Notes on Statistical Learning Theory

Ruixin Guo

February 16, 2024

Empirical Risk, Training Error and Test Error:

Let $S = \{(x_i, y_i)\}_{i=1}^m$ be the dataset where each $(x_i, y_i) \sim \mathcal{D}$. f be the machine learning model. $\hat{y}_i = f(x_i)$ be the prediction of x_i by f. $L(\hat{y}_i, y_i)$ be the loss.

The empirical risk is defined as

$$R^{\text{emp}}(f) = \frac{1}{m}\sum_{i=1}^{m} L(f(x_i), y_i)$$

The true risk is defined as

$$R^{\text{true}}(f) = \mathbb{E}_{(x_i,y_i)\sim\mathcal{D}}[L(f(x_i), y_i)]$$

The goal of machine learning is to find f that minimizes the true risk. Since $R^{\text{true}}(f)$ is not computable because $\mathcal{D}$ is unknown, we minimize $R^{\text{emp}}(f)$ instead. This process is called empirical risk minimization, also called training.

Let f_m be the solution of minimizing $R^{\text{emp}}(f)$ we found through training. This solution may not be optimal. We call $R^{\text{emp}}(f_m)$ training error, which means the average loss that f_m obtained on the training set S.

It is easy to prove that for any f,

$$R^{\text{true}}(f) = \mathbb{E}_{(x_i,y_i)\sim\mathcal{D}}[R^{\text{emp}}(f)]$$

By given $R^{\text{emp}}(f_m)$, we want to know how far $R^{\text{true}}(f_m)$ is away from $R^{\text{emp}}(f_m)$. The distance between $R^{\text{true}}(f_m)$ and $R^{\text{emp}}(f_m)$ is called the generalization error, usually bounded using concentration inequality.

Since $R^{\text{true}}(f_m)$ cannot be calculated, how do we evaluate the performance of a model f? We sample some data from $\mathcal{D}$. Let $S' = \{(x_i^t, y_i^t)\}_{i=1}^n$ and $(x_i^t, y_i^t) \sim \mathcal{D}$, and $S \cap S' = \varnothing$. The symbol t here means "test" and used to distinguish the test data from the training data. We call S' the test set.

The test error is defined as the average loss on the test set.

$$R^{\text{test}}(f) = \frac{1}{n}\sum_{i=1}^{n} L(f(x_i^t), y_i^t))$$

Note that

$$R^{\text{true}}(f) = \mathbb{E}_{(x_i,y_i)\sim\mathcal{D}}[R^{\text{test}}(f)]$$

That is, $R^{\text{test}}(f)$ is an unbiased estimator of $R^{\text{true}}(f)$. In practice, we consider the f that minimizes $R^{\text{test}}(f)$ as the best model, since $R^{\text{test}}(f)$ is much easier to calculate than $R^{\text{true}}(f)$. Note that our goal is to minimize $R^{\text{true}}(f)$, and the f that minimizes $R^{\text{test}}(f)$ is different from the one that minimizes $R^{\text{true}}(f)$.

We can consider $R^{\text{test}}(f)$ as the empirical risk of f on the test set S'. It satisfies the same generalization bound as $R^{\text{emp}}(f)$. The only difference is that they evaluate average loss on different samples from $\mathcal{D}$.

Therefore, both training error and test error can be empirical risk. Training error is the empirical risk obtained by empirical risk minimization.

Predictor and Posterior Distribution:

Let $p(x, y)$ be the PDF of $\mathcal{D}$ such that $p(x, y) = p(y|x)p(x)$. Let θ be the vector of parameters of the predictor f, then the true risk is

$$R^{\text{true}}(f) = \mathbb{E}_{(x_i,y_i)\sim\mathcal{D}}[L(f_\theta(x_i), y_i)]$$

For any x_i, suppose the true label y_i is generated according to a posterior distribution $p(y|x_i)$. This is a general case that the same x_i can have different labels with different probabilities. Then for the square loss and cross entropy loss, the predictor $f_\theta(x_i)$ approximates $\mathbb{E}_{y\sim p(y|x_i)}[y]$ and does not depend on $p(x)$.

In regression case[1], suppose $L(f_\theta(x_i) - y_i) = (f_\theta(x_i) - y_i)^2$, then

$$R^{\text{true}}(f) = \mathbb{E}_{x_i\sim p(x)}[\mathbb{E}_{y_i\sim p(y|x_i)}[(f_\theta(x_i) - y_i)^2]]$$

So minimizing $R^{\text{true}}(f)$ is to minimize $\mathbb{E}_{y_i\sim p(y|x_i)}[(f_\theta(x_i) - y_i)^2]$ for each x_i. Hence,

$$\frac{\partial}{\partial f_\theta(x_i)}\mathbb{E}_{y_i\sim p(y|x_i)}[(f_\theta(x_i) - y_i)^2] = \frac{\partial}{\partial f_\theta(x_i)}\int (f_\theta(x_i) - y)^2 p(y|x_i)dy = \int 2(f_\theta(x_i) - y)p(y|x_i)dy = 0$$

Since $\int p(y|x_i)dy = 1$, we have

$$f_\theta(x_i) = \int y\, p(y|x_i)dy = \mathbb{E}_{y\sim p(y|x_i)}[y]$$

In classification case, suppose we have d classes, the label y_i is often defined as one-hot labels $y_i = [y_i^1, y_i^2, ..., y_i^d]^T$. If x_i belongs to class $k, 1 \le k \le d$, then $y_i^k = 1$ and $y_i^j = 0$ for all $j \ne k$. This can be considered as a discrete distribution P. We let $f_\theta(x) = \hat{y}_i = [\hat{y}_i^1, \hat{y}_i^2, ..., \hat{y}_i^d]$ where $\sum_j \hat{y}_i^j = 1$, which is usually done by softmax. Then $f_\theta(x_i)$ can be considered as a distribution Q. Usually we define L as the cross entropy between P and Q:

$$L(f_\theta(x_i), y_i) = H(P, Q) = -\sum_{j=1}^{d} y_i^j \log \hat{y}_i^j$$

Here we treat the label y_i as a distribution P. It is not the same as the distribution $p(y|x_i)$ that generates the label y_i. For example, suppose $d = 2$, x_i can have two different labels $[1, 0]^T$ and $[0, 1]^T$ generated with different probability according to $p(y|x_i)$, each of them can be considered as a distribution. In general, consider we have a two dimensional distribution $p(x)$ where $x = [x_1, x_2]^T$ satisfies $x_1 \ge 0, x_2 \ge 0, x_1 + x_2 = 1$. Suppose we have a sample $x_0 = [0.3, 0.7]$, which can represent a Bernoulli distribution. The same x_0 can be sampled from any distribution $p(x)$. Any two samples from $p(x)$ can be regarded as two distributions such that we can use cross entropy as a metric to measure their distance.

Therefore, $R^{\text{true}}(f)$ is defined as

$$R^{\text{true}}(f) = \int H(P, Q)p(y|x_i)dy = \int (-\sum_{j=1}^{d} y^j \log \hat{y}_i^j)p(y|x_i)dy$$

Suppose $\hat{z}_i = [\hat{z}_i^1, \hat{z}_i^2, ..., \hat{z}_i^d]$ and $\hat{y}_i$ is obtained from $\hat{z}_i$ by applying softmax:

$$\hat{y}_i^k = \frac{e^{\hat{z}_i^k}}{\sum_j e^{\hat{z}_i^j}}$$

Then our goal becomes finding $\hat{z}_i$ to minimize $R^{\text{true}}(f)$. Using the derivative of cross entropy[2]:

$$\frac{\partial}{\partial \hat{z}_i} H(P, Q) = (\hat{y}_i - y)^T$$

[1] https://web.mit.edu/6.962/www/www_spring_2001/emin/slt.pdf

[2] https://stats.stackexchange.com/questions/277203/differentiation-of-cross-entropy

we have

$$\frac{\partial}{\partial \hat{z}_i} \int (-\sum_{j=1}^{d} y^j \log \hat{y}_i^j) p(y|x_i) dy = \int (\hat{y}_i - y)^T p(y|x_i) dy = 0$$

Therefore,

$$f_\theta(x_i) = \hat{y}_i = \int y\, p(y|x_i) dy = \mathbb{E}_{y \sim p(y|x_i)}[y]$$

$\mathbb{E}_{y \sim p(y|x_i)}[y]$ is known as the posterior mean of Bayes estimator.