ is our **generalization error**.

Theorem Empirical risk is an unbiased estimate of risk.

If (x_i,y_i) are sampled i.i.d. and w is *fixed* (independent of (x_i,y_i)). Compute the expectation of $R_T[w]$.

$$\begin{aligned}\mathbb{E}[R_T[w]] &= \frac{1}{n_T}\sum_{i=1}^{n}\mathbb{E}[\text{loss}(w,(x_i,y_i))] \\ &= \frac{1}{n_T}\sum_{i=1}^{n_T} R[w] \\ &= R[w]\end{aligned}$$

*Note: This assumes that the loss is bounded and positive. In other words, for some B, $0 \le \text{loss}(w,(x,y)) \le B$.

We now consider the following. $var(X - \mathbb{E}[X]) = \text{var}(X)$, where $X = R_T[w]$.

$$\begin{aligned}\text{var}(R_T[w] - R[w]) &= \frac{1}{n_T}\text{var}(\text{loss}(w,(x,y)) - R[w]) \\ &\leq \frac{B^2}{n_T}\end{aligned}$$

So, the more data you get (the greater n_T is), the less variance you have in your estimate for total risk.

2 Hoeffding's Inequality

Definition: Hoeffding's Inequality Let $z_1, z_2 ... z_n$ be i.i.d. random variables. Assume the mean is μ, $\mathbb{E}[z_i] = \mu$ and that all probabilities are bounded by some B, $\Pr(0 \leq z_i \leq B) = 1$. Then, we find that the estimate for our mean gets exponentially better with an increase in the amount of training data: $\Pr(\frac{1}{n}\sum_{i=1}^{n} z_i - \mu \geq t) \leq \exp(-\frac{2nt^2}{B^2})$.

We refer to this as a quantitative prediction, because it provides us with probability estimates. Applying Hoeffding's to the above, we find that the actual risk is bounded within a certain distance of our empirical risk.

$$R[w] \leq R_T[w] + t \text{ with probability } \exp(-\frac{2nt^2}{B^2})$$

Plug in $t = \frac{B}{\sqrt{n}}$, and we get the following.

$$R[w] \leq R_T[w] + \frac{B}{\sqrt{n}} \text{ with probability at least } 86\%$$

Theorem Given we take the maximum of k estimates for w, our empirical risk grows exponentially closer to the actual risk: $R[\widehat{w}] \leq R[\widehat{w}] + B\sqrt{\frac{\log k}{n_T}}$ with probability at least $1 - \frac{1}{k}$. More formally, we have that:

Suppose we have $w_1, w_2 ... w_k$.

$$\hat{w} = \text{argmin}_{1 \leq i \leq k} R_T[w_i]$$

We now consider the new risk.

$R[\hat{w}] \leq R_T[w] + B\sqrt{\frac{\log k}{n_T}}$ with probability at least $1 - \frac{1}{k}$. We can prove this using the union bound. Note that the union bound states the following.

$$\Pr(\cup_{i=1}^{k} \mathbb{E}_i) \leq \sum_{i=1}^{k} \Pr(\mathbb{E}_i)$$

Let us plug into the Hoeffding inequality, where $R_T[w_i] = \frac{1}{n_T} \sum_{i=1}^{n} z_i$.

$$\Pr(R[w_i] - R_T[w_i] \geq t) \leq \exp(-\frac{2nt^2}{B^2})$$
$$\Pr(\exists i \text{ s.t.} R[w_i] - R_T[w_i] \geq t) \leq \exp(-\frac{2nt^2}{B^2})$$

If you look at the test set and adapt, the term $\sqrt{\frac{\log k}{n_V}}$ actually approaches $\sqrt{\frac{k}{n_V}}$.

Theorem If you have d parameters (i.e., $x \in \mathbb{R}^d, w \in \mathbb{R}^d$), then $w_T = \text{argmin}_{\|w\| \leq M} R_T[w]$.

If you choose to minimize $R[w] + \lambda \|w\|^2$, we find that our w is bounded, $\|w_T\| \leq \frac{1}{\lambda} R_T[o]$, where $R_T[o] = R_T[w_T] + \lambda \|w_T\|^2 \geq \lambda \|w_T\|^2$. Let us consider an example.

Bound the norm of w by M, $w = \frac{1}{\sqrt{m}}(1, 1, \cdots 1)$. This gives us 2^d possible vectors. Using Hoeffding's, we get that $R[w_T] \leq R_T[w_T] + CM\sqrt{\frac{d}{n}}$. Intuitively, we can prove this by considering all points where norm is less than M. We know that every point within ϵ of a certain point must have similar loss. We can then extend this to a ball of radius M.

If you have more points (n) than parameters (d), your risk is bounded. If otherwise, where the number of parameters is much larger than the number of points, then the bound provided above does not give us much information. Instead, we now need a new technique and a new bound: cross-validation.

3 Cross Validation

We take a validation set $(x_1, y_1 \cdots (x_{n_V}, y_{n_V})$. We then take our k parameter settings, and using the training set, generate $w_1, w_2 ... w_k$. Then $R[\hat{w}] \leq R_V[\hat{w}] + B\sqrt{\frac{\log k}{n_V}}$. Validation error, or test error, then tells us which w_i gives the smallest risk.

In practice, we are not provided with a validation set. So, we take our data $(x_1, y_1) ... (x_n, y_n)$ and partition randomly it into the new training set $T = \{(x_i, y_i)\}_{i=1}^{n_T}$ and the validation set $V = \{(x_i, y_i)\}_{i=1}^{n_V}$, such that we don't lose data, $n = n_T + n_V$ and the test training set is larger than the validation set $n_T > n_V$.

It is highly recommended that you repeat this process.

Note that the training set must be large in order for the bound $\sqrt{d}n_T$ to be meaningful. Additionally, we need a large enough test set so that $B\sqrt{\frac{\log k}{n_V}}$ is likewise meaningful. This is why we split into validation and training sets.

Here are the takeaways from both this section and all of this course.

1. More points than parameters.
2. Don't adapt to the test data.
3. Overfitting can happen in many different ways – for one, how we cross-validate.