Generalization Bounds

Shaily Mishra Sujit Gujar

Contents

1 Abstract

Write Abstract here

2 Introduction

Given (training) data underlying probability distribution $\mathcal{D}$, and based on desired output, i.e., what we need to predict for these input data, we return a hypothesis, a mapping from (unseen) input data to desired labels. Few applications are Recommendation Systems, Object Recognition, Speech Processing, Natural Language Processing, etc. AI systems are now in integration with day-to-day lives. On the one hand, Machine Learning uses algorithms to learn the mapping from input to output data; Deep Learning learns via neural nodes. This mapping could represent binary classification, multi-class classification, regression, ranking, clustering, or even an algorithm. The learning can be supervised, semi-supervised or unsupervised.

While training our model, we generally consider empirical training and test error. However, with the heavy application of learning models in susceptible areas, we need to look beyond empirical error. We need to have bounds on how worst our learning algorithm can perform. Since AI is advancing so progressively, such as Self-driving cars or in Health Care, it is severe that we should already know beforehand how worse our trained model will perform; otherwise, things could lead to disaster. So what we are looking for is generalization error, i.e., given a training error, what is the upper bound on unseen data? Also, note that we have an essential assumption that our data comes from an underlying distribution. We cannot train our model with training data of some distribution and then expect it to predict unseen data of some other distribution. For example, if we are trying to classify animal images, if we are training on cats and dogs, we cannot then test our model on horses and fish. We will now formally define our model.

3 Preliminaries

Consider instance space $\mathcal{X}$, each data point $x \in \mathcal{X}$ is a feature vector that represents the raw data. For example, x can represent pixels for Image Classification or Valuation profile in resource allocation. In Speech processing, we can have acoustic features, or in Email Classification, we have linguistic features. Our goal is to learn the desired output from this data, i.e., to predict whether an email is spam or not (binary classification), classify an image correctly (multi-class classification), distribute a resource among interested agents correctly, or predict a function value, x^2 (regression). We call output space as $y \in \mathcal{Y}$. Ideally we want to learn a function c, such that $y = c(x)$, $c \in \mathcal{C}$. We call $\mathcal{C}$ concept class, i.e., the function that gives the relation between x and y, mapping of x and y, is part of concept class. We assume that there exists such a function. However, we might not know the actual concept class, and we consider a hypothesis space $\mathcal{H}$, to look for a h, such that $\hat{y} = h(x)$, i.e. on any seen/unseen data x, h gives the desired output y. For example, a concept class is quadratic functions, but we are looking among linear functions, i.e., our hypothesis class is linear functions. We desire to learn an optimal $h \in \mathcal{H}$ based on our goal. We have a sample data set s, containing m sample data points, i.e. $\{data_1, data_2, \ldots, data_m\}$. Note that when we have labeled data, the sample is of form (x, y), and for unlabelled data, the sample is of the form (x). We assume that there is some probability distribution $\mathcal{D}$ over our sample data points, and these data points are, i.i.d. drawn from it. It's essential to have this assumption, as we cannot train our model to differentiate between cat and dog and then use that model to distinguish between land and rivers. Or even we cannot train our model with just one type of river and one type of land, and then use that model to differentiate over any land or river. For example, a river can be muddy somewhere, and if you don't have that in train data, you cannot use it in test data. We need to consider the probability distribution of land and river in general. Or for example, trying to predict if a research paper is well written and notable, so in our training samples, we need to consider all different formatting and templates. Now this $\mathcal{D}$ may or may not be known to us.

We define loss, i.e., the loss occurred when h predicts $\hat{y}$, then the actual value is y. We have various loss functions accordingly for our problem, i.e., binary classification, multi-class, multi-label, regression, etc. loss function $l : y \times y \rightarrow [0, \inf)$ $l(y, \hat{y})$ is the loss incurred when true label is y and the predict label is $\hat{y}$. Note that Error is expected loss. Binary Classification loss function:

$$l(y, \hat{y}) = \begin{cases} 1 & \hat{y} \neq y \\ 0 & \text{otherwise} \end{cases}$$

Regression loss function - Mean Square Error

$$l(y, \hat{y}) = (y - \hat{y})^2$$

Regression loss function - Absolute Error Loss

$$l(y, \hat{y}) = |(y - \hat{y})|$$

For multi-class classification

$$l(y, \hat{y}) = -\sum_{j=1}^{c} y_j * \log(\hat{y}_j)$$

.

4

Generalization Error

We define the empirical risk, i.e. the empirical loss, for hypothesis h, and with sample set s. as $R_{emp}(h) = 1/m \sum_{i=1}^{m} l(y, \hat{y})$. and the true error, i.e. the true loss/true risk is $R(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[l(y, \hat{y})]$. The empirical risk, is calculated over test sample drawn from distribution $\mathcal{D}$. Since we don't know $\mathcal{D}$, we cannot calculate the true risk.

- Probablity Approximately Correct (PAC). When train error is zero, $R_{emp}(h) = 0$, We want our prediction function to be $\epsilon-$ approximately correct, i.e. $|R(h)| > \epsilon$, we want the probability for this to happen with less than δ probability. $P(R_h > \epsilon) \leq \delta$
- Agnostic PAC. When train error is not zero, $R_{emp}(h) \neq 0$, We want our prediction function to be $\epsilon-$ approximately correct, i.e. $|R(h) - R_{emp}(h)| > \epsilon$, we want the probability for this to happen with less than δ probability. $P(|R(h) - R_{emp}(h)| > \epsilon) \leq \delta$

4.1 PAC Learning

We consider binary classification. With probability $(1 - \delta)$, we want to be ϵ-approximately correct. So there are m samples, and probability that each sample got predicted correctly, with no loss. R_h, i.e. the expected loss, expected risk is $> \epsilon$. so the loss is ϵ. In binary classification, the loss is 1 if $\hat{y} \neq y$, i.e. with probability p, the loss is 1, i.e. mis-classified. now $R(h)$, i.e. expected true error should not be more than ϵ, i.e. the probability that it classified m samples correctly for a hypothesis h, is $(1 - \epsilon)^m$. so the probability. $(1 - \epsilon)^m \leq e^{-\epsilon m}$. now for all $h \in \mathcal{H}$, $Pr_{h\in\mathcal{H}}(R(h) > \epsilon) \leq |\mathcal{H}|e^{-\epsilon m}$

So if the hypothesis space $\mathcal{H}$ is finite, then the probability that true error $R(h) > \epsilon$ is less than $|\mathcal{H}|e^{-\epsilon m}$.

So now if we want to ensure ϵ- approximately with δ probability, then we need $m \geq \frac{1}{\epsilon}(ln(|\mathcal{H}| + ln(1/\delta)))$

Theorem 4.1. *When $|\mathcal{H}|$ is finite, and the training error is zero, $0 \leq \epsilon \leq 1$, the probability that true error $R(h) \geq \epsilon$, is less than $|\mathcal{H}|e^{-\epsilon m}$.*

4.2 Agnostic PAC Learning

When training error $R_{emp}(h) \neq 0$, $Pr_{h\in\mathcal{H}}(|R(h) - R_{emp}(h)| \geq \epsilon) =??$. $R_{emp}(h)$ is random variable depending on the sample data set s.

Definition 1. *(Hoeffding's Inequality) : Let $X_1, X_2, \ldots, X_m$ be bounded independent random variables on $\mathbb{R}$ with $a \leq X_i \leq b$, $X = \sum_{i=1}^{m} X_i$, then $Pr(|X - \mathbb{E}(X)| > \epsilon) \leq 2e^{\frac{-2m\epsilon^2}{(b-a)^2}}$*

When $0 \leq X_i \leq 1$, $Pr(|X - \mathbb{E}(X)| > \epsilon) \leq 2e^{-2m\epsilon^2}$ For binary classification, the loss we define is 1 when misclassified and 0 when classified correctly. so in our case $0 \leq l(y, \hat{y}) \leq 1$. $R_{emp}(h) = 1/m \sum_{i=1}^{m} l(y_i, \hat{y_i})$ is a random variable of error that depends on m training examples. Hoeffding can be applied when all are independent random variables, i.e. $X_1, X_2, \ldots, X_m$ are all i.i.d.

For a hypothesis $h \in \mathcal{H}$, we know intuitively that more the samples, better the training. and we can see it from the equation.

$$\begin{aligned}
Pr(|R_{emp}(h) - R(h)| > \epsilon) &\le Pr(\max_{h\in\mathcal{H}} |R_{emp}(h) - R(h)| > \epsilon) \\
&= Pr(\cup_{h\in\mathcal{H}} |R_{emp}(h) - R(h)| > \epsilon) \\
&\le \sum_{h\in\mathcal{H}} Pr(|R_{emp}(h) - R(h)| > \epsilon) \\
&\le \sum_{h\in\mathcal{H}} 2e^{-2m\epsilon^2} \\
&= 2|\mathcal{H}|e^{-2m\epsilon^2}
\end{aligned}$$

4.2.1 For finite hypothesis class

If we want ϵ-approximately with δ probability, $2|\mathcal{H}|e^{-2m\epsilon^2} = \delta$. So with probability δ, the generalization error is $\sqrt{\frac{\ln 2/\delta + \ln|\mathcal{H}|}{2m}}$

4.2.2 For infinite hypothesis class

When $|\mathcal{H}|$ is infinite, basically we are bounding our generalization error with infinite, which gives us literally no information on generalization error. We will actually look on how big is our actually our $|\mathcal{H}|$. For consider binary classifier in R^2, if we have n points, then the maximum number of ways we can classify these n points is 2^n, so even though we have infinite classifier, we just have 2^n unique classifier. Also among all the classifier, if we restrict our hypothesis space $\mathcal{H}$ to linear classifiers, $y = w \cdot x + b$, our possibly $\mathcal{H}$ further reduces. We can classify 3 data points in R^2 any way we can, i.e., all 2^3 classification is possible to make using linear classifier. however when we have 4 data points, no matter how we arrange them we cannot achieve all 2^4 distinct classification.

4.2.3 Growth Function

We define growth function $\prod_{\mathcal{H}}(m)$, as what is the effective hypothesis space $\mathcal{H}$, i.e., $|\mathcal{H}|$ given m data points. If a hypothesis space $\mathcal{H}$ is able to produce all different labels for a set of data set, then we say it is able to scatter this set. For example, for binary classification in R^2, space of linear classifier shatters a set of 3 data points. But we also saw that it is not able to shatters a set of 4 data points.

Definition 2. *We define VC (Vapnik-Chervonenkis) dimension d as the maximum number of points that $\mathcal{H}$ can shatter, i.e. any possible orientation of these points.*

Note that any orientation of these points may not be shattered. For example, in R^2, 3 points on a x-axis, we cannot shatter it using linear classifier. however, when these 3 points are in triangular oriented, it is possible to shatter. In R^2, for linear classifier, the VC dim $d = 3$, in general R^n, for linear classifier the VC dim $d = n + 1$.

Now with VC dimension d, we can have a polynomial bound on $|\mathcal{H}|$ using Sauer's Lemma.

Theorem 4.2. *(Sauer's Lemma) If the hypothesis space $\mathcal{H}$ has VC dimension - d, so the effective hypothesis*

$$\prod_{\mathcal{H}}(m) \le \sum_{i=0}^{d} \binom{m}{i}$$

Proof: Base Case : 1) $d = 0, m \ge 1$, $\prod_{\mathcal{H}}(m) = 1$, and $\sum_{i=0}^{d} \binom{m}{i} = 1$
2) $m = 1, d \ge 1, \prod_{\mathcal{H}}(m) = 2$, and $\sum_{i=0}^{d} \binom{m}{i} = 2$
Inductive Step : $m > 1, d > 0$
Assume it is true for $m' < m, d' < d$. Consider cases : $(m-1, d-1), (m-1, d)$.
Consider a sample $s, (x_1, x_2, \dots, x_m) \in \mathcal{X}$, and $Y_1 = \{h(x_1, h(x_2), \dots, h(x_m))|h \in \mathcal{H}\}$, which is a sequence of 1 and -1. $Y_2 = \{h(x_1, h(x_2), \dots, h(x_m - 1))|h \in \mathcal{H}\}$. VC dim is d. Consider Y_3, set of all sequences in Y_2, that appear twice in Y_1.

$$|Y_1| = |Y_2| + |Y_3|$$

So $|Y_1|$ is m points with VC dim $= d$, $|Y_2|$ is $m - 1$ points with VC dim $= d$, and $|Y_3|$ is $m - 1$ points with VC dim $= d - 1$.

$$\begin{aligned}
|Y_1| &\le \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d-1} \binom{m-1}{i} \\
&= \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d} \binom{m-1}{i-1} \\
&= \sum_{i=0}^{d} \binom{m}{i}
\end{aligned}$$

□

Now,

$$\begin{aligned}
\prod_{\mathcal{H}}(m) &\le \sum_{i=0}^{d} \binom{m}{i} \\
&\le \sum_{i=0}^{d} \binom{m}{i} \Big(\frac{m}{d}\Big)^{d-i} \qquad \Big(\frac{m}{d} > 1\Big) \\
&= \Big(\frac{m}{d}\Big)^{d} \sum_{i=0}^{m} \binom{m}{i} \Big(\frac{d}{m}\Big)^{i} (1)^{m-i} \\
&= \Big(\frac{m}{d}\Big)^{d} \Big(1 + \frac{d}{m}\Big)^{m} \\
&\le \Big(\frac{m}{d}\Big)^{d} e^{\frac{d}{m}\cdot m} = \Big(\frac{em}{d}\Big)^{d}
\end{aligned}$$

$$\prod_{\mathcal{H}}(m) = O(m^d)$$

Note that Union Bound probability has the assumption that each event, i.e., $h_1, h_2 \in \mathcal{H}$, should be independent of each other.

4.2.4 Symmetrization Lemma

Can we get a better bound than this?

Theorem 4.3. *In our generalization error we look over the true error $R(h)$, we can sample a set s' i.i.d from the distribution $\mathcal{D}$, i.e. a test set/ghost set. $R'_{emp}(h)$ is the error on this ghost set. Now we can prove that,*

$$Pr(\sup_{h\in\mathcal{H}} |R_{emp}(h)-R(h)| > \epsilon) \leq 2Pr(\sup_{h\in\mathcal{H}} |R_{emp}(h)-R'_{emp}(h)| > \epsilon/2)$$

Proof:
So consider a hypothesis h_w for the worst hypothesis among $\mathcal{H}$, i.e., $\sup_{h\in\mathcal{H}} |R_{emp}(h) - R(h)|$. The event $|R_{emp}(h)-R(h)| \geq \epsilon$, and $|R(h)-R'_{emp}(h)| < \epsilon/2$, means it got generalization well for the ghost set s', but not for training set s, and we will prove that the possibility of this happening is extremely rare.

$$\begin{aligned}
&Pr(|R_{emp}(h) - R(h)| \geq \epsilon) \cdot Pr(|R'_{emp}(h) - R(h)| < \epsilon/2) \\
&= Pr(|R_{emp}(h) - R(h)| \geq \epsilon | R'_{emp}(h) - R(h)| < \epsilon/2) \\
&= Pr(|R_{emp}(h) - R(h)| \geq \epsilon | R'_{emp}(h) - R(h)| > -\epsilon/2) \\
&\leq Pr(|R_{emp}(h) - R'_{emp}(h)| \geq \epsilon/2)
\end{aligned}$$

As we already know $A \implies B$, then $P(A) \leq P(B)$.

Definition 3. *Chebyshev Inequality*

$$Pr(|X - \mathbb{E}(X)| \geq t) \leq \frac{var(X)}{t^2}$$

,

So now

$$Pr(|R(h) - R'_{emp}(h)| \geq \epsilon/2) \leq \frac{4var(R'_{emp}(h))}{\epsilon^2}$$

. Now error in our case of binary classification is total misclassified samples by total number of samples. so our error is a $1/m$ times binomial random variable, i.e. the probability of misclassified. $\mathbb{E}(R_{emp}(h)) = 1/m(mp)$, $p = \mathbb{E}(R'_{emp}(h)) = (R'_{emp}(h))$. $var(R'_{emp}(h)) = 1/m^2 \cdot mp(1-p)$.

$$Pr(|R(h) - R'_{emp}(h)| \geq \epsilon/2) \leq \frac{4var(R'_{emp}(h))}{\epsilon^2} \leq \frac{1}{\epsilon^2}$$

$$Pr(|R(h) - R'_{emp}(h)| < \epsilon/2) \geq 1 - \frac{1}{\epsilon^2}$$

$$\begin{aligned}
&Pr(|R_{emp}(h) - R(h)| \geq \epsilon) \cdot \left(1 - \frac{1}{m\epsilon^2}\right) \\
&\leq Pr(|R_{emp}(h) - R(h)| \geq \epsilon) \cdot Pr(|R'_{emp}(h) - R(h)| < \epsilon/2) \\
&\leq Pr(|R_{emp}(h) - R'_{emp}(h)| > \epsilon/2)
\end{aligned}$$

$$\begin{aligned}
&Pr(|R_{emp}(h) - R(h)| \geq \epsilon) \cdot \left(1 - \frac{1}{m\epsilon^2}\right) \\
&\leq Pr(|R_{emp}(h) - R'_{emp}(h)| > \epsilon/2)
\end{aligned}$$

Assuming $m\epsilon^2 > 2$, we get $1/(1 - \frac{1}{m\epsilon^2}) \leq 2$ Hence,

$$Pr(\sup_{h\in\mathcal{H}} |R_{emp}(h)-R(h)| > \epsilon) \leq 2Pr(\sup_{h\in\mathcal{H}} |R_{emp}(h)-R'_{emp}(h)| > \epsilon/2)$$

□

Further, extending this we get

$$\begin{aligned}
&\leq Pr(|R_{emp}(h) - R'_{emp}(h)| > \epsilon) \\
&= Pr(|R(h) - R'_{emp}(h) + R_{emp}(h) - R(h)| > \epsilon) \\
&\leq Pr(|R(h) - R'_{emp}(h) > \epsilon/2|) + Pr(|R_{emp}(h) - R(h)| > \epsilon/2|) \\
&\leq e^{-2m(\frac{\epsilon}{2})^2} + e^{-2m(\frac{\epsilon}{2})^2} = 2e^{(\frac{-2m\epsilon^2}{4})}
\end{aligned}$$

So compiling everything we get,

$$\begin{aligned}
&\leq Pr(\sup_{h\in\mathcal{H}} |R(h) - R_{emp}(h)| > \epsilon) \\
&\leq 2Pr(\sup_{h\in\mathcal{H}} |R'_{emp}(h) - R_{emp}(h)| > \epsilon/2) \\
&\leq 2Pr(\sup_{h\in\mathcal{H}} |R'_{emp}(h) - R_{emp}(h)| > \epsilon/2) \\
&\leq 2\sum_{h\in\mathcal{H}} Pr(|R'_{emp}(h) - R_{emp}(h)| > \epsilon/2) \\
&\leq 2\sum_{h\in\mathcal{H}} 2 \cdot e^{-m\epsilon^2/8} \\
&= 4\prod_{\mathcal{H}}(2m) \cdot e^{-m\epsilon^2/8}
\end{aligned}$$

So with probability $1-\delta$, we get generalized error as

$$R(h) \leq R_{emp}(h) + \sqrt{\frac{8(d(\ln(2m)+1)) + \ln(4/\delta)}{m}}$$

.

Another way of calculating generalization bound for binary classification is as follows,

4.2.5 Rademacher Averages

Rademacher Averages is the measure of the richness of a class of real-valued functions, i.e. it is used to describe the complexity of a function class $\mathcal{H}$. The intuition is kind of correlation between rademacher random variables r_i and predicted $\hat{y_i} = h(x_i)$, and we take the maximum of that, which tells about the complexity of function class $\mathcal{H}$

Definition 4. *For a class $\mathcal{H} \subseteq R^{\mathcal{X}}$, and $X_1, X_2, \ldots, X_m$ are i.i.d. drawn from $\mathcal{X}$, we assume probability distribution $\mathcal{D}$ over $\mathcal{X}$ and Rademacher random variables, i.i.d, $r_1, r_2, \ldots, r_m$, which take value +1 with probability 0.5, and -1 with probability 0.5, we define the Rademacher Complexity for a sample of size m,*

$$\mathcal{R}_m(\mathcal{H}) = \mathbb{E}_{r\in\{+1,-1\}^m}\left[\sup_{hin\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} r_i h_i(x_i)\right]$$

. And Rademacher Averages/Complexity for a class $\mathcal{H}$,

$$\mathcal{R}_{m,\mathcal{D}}(\mathcal{H}) = \mathbb{E}_{X\sim D^m}[\mathcal{R}_m(\mathcal{H})]$$

Theorem 4.4. *Consider hypothesis space $\mathcal{H}$, an input space $\mathcal{X}$,$\mathcal{Y}$ is the actual label of $\mathcal{X}$, our prediction $\hat{\mathcal{Y}}$, and the loss function $l : \mathcal{Y} \times \hat{\mathcal{Y}} \to [0, B]$ We assume probability distribution $\mathbb{D}$ over $\mathcal{X} \times \mathcal{Y}$, then with probability $1 - \delta$,*

$$R(h) \leq R_{emp}(h) + 2\mathcal{R}_{m,\mathcal{D}}(l_{\mathcal{H}}) + B\sqrt{\frac{\ln 1/\delta}{2m}}$$

For Binary Classification:

$$R(h) \leq R_{emp}(h) + 2\mathcal{R}_{m,\mathcal{D}}(l_{\mathcal{H}}) + \sqrt{\frac{\ln 1/\delta}{2m}}$$

Definition 5. *(McDiarmid's Inequality) Consider independent random variables $X_1, X_2, \ldots, X_m \in \mathcal{X}$, and a mapping $f : \mathcal{X}^m \to \mathbb{R}$, then $\forall i \in [1, m], \forall x_1, x_2, \ldots, x_n, x'_i \in \mathcal{X}$,*

$$|f(x_1, \ldots, x_i, \ldots, x_m) - f(x_1, \ldots, x'_i, \ldots, x_m)| \leq c_i$$

, then $\forall \epsilon > 0$,

$$Pr(f(x_1, x_2, \ldots, x_m) - \mathbb{E}[f(X_1, X_2, \ldots, X_m)] \geq \epsilon) \leq \exp\Big(\frac{-2\epsilon^2}{\sum_{i=1}^m c_i^2}\Big)$$

McDiarmid's Inequality reduces to Hoeffding's Inequality, when $X_i \in [a_i, b_i]$, $f = 1/m \sum_{i=1}^m X_i, c_i = \frac{b_i - a_i}{m}$

$$R(h) \leq R_{emp}(h) + 2\mathcal{R}_{m,\mathcal{D}}(l_{\mathcal{H}}) + \sqrt{\frac{\ln 1/\delta}{2m}}$$

We will using McDiarmid to prove the above theorem, $\forall \in (\mathcal{X} \times \mathcal{Y})^m, f(s) = \sup_{h\in\mathcal{H}}(R(h) - R_{emp}(h))$, then $c_i = 1/m$, then we have

$$Pr(\sup_{h\in\mathcal{H}}(R(h) - R_{emp}(h)) - \mathbb{E}[\sup_{h\in\mathcal{H}}(R(h) - R_{emp}(h))]) \leq e^{-2m\epsilon^2}$$

. Now to have $\epsilon-$approximately correct, by probablity $1 - \delta$, we need to set $\delta = e^{-2m\epsilon^2}$.

$$\sup_{h\in\mathcal{H}}(R(h) - R_{emp}(h)) - \mathbb{E}[\sup_{h\in\mathcal{H}}(R(h) - R_{emp}(h))] + \sqrt{\frac{\ln 1/\delta}{2m}}$$

,

now we have,

$$\mathbb{E}_{s\sim\mathcal{D}^m}[\sup_{h\in\mathcal{H}}(R(h) - R_{emp}(h))]$$

$$= \mathbb{E}_{s\sim\mathcal{D}^m}[\sup_{h\in\mathcal{H}}[\mathbb{E}_{(\tilde{x},\tilde{y})\sim\mathcal{D}^m} 1/m \sum_{i=1}^m l(\tilde{x_i}, \tilde{y_i}) - 1/m \sum_{i=1}^m l(x_i, y_i)]]$$

$$= \mathbb{E}_{s\sim\mathcal{D}^m}[\sup_{h\in\mathcal{H}} \mathbb{E}_{(\tilde{x},\tilde{y})\sim\mathcal{D}^m}[1/m \sum_{i=1}^m l(\tilde{x_i}, \tilde{y_i}) - 1/m \sum_{i=1}^m l(x_i, y_i)]]$$

From Jensen Inequality,

$$\mathbb{E}(g(x) \geq g(\mathbb{E}(x))$$

Here $g = \sup_{h\in\mathcal{H}} 1/m$

$$\leq \mathbb{E}_{(s,\tilde{s})\sim(\mathcal{D}^m\times\mathcal{D}^m), r\in\{+1,-1\}^m}\Big[\sup_{h\in\mathcal{H}} 1/m \sum_{i=1}^m r_i[l_h(\tilde{x_i}, \tilde{y_i}) - l_h(x_i, y_i)]\Big]$$

Taking supremum jointly must be less than or equal to individual supremum, and $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$

$$\leq \mathbb{E}_{(s,\tilde{s})\sim(\mathcal{D}^m\times\mathcal{D}^m), r\in\{+1,-1\}^m}\Big[\sup_{h\in\mathcal{H}} 1/m \sum_{i=1}^m r_i[l_h(\tilde{x_i}, \tilde{y_i})]$$

$$+ \sup_{h\in\mathcal{H}} 1/m \sum_{i=1}^m -r_i[l_h(x_i, y_i)]\Big]$$

$$= \mathbb{E}_{(s,\tilde{s})\sim(\mathcal{D}^m\times\mathcal{D}^m), r\in\{+1,-1\}^m}\Big[\sup_{h\in\mathcal{H}} 1/m \sum_{i=1}^m r_i[l_h(\tilde{x_i}, \tilde{y_i})]\Big]$$

$$+ \mathbb{E}_{(s,\tilde{s})\sim(\mathcal{D}^m\times\mathcal{D}^m), r\in\{+1,-1\}^m}\Big[\sup_{h\in\mathcal{H}} 1/m \sum_{i=1}^m -r_i[l_h(x_i, y_i)]\Big]$$

$$= \mathcal{R}_{m,\mathcal{D}}(l_h) + \mathcal{R}_m(l_{\mathcal{H}})$$

$$= 2\mathcal{R}_{m,\mathcal{D}}(l_h)$$

□

Now, for binary classification,

$$\mathcal{R}_{m,\mathcal{D}}(l_{\mathcal{H}}) = 1/2\mathcal{R}_m(\mathcal{H})$$

$$\begin{aligned}
\mathcal{R}_{m,\mathcal{D}}(l_{\mathcal{H}}) &= \mathbb{E}_{s,r}[\sup_{h\in\mathcal{H}} 1/m \sum_{i=1}^m r_i[l_h(x_i, y_i)]] \\
&= \mathbb{E}_{s,r}[\sup_{h\in\mathcal{H}} 1/m \sum_{i=1}^m r_i[\frac{1 - y_i h(x_i)}{2}]] \\
&= 1/2\mathbb{E}_{s,r}[\sup_{h\in\mathcal{H}} 1/m \sum_{i=1}^m r_i h(x_i)]] \\
&= 1/2\mathcal{R}_{m,\mathcal{D}}(\mathcal{H})
\end{aligned}$$

Theorem 4.5. *(Massart's Bound) Assume $|\mathcal{H}|$ is finite, Let $s = \{x_1, x_2, \ldots, x_m\}$, and*

$$B = \max_{h\in\mathcal{H}} \Big(\sum_{i=1}^m h^2(x_i)\Big)^{1/2}$$

, then

$$\mathcal{R}_{m,D}(\mathcal{H}) \leq \frac{B\sqrt{2\ln|\mathcal{H}|}}{m}$$

And from the growth function, we know $|\mathcal{H}| = \Big(\frac{me}{d}\Big)^d$, and we have $\mathcal{R}_{m,\mathcal{D}} \leq \frac{\sqrt{m}\sqrt{2d\ln{(em/d)}}}{m}$, plugging this, we get

$$R(h) \leq R_{emp}(h) + \sqrt{\frac{2d\ln{(em/d)}}{m}} + \sqrt{\frac{\ln 1/\delta}{2m}}$$

References