Wang Yanan Institute for Studies in Economics

MICROECONOMETRICS COURSE REPORT#1

Notes on VC theory

Student
Baohao Wei
15220172202657

Supervisor
Jiaming Mao

Summary

In this report, we will introduce a fundmental concept in learning theory: VC dimension, which measues the complexity of statistical models. Larger VC dimension means more flexible models. Here, we start from basic classification problem and try to understand why we need VC dimension in a theoretical way.

Keywords: VC dimension; Growth funcion; Dichotomy; Hoeffding Inequality

1 Basic Settings

Let's start with binary classification problem. Given input space $\mathcal{X}$ and output space $\mathcal{Y} = \{+1, -1\}$, suppose there is some unknown probability distribution P_{XY} on $\mathcal{X} \times \mathcal{Y}$, some hypothesis space $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$, loss function $l : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$ and N independent sample points $\{(x_i, y_i)\}_{i=1}^N$that drawn from P_{XY}. Our target is to learn some function $g \in \mathcal{H}$ that minimizes the possible error. Here, the possible error is measured using ***out-of-sample error*** that defined as below:

$$E_{out}(g) = \mathbb{E}_{P_{XY}}[l(g, X, Y)], \tag{1.1}$$

where loss function is defined as binary error: $l(h, X, Y) = I\{h(x) \neq y\}$.

2 Hoeffding Inequality

However, since the joint distribution P_{XY} is unknown to us, $E_{out}(g)$ actually can not be estimated. But we have some independent sample points obtained from this unknown distribution. Now, our idea is to use these sample points to indirectly estimate $E_{out}(g)$. Here, we define the ***in-sample error***, which measures the goodness of g within these observed sample points:

$$E_{in}(g) = \frac{1}{N} \sum_{i=1}^{N} l(g, x_i, y_i). \tag{2.1}$$

Now, let's study the relationship between $E_{out}(g)$ and $E_{in}(g)$ and how accrately we can use $E_{in}(g)$ to estimate $E_{out}(g)$. Following inequality provides us a useful tool:

Hoeffding Inequity. *Suppose independent random variables $\xi_1, \xi_2, ..., \xi_N$ satisfies $\xi_i \in [a_i, b_i]$ for $i = 1, 2, ..., N$. Denote $\overline{\xi} = 1/N \sum_{i=1}^N \xi_i$, then for $\forall \epsilon > 0$,*

$$Pr(\overline{\xi} - E(\overline{\xi}) > \epsilon) \leq exp(-\frac{2N^2\epsilon^2}{\sum_{i=1}^N (b_i - a_i)^2}). \tag{2.2}$$

As we can see, Hoeffding inequality measures the difference between these two errors in a probability language. Actually, for any fixed function h, $\overline{\xi}$ is our in-sample error, $E(\overline{\xi})$ is our out-of-sample error. Since loss function is binary, then $a_i = 0, b_i = 1$. Then, we can write the Hoeffding inequality into the following form:

$$Pr(|E_{in} - E_{out}| > \epsilon) \leq 2exp(-2N\epsilon^2). \tag{2.3}$$

Note that the right hand of Hoeffding inequality is an exponential function, and as N goes to infinity, the bound shrinks to zero.

3 Symmetrization

So far, we have discussed how to use our sample points to meansure out-of-sample error for any fix function. However, since we are learning the unknown function from hypothesis space $\mathcal{H}$, we hope that for any funcion h in hypothesis space $\mathcal{H}$, above bound holds. In other words, we hope to show that $Pr(sup_{h \in \mathcal{H}}(E_{out}(h) - E_{in}(h) > \epsilon))$ is bounded.

Before that, we introduce some definitions. First, we define the loss class $\mathcal{F}$:

$$\mathcal{F} = \{f_h | h \in \mathcal{H}\} \tag{3.1}$$

where $f_h(x_i, y_i) = l(h, x_i, y_i) : \mathcal{Z} \to \mathcal{R}^+$. Note that the elements in $\mathcal{F}$ and $\mathcal{H}$ are one-to-one corresponding, so learning from $\mathcal{H}$ is actually learning from $\mathcal{F}$[1]. Then, we define the projection from $\mathcal{F}$ to $\{(x_i, y_i)\}_{i=1}^N$:

$$\mathcal{F}^P = \mathcal{F}((x_1, y_1), \cdots, (x_N, y_N)) = \{(f(x_1, y_1), \cdots, f(x_N, y_N)) | f \in \mathcal{F}\} \tag{3.2}$$

Note that, $\mathcal{F}^P$ is actually a set contains some N dimensional vector. Since we are using the binary loss, $f(x_i, y_i)$ actually can only take 0 or 1. Hence, $\mathcal{F}^P$, which contains N-dimensional binary vector, has finite elements (2^N at most) even if the original $\mathcal{F}$ (or hypothesis space $\mathcal{H}$) is infinite. Also, we can define the in-sample error for f: $E_{in}(h) = E_N(f_h) = \frac{1}{N}\sum_{i=1}^N f_h(x_i, y_i)$, similar definition for out-of-sample error $E_N(f)$. We also introduce the ghost sample $\{(x_i^*, y_i^*)\}_{i=1}^N$, which are also drawn independently from P_{XY}, and independent of the original sample. Now, we formally introduce the following lemma:

Symmetrization Lemma. *For any $\epsilon > 0$, and $N\epsilon^2 \geq 2$, we have*

$$Pr(\sup_{f\in\mathcal{F}}(E(f) - E_N(f) > \epsilon)) \leq 2Pr(\sup_{f\in\mathcal{F}}(E_N^*(f) - E_N(f) > \frac{\epsilon}{2})) \tag{3.3}$$

where $E_N^*(f)$ is defined on ghost sample.

Now, we can use two finite sets $\mathcal{F}^P$, $\mathcal{F}^{P^*}$ to get the uniform bound. The right hand side of the inequality is the supremum of the project set from $\mathcal{F}$ to our original sample and ghost sample, which is set of $2N$-dimensional binary vectors. We can use it to get our uniform bound:

$$\begin{aligned} Pr(\sup_{f\in\mathcal{F}}(E_N^*(f) - E_N(f) > \frac{\epsilon}{2})) &\leq Pr(\bigcup_{f\in\mathcal{F}}(E_N^*(f) - E_N(f) > \frac{\epsilon}{2})) \\ &\leq |\mathcal{F}^P| Pr(E_N^*(f) - E_N(f) > \frac{\epsilon}{2}) \\ &\leq 2^{2N} exp(-\epsilon^2 N/2) \\ &= exp(N(2ln2 - \epsilon^2/2)). \end{aligned} \tag{3.4}$$

It seems that we get a useful bound. However, since $2ln2 \approx 1.39$, actually any reansonable choice of ϵ could lead to a positive $(2ln2 - \epsilon^2/2)$, and as N goes to infinity, this bound is meaningless. The reason is that we use the maximum number of elements in $\mathcal{F}^P$, which is also an exponential function. Hence, the advantage that we get from Hoeffding inequality is offset. If the true number of elements in $\mathcal{F}^P$ is much smaller than 2^{2N}, then our problem can be solved.

4 Vapnik-Chervonenkis Dimension

First, we introduce the concept of ***shatter***. We say that $\mathcal{F}$ shatters the given sample points $\{(x_i, y_i)\}_{i=1}^N$ if the projection set $\mathcal{F}^P$ contains all possible 2^N binary vectors. Obviously, $|\mathcal{F}^P|$ depends on the choice of $\mathcal{F}$ and the layout of sample points $\{(x_i, y_i)\}_{i=1}^N$.

For example, given two independent points x_1, x_2 on x-axis, let $\mathcal{F}_1$ denote the positive rays[2] and $\mathcal{F}_2$ denote the positive intervals[3]. As we can see from the graph below, $\mathcal{F}_2$ can shatter these two points while $\mathcal{F}_1$ can not. Also, the choice of sample points can decide whether $\mathcal{F}$ can shatter the sample. For example, we consider using positive rectangles[4] to shatter several sample points on $\mathbb{R}^2$. The following graph shows two different types of layout of sample points, where the first one can be shattered and the second one can not. The reason is that, for the second sample, the

[1] Actually, it is possible that different h may lead to the same loss. That is, although $h_1 \neq h_2$, $l(h_1, x_i, y_i) = l(h_2, x_i, y_i)$ for $i = 1, ..., N$. Here, we view that these two functions are in the same equivalent class.

[2] Positive ray is defined as $f_\alpha(x) = I(x > \alpha)$, $\alpha \in \mathbb{R}$

[3] Positive interval is defined as $f_{\alpha,\beta}(x) = I(x \in [\alpha, \beta])$, $\alpha, \beta \in \mathbb{R}$

[4] Position rectangle function defined on $\mathbb{R}^2$ takes 1 if a point is within the rectangle, 0 otherwise. The funcion requires four parameters.

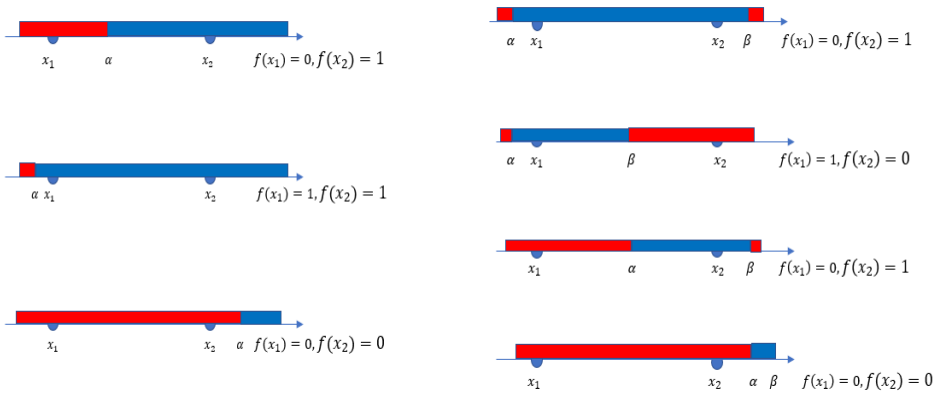

Figure 1: Positive rays and Positive intervals

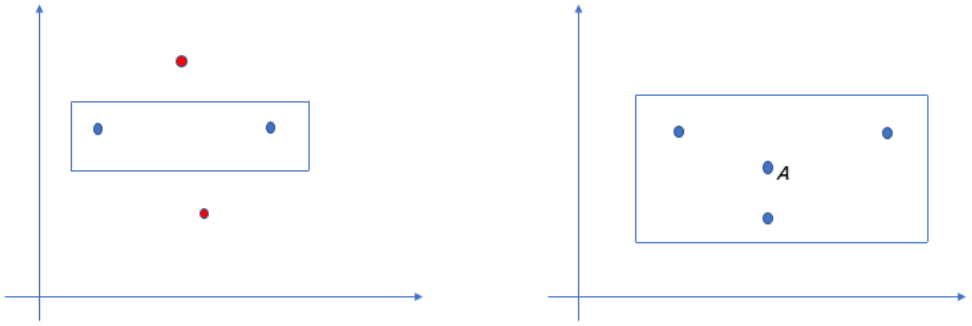

Figure 2: Different layouts give different $|\mathcal{F}^P|$

loss $f(A)$ is determined if other three sample points share the same loss. That is, if other three points are contained in a rectangle, point A must be contained in the same rectangle. Generally, the binary vector that generated by $f \in \mathcal{F}$ projected on a sample points is called a ***dichotomy***.

Now, let's introduce the definition of ***growth function***. Given positive integer N and funcion space $\mathcal{F}$, the value of growth function $S_{\mathcal{F}}(N)$ equals the maximum number of dichotomies that generated by $\mathcal{F}$.

$$S_{\mathcal{F}}(N) = \sup_{(x_i, y_i)} |\mathcal{F}((x_1, y_1), ..., (x_N, y_N))| = sup\, |\mathcal{F}^P| \tag{4.1}$$

Note that, if $\mathcal{F}$ shatters $\{(x_i, y_i)\}_{i=1}^N$, then we must have $S_F(N) = 2^N$; similarly, if we have $S_F(N) = 2^N$, there must exsist one sample $\{(x_i, y_i)\}_{i=1}^N$ shattered by $\mathcal{F}$. Recall the bound that we get in the previous section, we hope that $|\mathcal{F}^P|$ contains as fewer number of elements as possible. That is, we do not want $\mathcal{F}$ to shatter all the sample points, which lead to an exponential coefficient. Another important thing to notice is that, if for some $\mathcal{F}$, there exsists some N_0 such that $S_F(N_0) < 2^{N_0}$, then for any $N_1 > N_0$, we have $S_F(N_1) < 2^{N_1}$. Also, we call these N's which statisfies $S_F(N) < 2^N$ as ***break points***.

Finally, we can introduce the definition for VC dimension. Given some function space $\mathcal{F}$, the ***VC dimension*** of $\mathcal{F}$, denoted as $d_{\mathcal{F}}$, is the maximum integer N statisfies:

$$S_F(N) = 2^N \tag{4.2}$$

If there is no such N exsists, we let $d_{\mathcal{F}} = \infty$. Based on this definition, we can see that $\mathcal{F}$ can shatter at most N sample points and any sample contains more than N sample points can not be shattered by $\mathcal{F}$. That is, if we have samples whose number of points N is larger than $d_{\mathcal{F}}$, $|\mathcal{F}^P|$ contains smaller elements than 2^N, then we have a better coefficient. Then, we will give this better upper bound.

Sauer's Lemma. *Suppose the VC dimension of some function space $\mathcal{F}$ is $d_{\mathcal{F}} < \infty$. Then for any positive integer N, we have:*

$$S_F(N) \leq \sum_{i=0}^{d_{\mathcal{F}}} C_N^i \tag{4.3}$$

Then, we use this conclusion to modify our previous bound. For $N > d_{\mathcal{F}}(N)$:

$$\begin{aligned}
(\frac{d_{\mathcal{F}}}{N})^{d_{\mathcal{F}}} S_F(N) &\leq (\frac{d_{\mathcal{F}}}{N})^{d_{\mathcal{F}}} \sum_{i=0}^{d_{\mathcal{F}}} C_N^i \\
&\leq \sum_{i=0}^{d_{\mathcal{F}}} C_N^i (\frac{d_{\mathcal{F}}}{N})^i \\
&\leq \sum_{i=0}^{N} C_N^i (\frac{d_{\mathcal{F}}}{N})^i \\
&= (1 + \frac{d_{\mathcal{F}}}{N})^N \leq e^{d_{\mathcal{F}}}
\end{aligned} \tag{4.4}$$

That is,

$$S_F(N) \leq (\frac{eN}{d_{\mathcal{F}}})^{d_{\mathcal{F}}} \tag{4.5}$$

By the symmetrization lemma, our new bound becomes[1]:

$$\begin{aligned}
Pr(\sup_{f \in \mathcal{F}}(E(f) - E_N(f) > \epsilon)) &\leq 2Pr(\sup_{f \in \mathcal{F}}(E_N^*(f) - E_N(f) > \frac{\epsilon}{2})) \\
&\leq 2S_F(2N) exp(-\epsilon^2 N/2) \\
&\leq 2(\frac{2eN}{d_{\mathcal{F}}})^{d_{\mathcal{F}}} exp(-\epsilon^2 N/2)
\end{aligned} \tag{4.6}$$

Compared with (3.4), we can see that now, our new coefficient is a polynominal instead of an exponential. Let the right hand side be δ, then we can find that:

$$\epsilon = \sqrt{\frac{2}{N} log 2 + \frac{2d}{N} log(\frac{2Ne}{d}) + \frac{2}{N} log \frac{1}{\delta}} = \epsilon(\delta, d_{\mathcal{F}}, N) \tag{4.7}$$

It shows that no matter what final hypothesis funcion f that our learning algorithm gives, we have the following inequality with probability at least $1 - \delta$:

$$E(f) \leq E_N(f) + \epsilon(\delta, d_{\mathcal{F}}, N) \tag{4.8}$$

Intuitively, if our learning model is very simple, although its VC dimension is very small and we can bound the difference between E_{in} and E_{out} error very well (In other words, E_{in} and E_{out} are very close.), the E_{in} may be very high, which lead to a high E_{out} in the end. On the other hand, if we use a complex model, although it can lead to a low E_{in}, we can not bound the difference between E_{in} and E_{out} very well. That also may lead to a poor estimation.

[1] Since we use our original sample and ghost sample, then growth funcion should be $S_F(2N)$.

5 Mathematical proofs

5.1 Proof of Hoeffding Inequality

To prove the inequality, we need following lemmas.

Lemma 1. $I(f(x)) \leq e^{\eta f(x)}$ for any $\eta \leq 0$, where $I(x)$ is the indicator function, $I(x) = 1$ if $x > 0$ and $I(x) = 0$ otherwise.

Lemma 2. Suppose Z is a random variable, and $a \leq Z \leq b$. Then for any η, we have the following inequality:

$$\mathbb{E}\left[e^{\eta Z}\right] \leq \frac{b - \mathbb{E}[Z]}{b - a} e^{\eta a} + \frac{\mathbb{E}[Z] - a}{b - a} e^{\eta b} \tag{5.1}$$

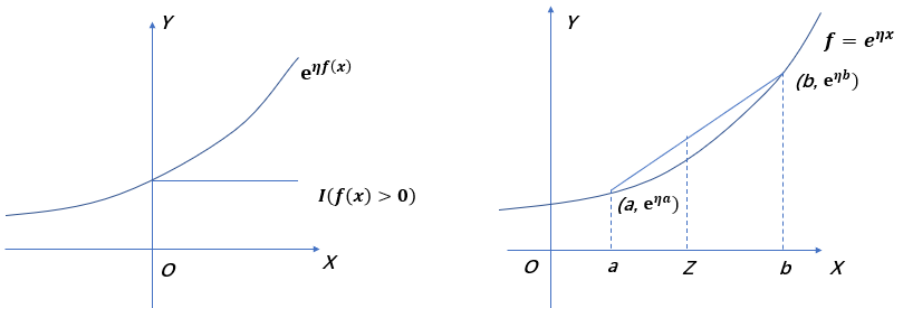

Figure 3: Figures for the proof of lemma

The proof of lemma 1 and lemma 2 is simple. As the figure shows, the indicator function can always be bounded by exponential function, then we have the lemma 1. Also, since the exponential functions are convex, they are always beneath the linear function that connects two different points, as the figure shows. Then, we have the following inequality:

$$e^{\eta Z} \leq e^{\eta a} + \frac{e^{\eta b} - e^{\eta a}}{b - a}(z - a) \tag{5.2}$$

Then, we take the expectation of left side and right side, we have our lemma 2. Now, we start the formal proof of Hoeffding inequality. Note that:

$$\begin{aligned} Pr(\bar{\xi} - \mathbb{E}[\bar{\xi}] > \epsilon) &= Pr(N\bar{\xi} - N\mathbb{E}[\bar{\xi}] > N\epsilon) \\ &= \mathbb{E}(I(N\bar{\xi} - N\mathbb{E}[\bar{\xi}] - N\epsilon)) \\ &\leq \mathbb{E}[\exp(\eta(N\bar{\xi} - N\mathbb{E}[\bar{\xi}] - N\epsilon))] \\ &= e^{-N\eta\epsilon} \prod_{i=1}^{N} \mathbb{E}\left[\exp\left(\eta\left(\xi_i - \mathbb{E}\left[\xi_i\right]\right)\right)\right] \end{aligned} \tag{5.3}$$

where the second transformation is based on lemma 1 and last trasformation is based on the indepedence of $\xi_i's$. That is, $\mathbb{E}[\exp(\eta(N\bar{\xi} - N\mathbb{E}[\bar{\xi}] - N\epsilon))] = e^{-N\eta\epsilon}\mathbb{E}\{exp[\eta(\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\mathbb{E}\xi_i)]\} = e^{-N\eta\epsilon}\mathbb{E}\{exp[\eta\sum_{i=1}^{N}(\xi_i - \mathbb{E}\xi_i)]\}$ Since $\xi_i's$ are independent with each other, by the definition of covariance, we have $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ if X and Y are independent. Now we have $e^{-N\eta\epsilon}\mathbb{E}\{exp[\eta\sum_{i=1}^{N}(\xi_i - \mathbb{E}\xi_i)]\} = e^{-N\eta\epsilon}\prod_{i=1}^{N}\mathbb{E}\left[\exp\left(\eta\left(\xi_i - \mathbb{E}\left[\xi_i\right]\right)\right)\right]$

Next, we try to show that $\mathbb{E}\left[\exp\left(\eta\left(\xi_i - \mathbb{E}\left[\xi_i\right]\right)\right)\right]$ is bounded using lemma 2. We have the

following:

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\eta\left(\xi_i - \mathbb{E}\left[\xi_i\right]\right)\right)\right] &= e^{-\eta\mathbb{E}\xi_i}\mathbb{E}\left[\exp\left(\eta\xi_i\right)\right] \\
&\leq e^{-\eta\mathbb{E}\xi_i}\left(\frac{b_i - \mathbb{E}\xi_i}{b_i - a_i}e^{\eta a_i} + \frac{\mathbb{E}\xi_i - a_i}{b_i - a_i}e^{\eta b_i}\right) \\
&= e^{-\eta(\mathbb{E}\xi_i - a_i)}\left(1 - \frac{\mathbb{E}\xi_i - a_i}{b_i - a_i} + \frac{\mathbb{E}\xi_i - a_i}{b_i - a_i}e^{\eta(b_i - a_i)}\right) \\
&= \exp\left(-\eta(b_i - a_i)p_i\right) \cdot \exp\log\left(1 - p_i + p_i e^{\eta(b_i - a_i)}\right) \\
&= \exp\left(-\eta_i p_i + \log\left(1 - p_i + p_i e^{\eta_i}\right)\right)
\end{aligned} \tag{5.4}$$

where $p_i = \frac{\mathbb{E}\xi_i - a_i}{b_i - a_i}$ and $\eta_i = \eta(b_i - a_i)$. Now consider the last line, which is a function of η_i and $p_i's$ are constants. We denote that $L(\eta_i) = \left(-\eta_i p_i + \log\left(1 - p_i + p_i e^{\eta_i}\right)\right)$ and try to show that it is bounded. We use Taylor expansion to control it, that is $L\left(\eta_i\right) = L(0) + L'(0)\eta_i + \frac{1}{2}L''(\zeta)\eta_i^2$. We take the derivatives of $L(\eta_i)$ and have the following:

$$\begin{aligned}
L'\left(\eta_i\right) &= -p_i + \frac{p_i e^{\eta_i}}{1 - p_i + p_i e^{\eta_i}} = -p_i + \frac{p_i}{\left(1 - p_i\right)e^{-\eta_i} + p_i} \\
L''\left(\eta_i\right) &= \frac{p_i\left(1 - p_i\right)e^{-\eta_i}}{\left[\left(1 - p_i\right)e^{-\eta_i} + p_i\right]^2} \leq \frac{\frac{1}{4}\left[\left(1 - p_i\right)e^{-\eta_i} + p_i\right]^2}{\left[\left(1 - p_i\right)e^{-\eta_i} + p_i\right]^2} = \frac{1}{4}
\end{aligned} \tag{5.5}$$

Now we know that $L(\eta_i) \leq L(0) + L'(0)\eta_i + \frac{1}{8}\eta_i^2 = \frac{1}{8}\eta^2(b_i - a_i)^2$. Here we have the bound for the last line of (20):

$$\exp\left(-\eta_i p_i + \log\left(1 - p_i + p_i e^{\eta_i}\right)\right) \leq \exp(\frac{1}{8}\eta^2(b_i - a_i)^2) \tag{5.6}$$

And back to the last line of (19) we have:

$$\begin{aligned}
Pr(\bar{\xi} - \mathbb{E}[\bar{\xi}] > \epsilon) &\leq e^{-N\eta\epsilon}\prod_{i=1}^{N}\exp(\frac{1}{8}\eta^2(b_i - a_i)^2) \\
&= exp(-N\eta\epsilon + \frac{1}{8}\eta^2\sum_{i=1}^{N}(b_i - a_i)^2)
\end{aligned} \tag{5.7}$$

Obviously, the last formula is a qudratic function of η and then we can use it to bound the whole probability. That is, the right hand side formula takes its optimal value at $\eta^* = \frac{4N\epsilon}{\sum_{i=1}^{N}(b_i - a_i)^2}$. Then, we have the Hoeffding inequality.

5.2 Proof of Symmetrization Lemma

For simplicity, we assume that the supremum on the left hand side can be achieved at $f_N \in \mathcal{F}$. Note that f_N depends on the choice of the original sample. Now,

$$\begin{aligned}
&I(E(f_N) - E_N(f_N) > \epsilon)I(E(f_N) - E_N^*(f_N) < \epsilon/2) \\
&= I(\{E(f_N) - E_N(f_N) > \epsilon\}\bigcap\{-E(f_N) + E_N^*(f_N) > -\epsilon/2\}) \\
&\leq I(E_N^*(f_N) - E_N(f_N) > \epsilon/2)
\end{aligned} \tag{5.8}$$

Then we take expectation for ghost sample[1]:

$$I(E(f_N) - E_N(f_N) > \epsilon)Pr(E(f_N) - E_N^*(f_N) < \epsilon/2) \leq Pr(E_N^*(f_N) - E_N(f_N) > \epsilon/2) \tag{5.9}$$

[1] Note that here,$E_N^*(f_N) = 1/N\sum_{i=1}^{N} f_N(x_i^*, y_i^*)$ is a random variable which depends on the choice of ghost sample.

Then we consider $Pr(E(f_N) - E_N^*(f_N) < \epsilon/2)$. By Chebyshev's Inequality,

$$\begin{aligned}
&Pr(E(f_N) - E_N^*(f_N) < \epsilon/2) \\
&\leq \frac{4Var(E_N^*(f_N))}{\epsilon^2} = \frac{Var(\frac{1}{N}\sum_{i=1}^N f_N(x_i^*, y_i^*))}{\epsilon^2} \\
&\leq \frac{4Var(\sum_{i=1}^N f_N(x_i^*, y_i^*))}{N^2\epsilon^2} \\
&\leq \frac{4Np(1-p)}{N^2\epsilon^2} \quad (\sum_{i=1}^N f_N(x_i^*, y_i^*) \sim Binominal(N, p), \text{ where } p = Pr(f_N = 0)) \\
&\leq \frac{1}{N\epsilon^2} \quad (p(1-p) \leq \frac{1}{4} \text{ for any } p \in (0,1))
\end{aligned} \tag{5.10}$$

Equivalently, we have:

$$Pr(E(f_N) - E_N^*(f_N) \geq \epsilon/2) \leq 1 - \frac{1}{N\epsilon^2} \leq \frac{1}{2} \quad (\text{From assumption, } N\epsilon^2 \geq 2) \tag{5.11}$$

Combine this conclusion with inequality (25) ann take expectation for our training sample, we have[1]:

$$Pr(E(f_N) - E_N(f_N) > \epsilon) \leq 2Pr(E_N^*(f_N) - E_N(f_N) > \epsilon/2) \tag{5.12}$$

Finally,

$$\begin{aligned}
Pr(\sup_{f\in\mathcal{F}}(E(f) - E_N(f) > \epsilon)) &= Pr((E(f_N) - E_N(f_N) > \epsilon)) \\
&\leq 2Pr(E_N^*(f_N) - E_N(f_N) > \epsilon/2) \\
&\leq 2Pr(\sup_{f\in\mathcal{F}}(E_N^*(f) - E_N(f)) > \epsilon/2)
\end{aligned} \tag{5.13}$$

5.3 Proof of Suaer's Lemma

We try to prove this lemma using mathematical induction. Obiviously, when $N = 0$ and $d_{\mathcal{H}}$[2]is arbitrary, $d_{\mathcal{H}} = 0$ and N is arbitrary, the inequality holds. Denote $\sum_{i=0}^{d_{\mathcal{H}}} C_N^i$ as $\Phi_{d_{\mathcal{H}}}(N)$. By induction we assume: $S_H(N-1) \leq \Phi_{d_{\mathcal{H}}}(N-1)$. Now, we try to show: $S_H(N) \leq \Phi_{d_{\mathcal{H}}}(N)$.

Given original sample $\{x_1, ..., x_N\}$ with size N, we try to divide our funcion space $\mathcal{H}$ into two parts $\mathcal{H}_1$ and $\mathcal{H}_2$, whose loss class is $\mathcal{F}_1$ and $\mathcal{F}_2$, respectively. Since $\mathcal{F}$ and $\mathcal{H}$ are equivalent, we use $\mathcal{F}$ to explain how we divide the funcion space. We use $D(A)$ to denote the total number of different elements in set A and use $S(A)$ to denote the set of all different elements in A [3]. Then, $\mathcal{F}_1$ statisfies $S(\{v^{(:N-1)}, v \in \mathcal{F}\}) = S(\{u^{(:N-1)}, u \in \mathcal{F}\})$ and $|\mathcal{F}_1| = D(\{u^{(:N-1)}, u \in \mathcal{F}\})$, where v, u are the N-dimensional binary vector in $\mathcal{F}_1$ and $\mathcal{F}$, $v^{(:N-1)}$ denotes the first $N-1$ dimensions in vector v.[4] And $\mathcal{F}_1 \bigcap \mathcal{F}_2 = \emptyset, \mathcal{F}_1 \bigcup \mathcal{F}_2 = \mathcal{F}$. Then, we have the following conclusion: for $\forall f \in \mathcal{F}_2$, there exsists $f^{'} \in \mathcal{F}_1$ such that $f(x_i) = f^{'}(x_i)$ for $i = 1, ..., N-1$ and $f(x_N) \neq f^{'}(x_N)$. Then, by the definition of $\mathcal{F}_1$ and $\mathcal{F}_2$, we have $|\mathcal{F}^P| = |\mathcal{F}_1^P| + |\mathcal{F}_2^P|$. Here, $\mathcal{F}_1^P = \mathcal{F}_1(x_1, ..., x_N)$.

Now, we want to show that

$$|\mathcal{F}_1^P| \leq \Phi_{d_{\mathcal{F}}}(N-1) \quad \text{and} \quad |\mathcal{F}_2^P| \leq \Phi_{d_{\mathcal{F}}-1}(N-1) \tag{5.14}$$

First, we have $\mathcal{F}_1(x_1, ..., x_{N-1}) = \mathcal{F}_1(x_1, ..., x_N)$. Since the first set is samller, it is clear that $\mathcal{F}_1(x_1, ..., x_{N-1}) \leq \mathcal{F}_1(x_1, ..., x_N)$. Another direction holds because $(x_1, ..., x_{N-1}, 0)$ and

[1] Similarly, since f_N is fixed, $E_N(f_N)$ is a random variable which depends on choice of training sample.
[2] To aviod confusion, we use $\mathcal{H}$ to denote our function space and $\mathcal{F}$ to denote our loss class.
[3] For example, if $A = \{1, 2, 3, 1, 3, 4\}$, then $D(A) = 4$. $S(A) = \{1, 2, 3, 4\}$
[4] For example, if $\mathcal{F} = \{(0,0,1), (0,1,0), (1,0,0), (1,1,0), (1,1,1), (0,1,1)\}$, then one choice for $\mathcal{F}_1$ and $\mathcal{F}_2$ is $\mathcal{F}_1 = \{(0,0,1), (0,1,0), (1,0,0), (1,1,0)\}$, $\mathcal{F}_2 = \{(1,1,1), (0,1,1)\}$

$(x_1, ..., x_{N-1}, 1)$ can not be both contained in $\mathcal{F}_1$, otherwise $|\mathcal{F}_1| > D(\{u^{(:N-1)}, u \in \mathcal{F}\})$. Also, since $\mathcal{F}_1 \subseteq \mathcal{F}$, we have $d_{\mathcal{F}_1} \leq d_{\mathcal{F}}$. Hence, we have showed $|\mathcal{F}_1^P| \leq \Phi_{d_{\mathcal{F}}}(N-1)$.

Now, we prove the second one in Equation (5.14) on the preceding page. Similarly as above, we have $\mathcal{F}_2(x_1, ..., x_{N-1}) = \mathcal{F}_2(x_1, ..., x_N)$ because if $(x_1, ..., x_{N-1}, 0)$, $(x_1, ..., x_{N-1}, 1)$ are both contained in $\mathcal{F}_2$, we get $|\mathcal{F}_1| < D(\{u^{(:N-1)}, u \in \mathcal{F}\})$. Then, we try to prove that $d_{\mathcal{F}_2} \leq d_{\mathcal{F}} - 1$. Note that if set T is shattered by $\mathcal{F}_2$, then $T \bigcup \{x_N\}$ is shattered by $\mathcal{F}$. This is because if T is shattered by $\mathcal{F}_2$, then x_N is not in T (same reason as above) and for every $f \in \mathcal{F}_2$, there is a "twin" $f^{'} \in \mathcal{F}_1$ that is indentical except x_N. Combine these two conclusions we can prove the second inequality.

Now, the lemma can be proved:

$$\begin{aligned}
|\mathcal{F}^P| &= |\mathcal{F}_1^P| + |\mathcal{F}_2^P| \\
&\leq \Phi_{d_{\mathcal{F}}}(N-1) + \Phi_{d_{\mathcal{F}}-1}(N-1) \\
&= \sum_{i=0}^{d_{\mathcal{F}}} C_{N-1}^i + \sum_{i=0}^{d_{\mathcal{F}}-1} C_{N-1}^i = 1 + \sum_{i=1}^{d_{\mathcal{F}}} C_{N-1}^i + \sum_{i=1}^{d_{\mathcal{F}}} C_{N-1}^{i-1} \qquad (5.15) \\
&= \sum_{i=0}^{d_{\mathcal{F}}} C_N^i = \Phi_{d_{\mathcal{F}}}(N) \quad (\text{Since } C_{N-1}^{i-1} + C_{N-1}^i = C_N^i)
\end{aligned}$$

References

[1] Kearns, M. J., & Vazirani, U. V. (1994). ***An introduction to Computational Learnning Theory***. Cambridge, MA, USA: MIT Press.

[2] Maria Florina Balcan. ***Machine Learning Theory***. Lecture notes at CMU.

[3] http://freemind.pluskid.org/category