Hoeffding Bound

Ruixin Guo

Department of Computer Science
Kent State University

February 24, 2023

Function Class

Recall: given a dataset $(x_i, y_i), i = 1, 2, ..., m$ where each $x_i \in \mathbb{R}^n$ and each $y_i \in \mathbb{R}$. Let $F : x_i \to y_i$ be a function which maps each x_i to y_i. F is unknown, and we want to use a function f to approximate F.

We need first make an assumption on what f should look like. In our assumption, the set of all possible f is called function class (or "hypothesis space" in some literature), denoted as $\mathcal{F}$.

For example, let $\mathcal{F} = \{f(x) = I(x \leq \theta), \theta \in \mathbb{R}\}$. Here we assume f looks like horizontal line. For each θ, $f(x)$ is a function. The set of $f(x)$ with respect to all possible θ forms the function class $\mathcal{F}$.

If f is a neural network, the architecture of f is fixed and the parameters are changeable. Each choice of parameters makes f a function. The set of f with respect to all possible choices of the parameters forms $\mathcal{F}$. The goal of machine learning is to search for the best f in $\mathcal{F}$ to approximate f^*.

Loss, Risk and Empirical Risk

We define the loss function on x_i as

$$L(x_i, y_i) = \mathbf{1}_{[f(x_i) \neq y_i]}$$

Suppose each (x_i, y_i) is sampled from a distribution D. The risk function is the average loss for the entire distribution D, defined as

$$R^{\text{true}}(f) = \mathbb{E}_{(x_i,y_i)\sim D}[L(x_i, y_i)] = \mathbb{E}_{(x_i,y_i)\sim D}[\mathbf{1}_{[f(x_i) \neq y_i]}]$$

However, we have no knowledge about the distribution of D. So usually we use empirical risk function instead. Empirical risk function is the average loss for the finite samples from D. For samples $x_1, x_2..., x_m$, the empirical risk function is defined as

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

Approximation Error and Estimation Error

Most problems in Machine Learning can be formalized as regulated empirical risk minimization:

$$f_m = \operatorname*{argmin}_{f \in \mathcal{F}} R^{\text{emp}}(f) + C\|f\|^2$$

Suppose the best function in $\mathcal{F}$ is f^*, which minimizes the true risk:

$$f^* = \operatorname*{argmin}_{f \in \mathcal{F}} R^{\text{true}}(f)$$

Let $R^* = \inf_f R^{\text{true}}(f)$, where f is from all measurable functions. R^* is the theoretical infimum risk, also known as Bayes Risk.

We would like to know how far $R^{\text{true}}(f_m)$ is from R^*:

$$R^{\text{true}}(f_m) - R^* = \underbrace{[R^{\text{true}}(f^*) - R^*]}_{\text{Approximation Error}} + \underbrace{[R^{\text{true}}(f_m) - R^{\text{true}}(f^*)]}_{\text{Estimation Error}}$$

Approximation Error and Estimation Error

$$R^{\text{true}}(f_m) - R^* = \underbrace{[R^{\text{true}}(f^*) - R^*]}_{\text{Approximation Error}} + \underbrace{[R^{\text{true}}(f_m) - R^{\text{true}}(f^*)]}_{\text{Estimation Error}}$$

We split the error $R^{\text{true}}(f_m) - R^*$ into two parts: Approximation Error and Estimation Error.

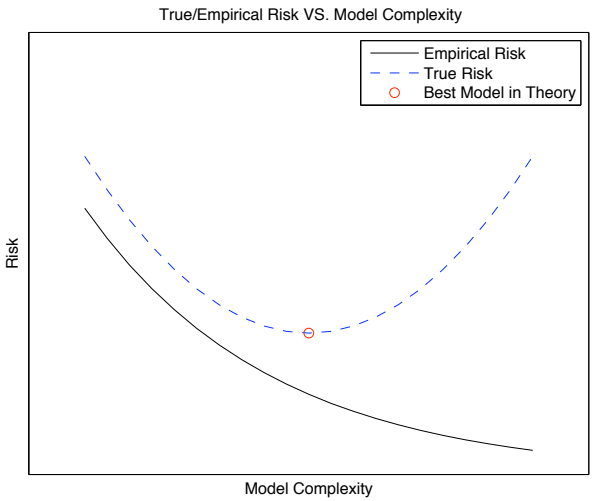
The Approximation Error is caused by the restriction of using $\mathcal{F}$. Because we have no knowledge about the distribution D, we cannot make an assumption that ensures f^* is included in $\mathcal{F}$. So the Approximation Error is difficult to measure.

The Estimation Error is caused by the usage of a finite sample that cannot completely represent the distribution D. This part of error is measurable and can be computed without any assumption. We will focus on this part of data.

Note that in both Approximation and Estimation error, we make no assumption of the data. In Statistical Learning Theory, generally there is no assumption made about the target, and it does not require any knowledge of the distribution D.

An Explanation of Estimation Error

Why we want to minimize the estimation error? The following figure gives us an illustration.



The $R^{\text{true}}(f^*)$ gives the best generalization of the model, i.e., when R^{true} is at its minimum, the gap between training error and test error is minimized.

If $R^{\text{true}}(f_m)$ is closer to $R^{\text{true}}(f^*)$, the model will have better generalizability.

We cannot compute $R^{\text{true}}(f_m)$ directly since we have no knowledge about R^{true}, however this value can be bounded. Consider another way to look at $R^{\text{true}}(f_m)$.

$$R^{\text{true}}(f_m) = R^{\text{emp}}(f_m) + [R^{\text{true}}(f_m) - R^{\text{emp}}(f_m)]$$

We already know $R^{\text{emp}}(f_m)$. Now we want to find an upper bound for $R^{\text{true}}(f_m) - R^{\text{emp}}(f_m)$.

Hoeffding's Inequality

For any $f \in \mathcal{F}$, we want to find a bound for $R^{\text{true}}(f) - R^{\text{emp}}(f)$.

Remember we defined $R^{\text{true}}(f)$ and $R^{\text{emp}}(f)$ as

$$R^{\text{true}}(f) = \mathbb{E}_{(x_i,y_i)\sim D}[\mathbf{1}_{[f(x_i)\neq y_i]}] \qquad R^{\text{emp}}(f) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}_{[f(x_i)\neq y_i]}$$

The following theorem gives us a probability bound for $R^{\text{true}}(f) - R^{\text{emp}}(f)$.

Theorem (Hoeffding): Let $Z_1, ..., Z_m$ be m iid random variables, and h is a bounded function, $h(Z) \in [a, b]$. Then for all $\epsilon > 0$ we have:

$$P_{\mathbf{Z}\sim D^m}\left[\left|\frac{1}{m}\sum_{i=1}^{m} h(Z_i) - \mathbb{E}_{\mathbf{Z}\sim D^m}[h(Z)]\right| \geq \epsilon\right] \leq 2\exp\left(-\frac{2m\epsilon^2}{(b-a)^2}\right)$$

This is a well known concentration inequality, i.e., the inequality bounding the deviation of a function of random variables from its average. Remember that $R^{\text{true}}(f) = \mathbb{E}R^{\text{emp}}(f)$ and $R^{\text{emp}}(f)$ converges to $R^{\text{true}}(f)$ when $m \to \infty$ (Law of Large Numbers).

Proof of Hoeffding's Inequality

Theorem (Hoeffding): Let $Z_1, ..., Z_m$ be m iid random variables, and h is a bounded function, $h(Z) \in [a, b]$. Then for all $\epsilon > 0$ we have:

$$P_{\mathbf{Z} \sim D^m} \left[\left| \frac{1}{m} \sum_{i=1}^{m} h(Z_i) - \mathbb{E}_{\mathbf{Z} \sim D^m}[h(Z)] \right| \geq \epsilon \right] \leq 2 \exp\left(-\frac{2m\epsilon^2}{(b-a)^2} \right)$$

Proof: This proof utilizes the Hoeffding's Lemma, suppose $X \in [a, b]$:

$$\mathbb{E}(e^{t(X - \mathbb{E}(X))}) \leq \exp(\frac{1}{8} t^2 (b-a)^2)$$

We will prove this in the Appendix. Let $X_1, X_2, ..., X_m$ be iid random variables, each $X_i \in [a, b]$, and $S_m = \sum_{i=1}^{n} X_i$, then for $\epsilon, t > 0$:

$$\begin{aligned}
P(S_m - \mathbb{E}(S_m) \geq \epsilon) &= P(\exp(t(S_m - \mathbb{E}(S_m))) \geq \exp(t\epsilon)) \\
&\leq \exp(t\epsilon)\mathbb{E}[\exp(t(S_m - \mathbb{E}(S_m)))] \quad \text{[Chernoff Bound]} \\
&= \exp(t\epsilon) \prod_{i=1}^{m} \mathbb{E}[\exp(t(S_i - \mathbb{E}(S_i)))] \\
&\leq \exp(t\epsilon) \prod_{i=1}^{m} \exp(\frac{t^2 (b-a)^2}{8}) \quad \text{[Hoeffding's Lemma]} \\
&= \exp(-t\epsilon + \frac{1}{8} t^2 m (b-a)^2)
\end{aligned}$$

Proof of Hoeffding's Inequality

Since $-t\epsilon + \frac{1}{8}t^2 m(b-a)^2$ is quadratic with respect to t, when $t = \frac{4\epsilon}{m(b-a)^2}$, we have $\min(-t\epsilon + \frac{1}{8}t^2 m(b-a)^2) = -\frac{2\epsilon^2}{m(b-a)^2}$, therefore

$$P(S_m - \mathbb{E}(S_m) \geq \epsilon) \leq \exp(-\frac{2\epsilon^2}{m(b-a)^2})$$

Let $Z = \frac{1}{m}S_m$, we have

$$P(Z - \mathbb{E}(Z) \geq \frac{\epsilon}{m}) \leq \exp(-\frac{2\epsilon^2}{m(b-a)^2}) \Rightarrow P(Z - \mathbb{E}(Z) \geq \epsilon) \leq \exp(-\frac{2m\epsilon^2}{(b-a)^2})$$

This is only one-side probability, for two-side probability:

$$P(|Z - \mathbb{E}(Z)| \geq \epsilon) = P(Z - \mathbb{E}(Z) \geq \epsilon) + P(Z - \mathbb{E}(Z) \leq -\epsilon) \leq 2\exp(-\frac{2m\epsilon^2}{(b-a)^2})$$

The second term again uses the Chernoff bound

$$P(Z - \mathbb{E}(Z) \leq -\epsilon) = P(\exp(t(Z - \mathbb{E}(Z))) \geq \exp(-t\epsilon)) \leq \exp(-t\epsilon)\mathbb{E}[\exp(t(Z - \mathbb{E}(Z)))]$$

where $t < 0$.

Bound $R^{\text{true}}(f) - R^{\text{emp}}(f)$ by Hoeffding's Inequality

Remember that

$$R^{\text{true}}(f) = \mathbb{E}_{(x_i,y_i)\sim D}[\mathbf{1}_{[f(x_i)\neq y_i]}] \qquad R^{\text{emp}}(f) = \frac{1}{m}\sum_{i=1}^{m}\mathbf{1}_{[f(x_i)\neq y_i]}$$

Since each $\mathbf{1}_{[f(x_i)\neq y_i]} = \{0,1\} \in [0,1]$, we can bound $R^{\text{true}}(f) - R^{\text{emp}}(f)$ by

$$P\left[\left|R^{\text{true}}(f) - R^{\text{emp}}(f)\right| \geq \epsilon\right] \leq 2\exp\left(-2m\epsilon^2\right) \quad \text{[two-side]}$$

$$P\left[R^{\text{true}}(f) - R^{\text{emp}}(f) \geq \epsilon\right] \leq \exp\left(-2m\epsilon^2\right) \quad \text{[one-side]}$$

By letting $\delta = 2\exp(-2m\epsilon^2) \Rightarrow \epsilon = \sqrt{\frac{\log\frac{2}{\delta}}{2m}}$, we have

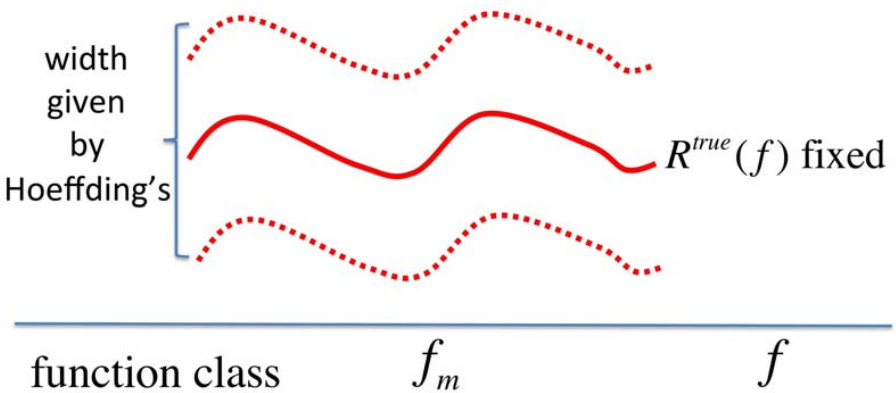
$$P\left[\left|R^{\text{true}}(f) - R^{\text{emp}}(f)\right| \leq \sqrt{\frac{\log\frac{2}{\delta}}{2m}}\right] \geq 1-\delta \quad \text{[two-side]}$$

$$P\left[R^{\text{true}}(f) - R^{\text{emp}}(f) \leq \sqrt{\frac{\log\frac{1}{\delta}}{2m}}\right] \geq 1-\delta \quad \text{[one-side]}$$

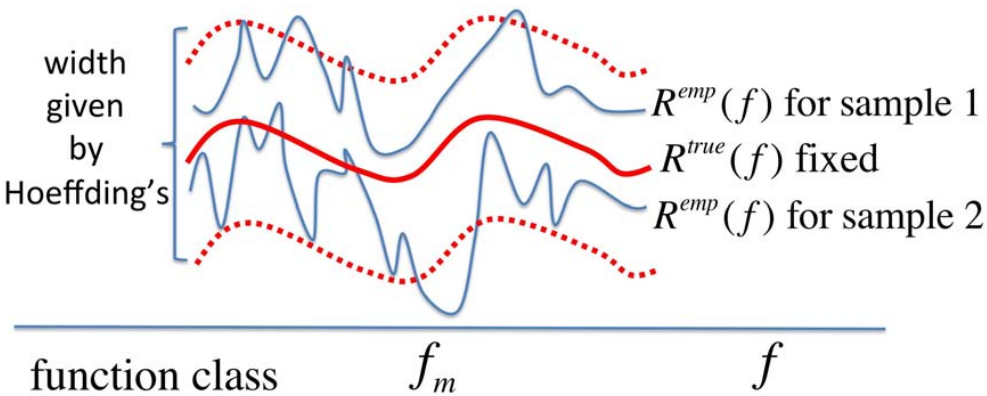
The Limitation of Hoeffding's Inequality

The Limitation of Hoeffding's Inequality:

- Hoeffding's inequality only consider a single function f.
- $R^{\text{emp}}(f)$ is more likely to deviate from $R^{\text{true}}(f)$ because our goal is to minimize $R^{\text{emp}}(f)$ with given data.
- When the function class $\mathcal{F}$ is larger, the deviation will happen more easily. (Consider $|\mathcal{F}|$ as a capacity measure of $\mathcal{F}$, greater $|\mathcal{F}|$ is usually caused by greater model complexity, and tends to make $R^{\text{emp}}(f)$ smaller.)



The $R^{\text{true}}(f)$ and its bounds for $R^{\text{emp}}(f)$.



Different $R^{\text{emp}}(f)$ based on different samples.

We need to build a bound considering the size of $|\mathcal{F}|$!

Uniform Bound

The Hoeffding's Inequality is a bound for a single f. Consider $|\mathcal{F}|$ be finite, the uniform bound is the bound for all fs in $|\mathcal{F}|$.

Suppose $Z_1, Z_2, ..., Z_N$ are iid random variables. Consider the union of probability:

$$P[\exists Z_i \in \{Z_1, Z_2, ..., Z_N\} : Z_i \geq \epsilon] = P[\bigcup_{i=1}^{N} (Z_i \geq \epsilon)] \leq \sum_{i=1}^{N} P(Z_i \geq \epsilon) = N \cdot P(Z_i \geq \epsilon)$$

The second inequality above uses the fact that, supposing A and B are two conditions:

$$P(A \bigcup B) = P(A) + P(B) - P(A \bigcap B) \leq P(A) + P(B)$$

For two-side bound, $P\left[\left|R^{\text{true}}(f) - R^{\text{emp}}(f)\right| \geq \epsilon\right] \leq 2\exp\left(-2m\epsilon^2\right)$ for a single $f \in \mathcal{F}$. Suppose $|\mathcal{F}| = N$ and N is finite, the union bound for all functions in $\mathcal{F}$ is

$$\begin{aligned} &P\left[\exists f_i \in \{f_1, f_2, ..., f_N\} : \left|R^{\text{true}}(f_i) - R^{\text{emp}}(f_i)\right| \geq \epsilon\right] \\ &= P\left[\bigcup_{i=1}^{m} \left|R^{\text{true}}(f_i) - R^{\text{emp}}(f_i)\right| \geq \epsilon\right] \leq 2N\exp\left(-2m\epsilon^2\right) \end{aligned}$$

Uniform Bound

$$P\left[\exists f_i \in \{f_1, f_2, ..., f_N\} : \left|R^{\text{true}}(f_i) - R^{\text{emp}}(f_i)\right| \geq \epsilon\right]$$
$$= P\left[\bigcup_{i=1}^{m} \left|R^{\text{true}}(f_i) - R^{\text{emp}}(f_i)\right| \geq \epsilon\right] \leq 2N \exp\left(-2m\epsilon^2\right)$$

By letting $\delta = 2N \exp\left(-2m\epsilon^2\right) \Rightarrow \epsilon = \sqrt{\frac{\log N + \log \frac{2}{\delta}}{2m}}$

$$P\left[\exists f_i \in \{f_1, f_2, ..., f_N\} : \left|R^{\text{true}}(f_i) - R^{\text{emp}}(f_i)\right| \geq \sqrt{\frac{\log N + \log \frac{2}{\delta}}{2m}}\right] \leq \delta$$

That is,

$$P\left[\forall f_i \in \{f_1, f_2, ..., f_N\} : \left|R^{\text{true}}(f_i) - R^{\text{emp}}(f_i)\right| \leq \sqrt{\frac{\log N + \log \frac{2}{\delta}}{2m}}\right] \geq 1 - \delta$$

Similarly, for one-side bound

$$P\left[\forall f_i \in \{f_1, f_2, ..., f_N\} : R^{\text{true}}(f_i) - R^{\text{emp}}(f_i) \leq \sqrt{\frac{\log N + \log \frac{1}{\delta}}{2m}}\right] \geq 1 - \delta$$

Estimation Error

Now let's find a bound for the estimation error. Suppose f_m is the minimizer of R^{emp}, we have:

$$R^{\text{true}}(f_m) = R^{\text{true}}(f_m) - R^{\text{true}}(f^*) + R^{\text{true}}(f^*)$$

Then will use the fact that $R^{\text{emp}}(f^*) - R^{\text{emp}}(f_m) \geq 0$. Then

$$\begin{aligned} R^{\text{true}}(f_m) &\leq [R^{\text{emp}}(f^*) - R^{\text{emp}}(f_m)] + R^{\text{true}}(f_m) - R^{\text{true}}(f^*) + R^{\text{true}}(f^*) \\ &\leq |R^{\text{true}}(f^*) - R^{\text{emp}}(f^*)| + |R^{\text{true}}(f_m) - R^{\text{emp}}(f_m)| + R^{\text{true}}(f^*) \\ &\leq 2 \sup_{f \in \mathcal{F}} |R^{\text{true}}(f) - R^{\text{emp}}(f)| + R^{\text{true}}(f^*) \end{aligned}$$

Since with probability at least $1 - \delta$, $|R^{\text{true}}(f) - R^{\text{emp}}(f)| \leq \sqrt{\frac{\log N + \log \frac{2}{\delta}}{2m}}$ we can bound the estimation error by

$$P\left(R^{\text{true}}(f_m) - R^{\text{true}}(f^*) \leq 2\sqrt{\frac{\log N + \log \frac{2}{\delta}}{2m}} \right) \geq 1 - \delta$$

Appendix: Hoeffding's Lemma

Lemma: Let X be any random variable in $[a, b]$, then for all $t \in \mathbb{R}$:

$$\mathbb{E}(e^{t(X-\mathbb{E}(X))}) \leq \exp(\frac{1}{8}t^2(b-a)^2)$$

Proof: Since e^{tx} is a convex function of x, we have that for all $x \in [a, b]$.

$$e^{tx} \leq \frac{b-x}{b-a}e^{ta} + \frac{x-a}{b-a}e^{tb} \Rightarrow \mathbb{E}[e^{tx}] \leq \frac{b-\mathbb{E}[x]}{b-a}e^{ta} + \frac{\mathbb{E}[x]-a}{b-a}e^{tb}$$

Replacing x by $x - \mathbb{E}(x)$, we know that $\mathbb{E}(x - \mathbb{E}(x)) = 0$, thus

$$\mathbb{E}[e^{t(x-\mathbb{E}(x))}] \leq \frac{b}{b-a}e^{ta} + \frac{-a}{b-a}e^{tb}$$

Let $u = t(b-a)$ and $e^{g(u)} = \frac{b}{b-a}e^{ta} + \frac{-a}{b-a}e^{tb}$, then

$$\begin{aligned} g(u) &= \log\left(\frac{b}{b-a}e^{ta} + \frac{-a}{b-a}e^{tb}\right) = \log\left(e^{ta}\left(\frac{b}{b-a} + \frac{-a}{b-a}e^{t(b-a)}\right)\right) \\ &= ta + \log\left(\gamma e^u + (1-\gamma)\right) = -\gamma u + \log\left(\gamma e^u + (1-\gamma)\right) \end{aligned}$$

Here we let $\gamma = -\frac{a}{a-b}$.

Appendix: Hoeffding's Lemma

Since $g(u) = -\gamma u + \log\left(\gamma e^u + (1-\gamma)\right)$, we have $g(0) = 0$, $g'(0) = 0$ and $g''(u) = \frac{\gamma e^u (1-\gamma)}{(\gamma e^u + (1-\gamma))^2} \leq \frac{1}{4}$ (Let $m = \gamma e^u$ and $n = 1-\gamma$, use the AMGM inequality $mn \leq \frac{(m+n)^2}{4}$).

Then use Taylor's Theorem, there is a $\xi \in [0, u]$ such that

$$g(u) = g(0) + g'(0)u + \frac{1}{2}g''(\xi)u^2 \leq \frac{1}{8}u^2$$

Therefore,

$$\mathbb{E}[e^{t(x-\mathbb{E}(x))}] \leq e^{g(u)} \leq e^{\frac{1}{8}u^2} = e^{\frac{t^2(b-a)^2}{8}}$$