Learning Theory

Generalizability and the Bias/Variance Trade-Off

Shan-Hung Wu
shwu@cs.nthu.edu.tw

Department of Computer Science,
National Tsing Hua University, Taiwan

NetDB-ML, Spring 2014

Outline

Outline

1. **Why Learning Theory?**
   - When Does Learning Work?
   - How Well Could We Learn?
2. **Preliminaries**
3. **When Does Learning Work?**
4. **The Consistency Bound**
   - Complexity Measure Revised
   - From Consistency Bound to VC Theorem
   - $P(\sup_{g\in\mathcal{H}} |R[g] - R_{emp}[g]| > \epsilon) < \delta$ for Finite Cases
   - $P(\sup_{g\in\mathcal{H}} |R[g] - R_{emp}[g]| > \epsilon) < \delta$ for Infinite Cases
5. **Generalization Error**
6. **Proofs***
   - Proof of Hoeffding's Inequality
   - Proof of Sauer's Lemma
   - Proof of Ghost Sample Bound

Outline

1. **Why Learning Theory?**
   - When Does Learning Work?
   - How Well Could We Learn?
2. **Preliminaries**
3. **When Does Learning Work?**
4. **The Consistency Bound**
   - Complexity Measure Revised
   - From Consistency Bound to VC Theorem
   - $P(\sup_{g \in \mathcal{H}} |R[g] - R_{emp}[g]| > \epsilon) < \delta$ for Finite Cases
   - $P(\sup_{g \in \mathcal{H}} |R[g] - R_{emp}[g]| > \epsilon) < \delta$ for Infinite Cases
5. **Generalization Error**
6. **Proofs***
   - Proof of Hoeffding's Inequality
   - Proof of Sauer's Lemma
   - Proof of Ghost Sample Bound

Empirical Risk Minimization (1)

- In supervised learning, we want to obtain a hypothesis $h \in \mathcal{H} : \mathcal{I} \to \mathcal{L}$ to make predictions
  - $\mathcal{I}$ and $\mathcal{L}$ are spaces of $\boldsymbol{x}$ and r respectively
- Assuming that h is parametrized by θ, we picked θ that minimizes $emp(\theta; \mathcal{X}) = \sum_{t=1}^{N} l(g(\boldsymbol{x}^{(t)}; \theta), r^{(t)})$
  - l is the loss function
- For simplicity, here we focus on the binary classifiers and the ***0-1 loss function*** $l(h(\boldsymbol{x}^{(t)}; \theta), r^{(t)}) = 1(h(\boldsymbol{x}^{(t)}; \theta) \neq r^{(t)})$, where $1(\cdot)$ is an indicator function
  - h is a perceptron
  - Theorems shown below for other h's can be obtained similarly

Empirical Risk Minimization (2)

- Give a function $g \in \mathcal{H}$, define formally its ***empirical error/risk*** over a training dataset $\mathcal{X}$ as $R_{emp}[g] := \frac{1}{N}\sum_{t=1}^{N} l(g(\boldsymbol{x}^{(t)}), r^{(t)})$
  - R_{emp} is a functional of h
- h is obtained via ***empirical risk minimization***, i.e., $h = \arg\inf_{g\in\mathcal{H}} R_{emp}[g]$

Generalization Error

- The prediction error made by h over the instances unseen during the training process is call the ***generalization error***

$$R[h] := \int p(\boldsymbol{x}, r) l(h(\boldsymbol{x}), r) d(\boldsymbol{x}, r) = E_{\mathcal{X} \times \mathcal{L}}[l(h(\boldsymbol{x}), r)]$$

- Does a low $R_{emp}[h]$ always imply a low $R[h]$? No!
- Let

$$h(\boldsymbol{x}) = \begin{cases} r^{(t)} & \text{if } \boldsymbol{x} = \boldsymbol{x}^{(t)} \text{ for some } t \\ 1 & \text{otherwise} \end{cases},$$

  then we have $R_{emp}[h] = 0$ but high $R[h]$
  - Actually, h does not lean anything from $\mathcal{X}$

When Does Learning Work? (1)

- ***No free lunch theorem***: learning is impossible if we allow $\mathcal{H}$ to contain all functions from $\mathcal{I}$ to $\mathcal{L}$ 那麼一定有一個很複雜的function empirical error=0 但generalization error很大
  - Since h could have arbitrary shape, its values at $\boldsymbol{x}^{(t)}$'s carry no information about the values at other points
- We need to assume ***inductive bias*** that restricts the "capacity of $\mathcal{H}$," called ***model complexity***
  歸納的偏見(可想成是經驗) 限制模型不能太複雜
  - E.g., $\mathcal{H}$ as a collection of hyperplanes, polynomials of degree k, or "smooth" functions, etc.

When Does Learning Work? (2)

- Let $h^* = \arg\inf_{g \in \mathcal{H}} R[g]$ hypothesis class 裡面最好的那一個
- We say that the empirical risk minimization works (or h is ***consistent***) iff for all ϵ, $\lim_{N\to\infty} P(R[h] - R[h^*] > \epsilon\}) = 0$ 如果觀察了無限多的樣本 那就一定可以做出最好的假設

Theorem

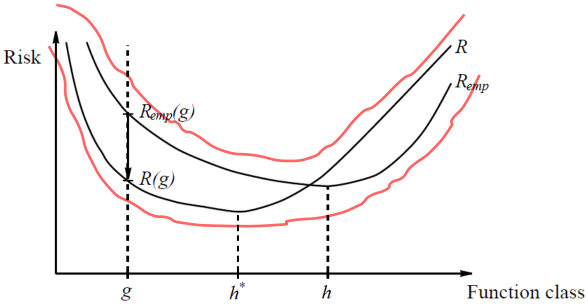
The the one-side uniform convergence

$$\lim_{N\to\infty} P(\sup_{g \in \mathcal{H}} \{R[g] - R_{emp}[g] > \epsilon\}) = 0$$

is a necessary and sufficient condition for h to be consistent.

- Here P is taken over $\mathcal{X}$, i.e., $P(\cdot) = \int_{\mathcal{X}} 1(\cdot) P(\mathcal{X}) d\mathcal{X} = E_{\mathcal{X}}[1(\cdot)]$ where $1(\cdot)$ is an indicator function
- $R[h^*]$ and $R[g]$ are constants; while $R[h]$ and $R_{emp}[g]$ depends on $\mathcal{X}$

Graphical Interpretation



- Consistency: $R[h]$ approaches to $R[h^*]$ as N grows amounts to
- (One-side) uniform convergence: the whole R_{emp} approaches to R as N grows

Outline

1. **Why Learning Theory?**
   - When Does Learning Work?
   - How Well Could We Learn?
2. **Preliminaries**
3. **When Does Learning Work?**
4. **The Consistency Bound**
   - Complexity Measure Revised
   - From Consistency Bound to VC Theorem
   - $P(\sup_{g\in\mathcal{H}}|R[g]-R_{emp}[g]|>\epsilon)<\delta$ for Finite Cases
   - $P(\sup_{g\in\mathcal{H}}|R[g]-R_{emp}[g]|>\epsilon)<\delta$ for Infinite Cases
5. **Generalization Error**
6. **Proofs***
   - Proof of Hoeffding's Inequality
   - Proof of Sauer's Lemma
   - Proof of Ghost Sample Bound

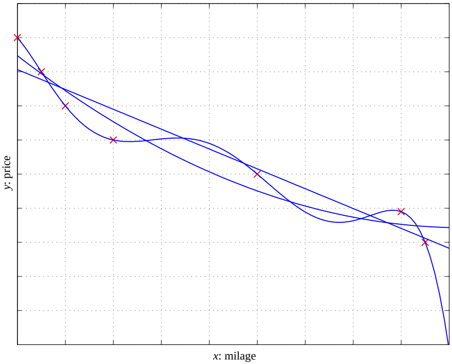
Can Learn $\neq$ Learn Well

- Given a hypothesis class $\mathcal{H}$, the above discusses whether it can learn from ***infinite*** examples 我們確保從無限大Dataset一定可以學到東西
- In practice, we have the number of examples is limited
  - Due to limited time, budget, etc.
- Does not tell how well will $\mathcal{H}$ learn with ***finite*** examples

但我們在意的是從有限大Dataset可以學得多好

Model Complexity and Bias/Variance Trade-Off

- Model complexity matters!
- Consider fitting polynomials of different degrees to examples
- A too simple a model causes ***underfitting***
  - Large ***bias***: the expected error made by a model across different training sets
  - h falls to capture the trend between $\mathcal{I}$ and $\mathcal{L}$
- A too complex a model causes ***overfitting***
  - Large ***variance***: the error made by the model due to the particularity of a training set
  - h captures not only the trend



VC Dimension

- Vapnik-Chervonenkis (VC) dimension is a measure of the capacity of a classification model

Definition (Vapnik-Chervonenkis Dimension)

We say that a hypothesis class $\mathcal{H}$ can ***shatters*** n points iff there exists a way to place the n points in the feature space such that for any of the 2^n possible labelings, we can find a hypothesis $g \in \mathcal{H}$ that separates the positive examples from negative. The maximum number of points $\mathcal{H}$ can shatter is called the ***Vapnik-Chervonenkis (VC) dimension***, denoted by $VC(\mathcal{H})$.

- What's the VC dimension of the rectangles? 4

空間中有n個點，有一個Hypothesis class H
存在一種這n個點的排列方式，使得無論這n個點是什麼label(共有2^n)種label方式，
在H中都可以找到一個g，使得g可以將這n個點依照它們的label分開
那麼H的VC dimension，記為VC(H)就至少是n

The Consistency Bound[1]

Theorem

Given $\mathcal{H}$ and $h \in \mathcal{H}$ obtained from empirical risk minimization. Then with probability at least $1-\delta$, we have that

$$\begin{aligned} R[h] &\leqslant R[h^*] + 2\sqrt{\frac{32}{N}\left(VC(\mathcal{H})\log\frac{Ne}{VC(\mathcal{H})} + \log\frac{4}{\delta}\right)} \\ &= R[h^*] + O\left(\sqrt{\frac{VC(\mathcal{H})}{N}\left(\log\frac{N}{VC(\mathcal{H})}\right) + \frac{1}{N}\log\frac{1}{\delta}}\right). \end{aligned}$$

- Bias: $R[h^*]$
- Variance: $O\left(\sqrt{\frac{VC(\mathcal{H})}{N}\left(\log\frac{N}{VC(\mathcal{H})}\right) + \frac{1}{N}\log\frac{1}{\delta}}\right)$
- Also tells how many samples N should be to learn properly

[1] This is a ***Probably Approximately Correct (PAC)*** bound of the form $P(A \leqslant \epsilon) \geqslant 1-\delta$ for some event A

Outline

The Union Bound

Lemma (Union Bound)

Let $A_1, A_2, \cdots, A_k$ be k different events (may not be independent with each other). Then

$$P(A_1 \cup A_2 \cup \cdots \cup A_k) \leqslant P(A_1) + P(A_2) + \cdots + P(A_k).$$

- $P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$

Hoeffding's Inequality

Lemma (Hoeffding's Inequality)

Let $Z_1, Z_2, \cdots, Z_n$ be n i.i.d. random variables sampled from Z. Then for any real-valued function f with values $f(Z) \in [a, b]$ and $\epsilon > 0$,

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} f(Z_i) - E[f(Z)]\right| > \epsilon\right) \leqslant 2\exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right).$$

- For any function $g \in \mathcal{H}$ that is independent with the dataset $\mathcal{X}$, we have $P(\left|R_{emp}[g] - R[g]\right| > \epsilon) \leqslant 2\exp(-2N\epsilon^2)$
  - $Z_t := 1(g(\boldsymbol{x}^{(t)}) \neq r^{(t)})$'s are i.i.d. and $f(Z) = Z \in [0, 1]$
- Applicable to $h \in \mathcal{H}$ obtained by empirical risk minimization over $\mathcal{X}$?

Hoeffding's Inequality

Lemma (Hoeffding's Inequality)

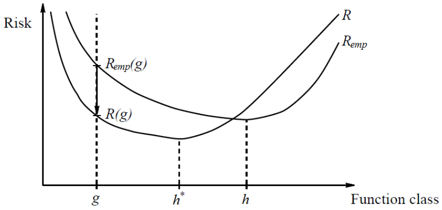
Let $Z_1, Z_2, \cdots, Z_n$ be n i.i.d. random variables sampled from Z. Then for any real-valued function f with values $f(Z) \in [a, b]$ and $\epsilon > 0$,

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} f(Z_i) - E[f(Z)]\right| > \epsilon\right) \leqslant 2\exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right).$$

- For any function $g \in \mathcal{H}$ that is independent with the dataset $\mathcal{X}$, we have $P(\left|R_{emp}[g] - R[g]\right| > \epsilon) \leqslant 2\exp(-2N\epsilon^2)$
  - $Z_t := 1(g(\boldsymbol{x}^{(t)}) \neq r^{(t)})$'s are i.i.d. and $f(Z) = Z \in [0, 1]$
- Applicable to $h \in \mathcal{H}$ obtained by empirical risk minimization over $\mathcal{X}$? ***No!***

Limitations

- The Hoeffding's Inequality tells us that, for any ***fixed*** function that does not change with $\mathcal{X}$, $R_{emp}[g]$ approaches to $R[g]$ as N grows to infinity
- h^* is fixed, but h is not
  - As R_{emp} changes with $\mathcal{X}$
- So, Hoeffding's Inequality does ***not*** imply $P(\left|R_{emp}[h]-R[h]\right|>\epsilon)\leqslant 2\exp(-2N\epsilon^2)$
  - In particular, $\lim_{N\to\infty}P(R_{emp}[h]-R[h]>\epsilon)\neq 0$



Outline

From Consistency to Uniform Convergence (1)

- We want to prove

$$\lim_{N\to\infty} P(R[h]-R[h^*] > \epsilon\}) = 0,$$

iff

$$\lim_{N\to\infty} P(\sup_{g\in\mathcal{H}}\{R[g]-R_{emp}[g] > \epsilon\}) = 0$$

where $h^* = \arg\inf_{g\in\mathcal{H}} R[g]$

- By definition, we have $R[h]-R[h^*] \geqslant 0$ and $R_{emp}[h^*]-R_{emp}[h] \geqslant 0$
- Therefore, $0 \leqslant R[h]-R[h^*]+R_{emp}[h^*]-R_{emp}[h]$

From Consistency to Uniform Convergence (2)

- We have

$$
\begin{aligned}
0 &\leqslant R[h]-R[h^*]+R_{emp}[h^*]-R_{emp}[h] \\
&= (R[h]-R_{emp}[h])+(R_{emp}[h^*]-R[h^*]) \\
&\leqslant \sup_{g\in\mathcal{H}}(R[g]-R_{emp}[g])+(R_{emp}[h^*]-R[h^*])
\end{aligned}
$$

- By Hoeffding's Inequality, we know that $\lim_{N\to\infty} P(R_{emp}[h^*]-R[h^*]>\epsilon)=0$
  - h^* is independent with $\mathcal{X}$
- So, $\lim_{N\to\infty} P(R[h]-R[h^*]+R_{emp}[h^*]-R_{emp}[h]>\epsilon)=0$ (and therefore $\lim_{N\to\infty} P(R[h]-R[h^*]>\epsilon)=0$) iff $\lim_{N\to\infty} P(\sup_{g\in\mathcal{H}}(R[g]-R_{emp}[g])>\epsilon)=0$

Outline

1. Why Learning Theory?
   - When Does Learning Work?
   - How Well Could We Learn?
2. Preliminaries
3. When Does Learning Work?
4. The Consistency Bound
   - Complexity Measure Revised
   - From Consistency Bound to VC Theorem
   - $P(\sup_{g\in\mathcal{H}}|R[g]-R_{emp}[g]|>\epsilon)<\delta$ for Finite Cases
   - $P(\sup_{g\in\mathcal{H}}|R[g]-R_{emp}[g]|>\epsilon)<\delta$ for Infinite Cases
5. Generalization Error
6. Proofs*
   - Proof of Hoeffding's Inequality
   - Proof of Sauer's Lemma
   - Proof of Ghost Sample Bound

Outline

Shattering Coefficient

- Recall that the VC dimension is a complexity measure of a hypothesis class $\mathcal{H}$ we assumed
  - Independent with $\mathcal{X}$; no need to know $\mathcal{X}$ before making the assumption
- However, during our analysis we need other complexity measures that are dependent with $\mathcal{X}$
- Given $\mathcal{X}$ and $\mathcal{H}$, define the ***shattering coefficient*** of $\mathcal{H}$ as

$$\mathcal{N}(\mathcal{H},\mathcal{X}) := |\{(g(\boldsymbol{x}^{(1)}),\cdots,g(\boldsymbol{x}^{(N)})) \in \{0,1\}^N : g \in \mathcal{H}\}|$$

  - Measures the number of sequences $(g(\boldsymbol{x}^{(1)}),\cdots,g(\boldsymbol{x}^{(N)}))$'s that the functions in $\mathcal{H}$ can give

Growth Function

- Define the ***growth function*** as

$$\mathcal{S}(\mathcal{H},n) := \sup_{(\boldsymbol{x}^{(1)} r^{(1)}),\cdots,(\boldsymbol{x}^{(n)},r^{(n)})\in\mathcal{I}\times\mathcal{L}} \mathcal{N}(\mathcal{H},(\boldsymbol{x}^{(1)} r^{(1)}),\cdots,(\boldsymbol{x}^{(n)},r^{(n)}))$$

- If $\mathcal{H}$ shatters n points, then $\mathcal{S}(\mathcal{H},n) = 2^n$
  - $VC(\mathcal{H})$ is the largest n such that $\mathcal{S}(\mathcal{H},n) = 2^n$

Sauer's Lemma

Theorem (Sauer's Lemma)

$\mathcal{S}(\mathcal{H}, n) \leqslant \sum_{k=0}^{VC(\mathcal{H})} \binom{n}{k}$.

Corollary

$\mathcal{S}(\mathcal{H}, n) \leqslant \left(\frac{ne}{VC(\mathcal{H})}\right)^{VC(\mathcal{H})}$ *[Homework]*

Outline

The Consistency Bound

Theorem

Given $\mathcal{H}$ and $h \in \mathcal{H}$ obtained from empirical risk minimization based on the 0-1 loss function. Then with probability at least $1-\delta$, we have that

$$\begin{aligned} R[h] &\leqslant R[h^*] + 2\sqrt{\frac{32}{N}\left(\log \mathcal{S}(\mathcal{H},N) + \log\frac{8}{\delta}\right)} \\ &\leqslant R[h^*] + 2\sqrt{\frac{32}{N}\left(VC(\mathcal{H})\log\frac{Ne}{VC(\mathcal{H})} + \log\frac{4}{\delta}\right)} \end{aligned}$$

- Although looser, the VC-dimension version is more "user friendly"
- Note that if we have $P(\sup_{g\in\mathcal{H}}|R[g]-R_{emp}[g]| \leqslant \epsilon) \geqslant 1-\delta$, then, with probability at least $1-\delta$,

$$\begin{aligned} R[h] &\leqslant R_{emp}[h] + \epsilon \\ &\leqslant R_{emp}[h^*] + \epsilon \\ &\leqslant R[h^*] + 2\epsilon \end{aligned}$$

VC Theorem

Theorem

Given $\mathcal{H}$ and $h \in \mathcal{H}$ obtained from empirical risk minimization based on the 0-1 loss function. We have

$$P(\sup_{g \in \mathcal{H}} |R[g] - R_{emp}[g]| > \epsilon) < 8\mathcal{S}(\mathcal{H}, N)\exp(-N\epsilon^2/32).$$

- Given a fixed δ, what is ϵ such that $P(\sup_{g \in \mathcal{H}} |R[g] - R_{emp}[g]| \leqslant \epsilon) \geqslant 1 - \delta$?

VC Theorem

Theorem

Given $\mathcal{H}$ and $h \in \mathcal{H}$ obtained from empirical risk minimization based on the 0-1 loss function. We have

$$P(\sup_{g \in \mathcal{H}} |R[g] - R_{emp}[g]| > \epsilon) < 8\mathcal{S}(\mathcal{H}, N) \exp(-N\epsilon^2/32).$$

- Given a fixed δ, what is ϵ such that $P(\sup_{g \in \mathcal{H}} |R[g] - R_{emp}[g]| \leqslant \epsilon) \geqslant 1 - \delta$?
  - The above amounts to $P(\sup_{g \in \mathcal{H}} |R[g] - R_{emp}[g]| \leqslant \epsilon) \geqslant 1 - 8\mathcal{S}(\mathcal{H}, N) \exp(-N\epsilon^2/32)$
  - $\epsilon = \sqrt{\frac{32}{N}\left(\log \mathcal{S}(\mathcal{H}, N) + \log \frac{8}{\delta}\right)}$, the consistency bound holds [Proof]

Outline

$P(\sup_{g\in\mathcal{H}}|R[g]-R_{emp}[g]|>\epsilon)<\delta$ for Finite Cases

- If $|\mathcal{H}|$ is finite, we can easily obtain δ for a given ϵ
- Let $\mathcal{H}=\bigcup_{i=1}^{|\mathcal{H}|} g_i$ and A_i be the event that $|R[g_i]-R_{emp}[g_i]|>\epsilon$
- By the union bound and Hoeffding's Inequality, we have $P(\sup_{g\in\mathcal{H}}|R[g]-R_{emp}[g]|>\epsilon)=P(\exists g\in\mathcal{H},\,|R[g]-R_{emp}[g]|>\epsilon)=$ $P(A_1\cup\cdots\cup A_{|\mathcal{H}|})\leqslant\sum_{i=1}^{|\mathcal{H}|}P(A_i)\leqslant\sum_{i=1}^{|\mathcal{H}|}2\exp(-2N\epsilon^2)=$ $2|\mathcal{H}|\exp(-2N\epsilon^2)$

$P(\sup_{g\in\mathcal{H}}|R[g]-R_{emp}[g]|>\epsilon)<\delta$ for Finite Cases

- If $|\mathcal{H}|$ is finite, we can easily obtain δ for a given ϵ
- Let $\mathcal{H}=\bigcup_{i=1}^{|\mathcal{H}|} g_i$ and A_i be the event that $|R[g_i]-R_{emp}[g_i]|>\epsilon$
- By the union bound and Hoeffding's Inequality, we have $P(\sup_{g\in\mathcal{H}}|R[g]-R_{emp}[g]|>\epsilon)=P(\exists g\in\mathcal{H}, |R[g]-R_{emp}[g]|>\epsilon)=$ $P(A_1\cup\cdots\cup A_{|\mathcal{H}|})\leqslant\sum_{i=1}^{|\mathcal{H}|}P(A_i)\leqslant\sum_{i=1}^{|\mathcal{H}|}2\exp(-2N\epsilon^2)=$ $2|\mathcal{H}|\exp(-2N\epsilon^2)$
- Unfortunately, $|\mathcal{H}|$ is infinite!
  - Here $\mathcal{H}$ denotes the collection of binary functions
- But if we can partition $\mathcal{H}$ into finite groups $\mathcal{H}=\mathcal{H}_1\cup\cdots$ such that each $P(\sup_{g\in\mathcal{H}_i}|R[g]-R_{emp}[g]|>\epsilon)$ can be bounded the same by Hoeffding's inequality, then we can still apply the union bound

Outline

Step 1: Symmetrization by Ghost Samples (1)

Theorem (Ghost Sample Bound)

For $N\epsilon^2 \geqslant 2$, we have

$$P(\sup_{g\in\mathcal{H}} |R[g]-R_{emp}[g]| > \epsilon) \leqslant 2P(\sup_{g\in\mathcal{H}} |R_{emp}[g]-R'_{emp}[g]| > \epsilon/2),$$

*where $R'_{emp}[g]$ is the empirical risk of g over another dataset consisting of N i.i.d. **ghost samples**.*

- We have

$$\begin{aligned}
&P(\sup_{g\in\mathcal{H}} |R_{emp}[g]-R'_{emp}[g]| > \tfrac{\epsilon}{2}) = P(\exists g, |R_{emp}[g]-R'_{emp}[g]| > \tfrac{\epsilon}{2}) \\
&= P(\exists g, \tfrac{1}{N}|\textstyle\sum_{t=1}^{N} 1(g(\boldsymbol{x}^{(t)}) \neq r^{(t)}) - 1(g(\boldsymbol{x}^{(t)'}) \neq r^{(t)'})| > \tfrac{\epsilon}{2}) \\
&\leqslant P(\exists g, \tfrac{1}{N}|\textstyle\sum_{t=1}^{N} 1(g(\boldsymbol{x}^{(t)}) \neq r^{(t)})| > \tfrac{\epsilon}{4} \\
&\qquad \text{or } \exists g, \tfrac{1}{N}|\textstyle\sum_{t=1}^{N} 1(g(\boldsymbol{x}^{(t)'}) \neq r^{(t)'})| > \tfrac{\epsilon}{4}) \\
&\leqslant 2P(\exists g, \tfrac{1}{N}|\textstyle\sum_{t=1}^{N} 1(g(\boldsymbol{x}^{(t)}) \neq r^{(t)})| > \tfrac{\epsilon}{4}) \\
&= 2P(\sup_{g\in\mathcal{H}} \tfrac{1}{N}|\textstyle\sum_{t=1}^{N} 1(g(\boldsymbol{x}^{(t)}) \neq r^{(t)})| > \tfrac{\epsilon}{4})
\end{aligned}$$

Step 1: Symmetrization by Ghost Samples (2)

- So far, we have $P(\sup_{g\in\mathcal{H}}|R[g]-R_{emp}[g]|>\epsilon)\leqslant 4P(\sup_{g\in\mathcal{H}}\frac{1}{N}|\sum_{t=1}^{N}1(g(\boldsymbol{x}^{(t)})\neq r^{(t)})|>\frac{\epsilon}{4})$
- Recall that, given $\mathcal{H}$, the sequence $(g(\boldsymbol{x}^{(1)}),\cdots,g(\boldsymbol{x}^{(N)}))$ can take at most $\mathcal{S}(\mathcal{H},N)$ values
- The sequence $(1(g(\boldsymbol{x}^{(1)})\neq r^{(1)}),\cdots,1(g(\boldsymbol{x}^{(N)})\neq r^{(N)}))$ can take at most $\mathcal{S}(\mathcal{H},N)$ values
- Let $\mathcal{H}_{\mathcal{X}}\subseteq\mathcal{H}$ be the smallest subset of $\mathcal{H}$ that gives rise of all the sequences $(1(g(\boldsymbol{x}^{(1)})\neq r^{(1)}),\cdots,1(g(\boldsymbol{x}^{(N)})\neq r^{(N)}))$'s, we have $P(\sup_{g\in\mathcal{H}}\frac{1}{N}|\sum_{t=1}^{N}1(g(\boldsymbol{x}^{(t)})\neq r^{(t)})|>\frac{\epsilon}{4})=$ $P(\max_{g\in\mathcal{H}_{\mathcal{X}}}\frac{1}{N}|\sum_{t=1}^{N}1(g(\boldsymbol{x}^{(t)})\neq r^{(t)})|>\frac{\epsilon}{4})$
- Note that $|\mathcal{H}_{\mathcal{X}}|\leqslant\mathcal{S}(\mathcal{H},N)$
- Since $\mathcal{H}_{\mathcal{X}}$ is finite, we are ready to apply the union bound + Hoeffding's Inequality ☺

Step 2: Union Bound

$$
\begin{aligned}
& P(\sup_{g\in\mathcal{H}} \tfrac{1}{N}|\textstyle\sum_{t=1}^{N} 1(g(\boldsymbol{x}^{(t)}) \neq r^{(t)})| > \tfrac{\epsilon}{4}) \\
& = P(\max_{g\in\mathcal{H}_{\mathcal{X}}} \tfrac{1}{N}|\textstyle\sum_{t=1}^{N} 1(g(\boldsymbol{x}^{(t)}) \neq r^{(t)})| > \tfrac{\epsilon}{4}) \\
& = P(\textstyle\bigcup_{g\in\mathcal{H}_{\mathcal{X}}} \tfrac{1}{N}|\textstyle\sum_{t=1}^{N} 1(g(\boldsymbol{x}^{(t)}) \neq r^{(t)})| > \tfrac{\epsilon}{4}) \\
& \leqslant \textstyle\sum_{g\in\mathcal{H}_{\mathcal{X}}} P(\tfrac{1}{N}|\textstyle\sum_{t=1}^{N} 1(g(\boldsymbol{x}^{(t)}) \neq r^{(t)})| > \tfrac{\epsilon}{4}) \\
& \leqslant |\mathcal{H}_{\mathcal{X}}| \sup_{g\in\mathcal{H}_{\mathcal{X}}} P(\tfrac{1}{N}|\textstyle\sum_{t=1}^{N} 1(g(\boldsymbol{x}^{(t)}) \neq r^{(t)})| > \tfrac{\epsilon}{4}) \\
& \leqslant \mathcal{S}(\mathcal{H}, N) \sup_{g\in\mathcal{H}_{\mathcal{X}}} P(\tfrac{1}{N}|\textstyle\sum_{t=1}^{N} 1(g(\boldsymbol{x}^{(t)}) \neq r^{(t)})| > \tfrac{\epsilon}{4}) \\
& \leqslant \mathcal{S}(\mathcal{H}, N) \sup_{g\in\mathcal{H}} P(\tfrac{1}{N}|\textstyle\sum_{t=1}^{N} 1(g(\boldsymbol{x}^{(t)}) \neq r^{(t)})| > \tfrac{\epsilon}{4})
\end{aligned}
$$

- We "pull" the supremum outside the probability
- Then, how to apply the Hoeffding's Inequality?

Step 2: Union Bound

$$
\begin{aligned}
&P(\sup_{g\in\mathcal{H}} \tfrac{1}{N}|\textstyle\sum_{t=1}^{N} 1(g(\boldsymbol{x}^{(t)}) \neq r^{(t)})| > \tfrac{\epsilon}{4}) \\
&= P(\max_{g\in\mathcal{H}_\mathcal{X}} \tfrac{1}{N}|\textstyle\sum_{t=1}^{N} 1(g(\boldsymbol{x}^{(t)}) \neq r^{(t)})| > \tfrac{\epsilon}{4}) \\
&= P(\bigcup_{g\in\mathcal{H}_\mathcal{X}} \tfrac{1}{N}|\textstyle\sum_{t=1}^{N} 1(g(\boldsymbol{x}^{(t)}) \neq r^{(t)})| > \tfrac{\epsilon}{4}) \\
&\leqslant \textstyle\sum_{g\in\mathcal{H}_\mathcal{X}} P(\tfrac{1}{N}|\textstyle\sum_{t=1}^{N} 1(g(\boldsymbol{x}^{(t)}) \neq r^{(t)})| > \tfrac{\epsilon}{4}) \\
&\leqslant |\mathcal{H}_\mathcal{X}| \sup_{g\in\mathcal{H}_\mathcal{X}} P(\tfrac{1}{N}|\textstyle\sum_{t=1}^{N} 1(g(\boldsymbol{x}^{(t)}) \neq r^{(t)})| > \tfrac{\epsilon}{4}) \\
&\leqslant \mathcal{S}(\mathcal{H},N) \sup_{g\in\mathcal{H}_\mathcal{X}} P(\tfrac{1}{N}|\textstyle\sum_{t=1}^{N} 1(g(\boldsymbol{x}^{(t)}) \neq r^{(t)})| > \tfrac{\epsilon}{4}) \\
&\leqslant \mathcal{S}(\mathcal{H},N) \sup_{g\in\mathcal{H}} P(\tfrac{1}{N}|\textstyle\sum_{t=1}^{N} 1(g(\boldsymbol{x}^{(t)}) \neq r^{(t)})| > \tfrac{\epsilon}{4})
\end{aligned}
$$

- We “pull” the supremum outside the probability
- Then, how to apply the Hoeffding's Inequality? No you can't ☹
    - $E[1(g(\boldsymbol{x}) \neq r)]$ does not have zero mean satisfying $P(|\frac{1}{N}\sum_{t=1}^{N} 1(g(\boldsymbol{x}^{(t)}) \neq r^{(t)}) - 0| > \frac{\epsilon}{4}) \leqslant 2\exp(-\frac{2N(\epsilon/4)^2}{(1-0)^2}) = 2\exp(-\frac{N\epsilon^2}{8})$

Step 1 Revisited

- Ghost sample bound:

$$
\begin{aligned}
&P(\sup_{g\in\mathcal{H}}|R[g]-R_{emp}[g]|>\epsilon)\\
&\leqslant 2P(\sup_{g\in\mathcal{H}}|R_{emp}[g]-R'_{emp}[g]|>\epsilon/2)\\
&=2P(\sup_{g\in\mathcal{H}}\tfrac{1}{N}|\textstyle\sum_{t=1}^{N}1(g(\boldsymbol{x}^{(t)})\neq r^{(t)})-1(g(\boldsymbol{x}^{(t)'})\neq r^{(t)'})|>\epsilon/2)
\end{aligned}
$$

- $1(g(\boldsymbol{x}^{(t)})\neq r^{(t)})$ and $1(g(\boldsymbol{x}^{(t)'})\neq r^{(t)'})$ have the same distribution
- $1(g(\boldsymbol{x}^{(t)})\neq r^{(t)})-1(g(\boldsymbol{x}^{(t)'})\neq r^{(t)'})$ has zero mean and a symmetric distribution
- So, the probability won't change if we change the sign of each $1(g(\boldsymbol{x}^{(t)})\neq r^{(t)})-1(g(\boldsymbol{x}^{(t)'})\neq r^{(t)'})$

Step 1b: Symmetrization by Random Signs

- Let $\sigma^{(1)}, \cdots, \sigma^{(N)}$ be N i.i.d. random variables, independent with $\mathcal{X}$ and $\mathcal{X}'$, such that $P(\sigma^{(t)} = 1) = P(\sigma^{(t)} = -1) = 1/2$
  - These are called ***Rademacher random variables***
- Step 1 [Proof]:

$$
\begin{aligned}
&P(\sup_{g\in\mathcal{H}} |R_{emp}[g] - R'_{emp}[g]| > \tfrac{\epsilon}{2}) = P(\exists g, |R_{emp}[g] - R'_{emp}[g]| > \tfrac{\epsilon}{2}) \\
&= P(\exists g, \tfrac{1}{N}|\textstyle\sum_{t=1}^{N} 1(g(\boldsymbol{x}^{(t)}) \neq r^{(t)}) - 1(g(\boldsymbol{x}^{(t)\prime}) \neq r^{(t)\prime})| > \tfrac{\epsilon}{2}) \\
&= P(\exists g, \tfrac{1}{N}|\textstyle\sum_{t=1}^{N} \sigma^{(t)}\{1(g(\boldsymbol{x}^{(t)}) \neq r^{(t)}) - 1(g(\boldsymbol{x}^{(t)\prime}) \neq r^{(t)\prime})\}| > \tfrac{\epsilon}{2}) \\
&\vdots \\
&\leqslant 2P(\sup_{g\in\mathcal{H}} \tfrac{1}{N}|\textstyle\sum_{t=1}^{N} \sigma^{(t)} 1(g(\boldsymbol{x}^{(t)}) \neq r^{(t)})| > \tfrac{\epsilon}{4})
\end{aligned}
$$

- Step 2 [Proof]:

$$
\begin{aligned}
&P(\sup_{g\in\mathcal{H}} \tfrac{1}{N}|\textstyle\sum_{t=1}^{N} \sigma^{(t)} 1(g(\boldsymbol{x}^{(t)}) \neq r^{(t)})| > \tfrac{\epsilon}{4}) \\
&\vdots \\
&\leqslant \mathcal{S}(\mathcal{H}, N) \sup_{g\in\mathcal{H}} P(\tfrac{1}{N}|\textstyle\sum_{t=1}^{N} \sigma^{(t)} 1(g(\boldsymbol{x}^{(t)}) \neq r^{(t)})| > \tfrac{\epsilon}{4})
\end{aligned}
$$

Step 3: Hoeffding's Inequality (1)

- So far, we have $P(\sup_{g\in\mathcal{H}}|R[g]-R_{emp}[g]|>\epsilon)\leqslant 4\mathcal{S}(\mathcal{H},N)\sup_{g\in\mathcal{H}}P(\frac{1}{N}|\sum_{t=1}^{N}\sigma^{(t)}1(g(\boldsymbol{x}^{(t)})\neq r^{(t)})|>\frac{\epsilon}{4})$
- Note that in $P(\frac{1}{N}|\sum_{t=1}^{N}\sigma^{(t)}1(g(\boldsymbol{x}^{(t)})\neq r^{(t)})|>\frac{\epsilon}{4})$ both $\mathcal{X}$ and $\sigma^{(t)}$'s are random variables
- ***Conditioning on*** $\mathcal{X}$ (i.e., treating $(\boldsymbol{x}^{(t)},r^{(t)})$'s as constants), we can apply Hoeffding's inequality: $P(\frac{1}{N}|\sum_{t=1}^{N}\sigma^{(t)}1(g(\boldsymbol{x}^{(t)})\neq r^{(t)})|>\frac{\epsilon}{4})\leqslant 2\exp(-\frac{2N(\epsilon/4)^2}{(1-(-1))^2})=2\exp(-\frac{N\epsilon^2}{32})$
  - $\sigma^{(t)}1(g(\boldsymbol{x}^{(t)})\neq r^{(t)})$'s are i.i.d. and have zero mean

Step 3: Hoeffding's Inequality (2)

- We have the conditioned version of step 2 and step 3:

$$
\begin{aligned}
& P(\sup_{g\in\mathcal{H}}|R[g]-R_{emp}[g]|>\epsilon) \\
& =\int_{\mathcal{X},\boldsymbol{\sigma}} 1(\sup_{g\in\mathcal{H}}\tfrac{1}{N}|\sum_{t=1}^{N}\sigma^{(t)}1(g(\boldsymbol{x}^{(t)})\neq r^{(t)})|>\tfrac{\epsilon}{4})p(\mathcal{X},\boldsymbol{\sigma})d(\mathcal{X},\boldsymbol{\sigma}) \\
& =\int_{\mathcal{X},\boldsymbol{\sigma}} 1(\sup_{g\in\mathcal{H}}\tfrac{1}{N}|\sum_{t=1}^{N}\sigma^{(t)}1(g(\boldsymbol{x}^{(t)})\neq r^{(t)})|>\tfrac{\epsilon}{4})P(\boldsymbol{\sigma}|\mathcal{X})p(\mathcal{X})d(\mathcal{X},\boldsymbol{\sigma}) \\
& =\int_{\mathcal{X}} P_{\boldsymbol{\sigma}|\mathcal{X}}(\sup_{g\in\mathcal{H}}\tfrac{1}{N}|\sum_{t=1}^{N}\sigma^{(t)}1(g(\boldsymbol{x}^{(t)})\neq r^{(t)})|>\tfrac{\epsilon}{4})p(\mathcal{X})d\mathcal{X} \\
& =\int_{\mathcal{X}} \mathcal{S}(\mathcal{H},N)\sup_{g\in\mathcal{H}}P_{\boldsymbol{\sigma}|\mathcal{X}}(\tfrac{1}{N}|\sum_{t=1}^{N}\sigma^{(t)}1(g(\boldsymbol{x}^{(t)})\neq r^{(t)})|>\tfrac{\epsilon}{4})p(\mathcal{X})d\mathcal{X} \\
& \qquad \text{(see old step 2)} \\
& \leqslant \int_{\mathcal{X}} \mathcal{S}(\mathcal{H},N)\sup_{g\in\mathcal{H}}2\exp(-\tfrac{N\epsilon^2}{32})p(\mathcal{X})d\mathcal{X} \\
& =\mathcal{S}(\mathcal{H},N)\sup_{g\in\mathcal{H}}2\exp(-\tfrac{N\epsilon^2}{32})=2\mathcal{S}(\mathcal{H},N)\exp(-\tfrac{N\epsilon^2}{32})
\end{aligned}
$$

- Finally, $P(\sup_{g\in\mathcal{H}}|R[g]-R_{emp}[g]|>\epsilon)\leqslant 8\mathcal{S}(\mathcal{H},N)\exp(-N\epsilon^2/32)$

Remark

- To learn something, we have to make hypothesis $\mathcal{H}$
- To learn well in the presence of finite examples, we need to pick $\mathcal{H}$ with right complexity to prevent both underfitting and overfitting
  - Note that the consistency bound (and the bias/variance trade-off) holds for ***any*** distribution of $f(Z)$
  - There are similar bounds for classifiers/regressors other than perceptron

Outline

Generalization Error Revisited (1)

- Assuming a hypothesis class $\mathcal{H}$, let $h \in \mathcal{H}$ be the hypothesis trained from the dataset $\mathcal{X} = \{(\boldsymbol{x}^{(t)}, r^{(t)})\}_{t=1}^{N}$ by minimizing the empirical error: $R_{emp}[h] := \frac{1}{N}\sum_{t=1}^{N} l(h(\boldsymbol{x}^{(t)}), r^{(t)})$
  - l is the loss function
- Generalization error of h: $R[h] := \int p(\boldsymbol{x}, r) l(h(\boldsymbol{x}), r) d(\boldsymbol{x}, r) = E_{\mathcal{I}\times\mathcal{L}}[l(h(\boldsymbol{x}), r)]$
- Let $h^* := \arg\inf_{g\in\mathcal{H}} R[g]$
- The consistency bound

$$R[h] \leqslant R[h^*] + O\left(\sqrt{\frac{VC(\mathcal{H})}{N}\left(\log\frac{N}{VC(\mathcal{H})}\right) + \frac{1}{N}\log\frac{1}{\delta}}\right)$$

tells us that as long as $VC(\mathcal{H})$ is finite, we have $R[h] \to R[h^*]$ as $N \to \infty$

Generalization Error Revisited (2)

- Let $R^* := \inf_{f:\mathcal{I}\to\mathcal{L}} R[f]$, the "true" function that generates $\mathcal{X}$
- Consistency $R[h] \to R[h^*]$ is ***not*** our ultimate goal
- Instead, our ultimate goal is to have $R[h] \to R^*$!
- $R[h] - R^* = (R[h^*] - R^*) + (R[h] - R[h^*])$
  - $R[h^*] - R^*$ is called the ***approximation error***
  - $R[h] - R[h^*]$ is called the ***estimation error***

Components of Generalization Error

- ***Approximation error*** $(R[h^*] - R^*)$:
  - Measures how well the "true" function can be approximated by the best function in our hypothesis class $\mathcal{H}$
  - Corresponds to the ***bias*** from the statistics point of view
- ***Estimation error*** $(R[h] - R[h^*])$:
  - Measures how accurately we can determine the best function implementable by our learning system using a finite training set instead of the unseen testing examples
  - Corresponds to the ***variance*** from the statistics point of view
- ***Optimization error***
  - Error introduced when solving the objective (e.g., using numeric methods)
  - Measures how precisely we can compute the function, given limited resources such as CPU and/or memory

Trade-Offs

- Too simple a model $\mathcal{H}$ causes large approximation error and ***underfitting***
  - h falls to capture the trend between $\mathcal{I}$ and $\mathcal{L}$
- Too complex a model $\mathcal{H}$ causes large estimation error and ***overfitting***
  - h captures not only the trend but some spurious patterns (e.g., noise) local to a particular $\mathcal{X}$
- The right complexity can be determined using ***model selection*** techniques, to be discussed later
- Estimation error is also determined by #training examples
- Optimization error can be reduced at the cost of computation time (e.g, more iterations)

Small-Scale vs. Large-Scale Learning

- In practice, we have budget for a given problem
  - Number of examples, computation time, memory, etc.
- Small-scale learning problems:
  - Constrained by #training examples
  - Generalization error dominated by the approximation and estimation errors
  - Optimization error insignificant since the computation time not limited
- Large-scale learning problems:
  - Constrained by computation time
  - Besides adjusting the approximation capacity of the family of function, one can also adjust #training examples
  - Example sampling or approximate optimization (e.g. early-termination)?

Small-Scale vs. Large-Scale Learning

- In practice, we have budget for a given problem
  - Number of examples, computation time, memory, etc.
- Small-scale learning problems:
  - Constrained by #training examples
  - Generalization error dominated by the approximation and estimation errors
  - Optimization error insignificant since the computation time not limited
- Large-scale learning problems:
  - Constrained by computation time
  - Besides adjusting the approximation capacity of the family of function, one can also adjust #training examples
  - Example sampling or approximate optimization (e.g. early-termination)? Always ***try the latter first*** because optimization error usually decreases exponentially (or at least faster) with time

Outline

Outline

Hoeffding's Inequality

Lemma (Hoeffding's Inequality)

Let $Z_1, Z_2, \cdots, Z_n$ be n i.i.d. random variables sampled from Z. Then for any real-valued function f with values $f(Z) \in [a,b]$ and $\epsilon > 0$,

$$P(\left|\frac{1}{n}\sum_{i=1}^{n} f(Z_i) - E[f(Z)]\right| > \epsilon) \leqslant 2\exp(-\frac{2n\epsilon^2}{(b-a)^2}).$$

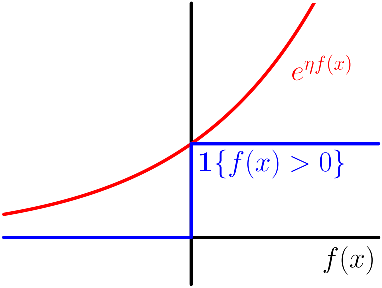
- We show that

$$P(\frac{1}{n}\sum_{i=1}^{n} f(Z_i) - E[f(Z)] > \epsilon) \leqslant \exp(-\frac{2n\epsilon^2}{(b-a)^2}),$$

which implies $P(\left|\frac{1}{n}\sum_{i=1}^{n} f(Z_i) - E[f(Z)]\right| > \epsilon) =$
$P(\frac{1}{n}\sum_{i=1}^{n} f(Z_i) - E[f(Z)] > \epsilon) + P(-\frac{1}{n}\sum_{i=1}^{n} f(Z_i) + E[f(Z)] > \epsilon) =$
$P(\frac{1}{n}\sum_{i=1}^{n} f(Z_i) - E[f(Z)] > \epsilon) + P(\frac{1}{n}\sum_{i=1}^{n} -f(Z_i) - E[-f(Z)] > \epsilon) \leqslant$
$2\exp(-\frac{2n\epsilon^2}{(b-a)^2}) + 2\exp(-\frac{2n\epsilon^2}{(-a+b)^2}) = 2\exp(-\frac{2n\epsilon^2}{(b-a)^2})$

Proof (1)

- $P(\frac{1}{n}\sum_{i=1}^{n} f(Z_i) - E[f(Z)] > \epsilon) = P(\sum_{i=1}^{n} f(Z_i) - nE[f(Z)] > n\epsilon) = E_Z[1(\sum_{i=1}^{n} f(Z_i) - nE[f(Z)] > n\epsilon)]$
- Note that for any positive real number η, we have $E_Z[1(\sum_{i=1}^{n} f(Z_i) - nE[f(Z)] - n\epsilon > 0)] \leqslant E_Z[\exp(\eta(\sum_{i=1}^{n} f(Z_i) - nE[f(Z)] - n\epsilon))] = e^{-\eta n\epsilon} \prod_{i=1}^{n} E[\exp(\eta(f(Z_i) - E[f(Z)]))]$



Proof (2)

Lemma

Given a random variable Z and a real-valued function f with values $f(Z) \in [a,b]$, for any real number η, we have

$$E[e^{\eta f(Z)}] \leqslant \frac{b - E[f(Z)]}{b-a} e^{\eta a} + \frac{E[f(Z)] - a}{b-a} e^{\eta b}.$$

Proof.

[Homework] □

Proof (3)

- From above we have $E[\exp(\eta(f(Z_i)-E[f(Z)]))] = e^{-\eta E[f(Z)]}E[\exp(\eta f(Z_i))] \leqslant e^{-\eta E[f(Z)]}\left(\frac{b-E[f(Z_i)]}{b-a}e^{\eta a}+\frac{E[f(Z_i)]-a}{b-a}e^{\eta b}\right) = e^{-\eta(E[f(Z)]-a)}\left(1-\frac{E[f(Z_i)]-a}{b-a}+\frac{E[f(Z_i)]-a}{b-a}e^{\eta(b-a)}\right) = \exp(-\eta(b-a)p_i)\cdot \exp\log(1-p_i+p_ie^{\eta(b-a)}) = \exp(-\kappa p_i+\log(1-p_i+p_ie^{\kappa}))$
- Let $L(\kappa) = -\kappa p_i+\log(1-p_i+p_ie^{\kappa})$. By Taylor's theorem, we can expend it at 0 and get $L(\kappa) = L(0)+L'(0)\kappa+\frac{1}{2}L''(\zeta)\kappa^2$, where $\zeta \in (0,\kappa)$, $L'(\kappa) = -p_i+\frac{p_ie^{\kappa}}{1-p_i+p_ie^{\kappa}} = -p_i+\frac{p_i}{(1-p_i)e^{-\kappa}+p_i}$, and $L''(\kappa) = \frac{p_i(1-p_i)e^{-\kappa}}{((1-p_i)e^{-\kappa}+p_i)^2}$ [Proof]
- By inequality of arithmetic and geometric means, we have $L''(\kappa) = \frac{p_i(1-p_i)e^{-\kappa}}{((1-p_i)e^{-\kappa}+p_i)^2} = \frac{\left(\sqrt{p_i(1-p_i)e^{-\kappa}}\right)^2}{((1-p_i)e^{-\kappa}+p_i)^2} \leqslant \frac{\left(\frac{p_i+(1-p_i)e^{-\kappa}}{2}\right)^2}{((1-p_i)e^{-\kappa}+p_i)^2} = \frac{1}{4}$, implying $L(\kappa) \leqslant L(0)+L'(0)\kappa+\frac{1}{8}\kappa^2 = \frac{1}{8}\eta^2(b-a)^2$
- So $E[\exp(\eta(f(Z_i)-E[f(Z)]))] \leqslant e^{L(\kappa)} \leqslant e^{\frac{1}{8}\eta^2(b-a)^2}$

Proof (4)

- Now we have $P(\frac{1}{n}\sum_{i=1}^{n} f(Z_i) - E[f(Z)] > \epsilon) \leqslant e^{-\eta n\epsilon}\prod_{i=1}^{n} E[\exp(\eta(f(Z_i) - E[f(Z)]))] \leqslant e^{-\eta n\epsilon}\prod_{i=1}^{n} e^{\frac{1}{8}\eta(b-a)^2} = \exp(\frac{n}{8}(b-a)^2\eta^2 - n\epsilon\eta)$
- Note that $P(\frac{1}{n}\sum_{i=1}^{n} f(Z_i) - E[f(Z)] > \epsilon) \leqslant \exp(\frac{n}{8}(b-a)^2\eta - n\epsilon\eta)$ holds for all $\eta > 0$, so we can simply find the best η that gives the tightest bound
- How?

Proof (4)

- Now we have $P(\frac{1}{n}\sum_{i=1}^{n} f(Z_i) - E[f(Z)] > \epsilon) \leqslant e^{-\eta n\epsilon}\prod_{i=1}^{n} E[\exp(\eta(f(Z_i) - E[f(Z)]))] \leqslant e^{-\eta n\epsilon}\prod_{i=1}^{n} e^{\frac{1}{8}\eta(b-a)^2} = \exp(\frac{n}{8}(b-a)^2\eta^2 - n\epsilon\eta)$
- Note that $P(\frac{1}{n}\sum_{i=1}^{n} f(Z_i) - E[f(Z)] > \epsilon) \leqslant \exp(\frac{n}{8}(b-a)^2\eta - n\epsilon\eta)$ holds for all $\eta > 0$, so we can simply find the best η that gives the tightest bound
- How? $\frac{d(\frac{n}{8}(b-a)^2\eta^2 - n\epsilon\eta)}{d\eta} = 0 \Rightarrow \eta = \frac{4\epsilon}{(b-a)^2} > 0$, at which the bound $P(\frac{1}{n}\sum_{i=1}^{n} f(Z_i) - E[f(Z)] > \epsilon) \leqslant \exp(-\frac{2n\epsilon^2}{(b-a)^2})$ is the tightest

Outline

Sauer's Lemma

Theorem (Sauer's Lemma)

$\mathcal{S}(\mathcal{H},n) \leqslant \sum_{k=0}^{VC(\mathcal{H})} \binom{n}{k}$ *for all* $\mathcal{H}$ *and* n.

- The proof proceeds by induction on both $VC(\mathcal{H})$ and n
- Base case 1: $n = 0$ and $VC(\mathcal{H})$ is arbitrary
  - When $n = 0$, there can only be one subset, hence $\mathcal{S}(\mathcal{H},n) = 1 = 1 + 0 + 0 + \cdots = \sum_{k=0}^{VC(\mathcal{H})} \binom{0}{k}$
- Base case 2: $VC(\mathcal{H}) = 0$ and n is arbitrary
  - When $VC(\mathcal{H}) = 0$, no set of points can be shattered, hence all points can be labeled only one way, implying that $\mathcal{S}(\mathcal{H},n) = 1 = \sum_{k=0}^{0} \binom{n}{k}$

Proof (1)

- Assume for induction that for all $\mathcal{H}'$ and m such that $VC(\mathcal{H}') \leqslant VC(\mathcal{H})$ and $m \leqslant n$, and at least one of these inequalities is strict, we have $\mathcal{S}(\mathcal{H}', m) \leqslant \sum_{k=0}^{VC(\mathcal{H}')} \binom{m}{k}$
- Now suppose we have a data set $\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_m\}$. Let $\mathcal{G}$ be a hypothesis class defined over $\mathcal{X}$ such that $\{(g(\boldsymbol{x}^{(1)}), \cdots, g(\boldsymbol{x}^{(m)})) \in \{0,1\}^m : g \in \mathcal{G}\} = \{(h(\boldsymbol{x}^{(1)}), \cdots, h(\boldsymbol{x}^{(m)})) : h \in \mathcal{H}\}$ and $|\mathcal{G}| = \mathcal{N}(\mathcal{H}, \mathcal{X})$
- We have $VC(\mathcal{G}) \leqslant VC(\mathcal{H})$ since any subset of $\mathcal{X}$ that is shattered by $\mathcal{G}$ is also shattered by $\mathcal{H}$

Proof (2)

- We now construct $\mathcal{G}_1$ and $\mathcal{G}_2$ as follows on which we can apply our induction hypothesis:
  - For each possible labeling of $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{m-1}\}$ induced by $\mathcal{G}$, we add a representative function from $\mathcal{G}$ to $\mathcal{G}_1$
  - Let $\mathcal{G}_2 = \mathcal{G} \backslash \mathcal{G}_1$
- So for each $g \in \mathcal{G}_2$, there exists $\tilde{g} \in \mathcal{G}_1$ such that $g(\boldsymbol{x}_i) = \tilde{g}(\boldsymbol{x}_i)$ for $i \in \{1, \cdots, m-1\}$ and $g(\boldsymbol{x}_m) \neq \tilde{g}(\boldsymbol{x}_m)$
  - For convenience, let's choose the representatives such that $g(\boldsymbol{x}_m)$ remains the same for all $g \in \mathcal{G}_2$
  - We have $\mathcal{N}(\mathcal{G}_1, \mathcal{X}) = \mathcal{N}(\mathcal{G}_1, \mathcal{X} \backslash \{\boldsymbol{x}_m\})$ and $\mathcal{N}(\mathcal{G}_2, \mathcal{X}) = \mathcal{N}(\mathcal{G}_2, \mathcal{X} \backslash \{\boldsymbol{x}_m\})$
- By construction we have $\mathcal{N}(\mathcal{H}, \mathcal{X}) = \mathcal{N}(\mathcal{G}, \mathcal{X}) = \mathcal{N}(\mathcal{G}_1, \mathcal{X}) + \mathcal{N}(\mathcal{G}_2, \mathcal{X}) = \mathcal{N}(\mathcal{G}_1, \mathcal{X} \backslash \{\boldsymbol{x}_m\}) + \mathcal{N}(\mathcal{G}_2, \mathcal{X} \backslash \{\boldsymbol{x}_m\})$

Proof (3)

- Since $\mathcal{G}_1 \subseteq \mathcal{G}$, we have $VC(\mathcal{G}_1) \leqslant VC(\mathcal{G}) \leqslant VC(\mathcal{H})$
- By induction, we obtain $\mathcal{N}(\mathcal{G}_1, \mathcal{X} \backslash \{\boldsymbol{x}_m\}) \leqslant \sum_{k=0}^{VC(\mathcal{H})} \left(\begin{array}{c} m-1 \\ k \end{array} \right)$

Proof (4)

- Note that if a dataset $\mathcal{Y}$ is shattered by $\mathcal{G}_2$ then
  - $\boldsymbol{x}_m \notin \mathcal{Y}$, since for all $g \in \mathcal{G}_2$, $g(\boldsymbol{x}_m)$ remains the same
  - In addition, $\mathcal{Y} \cup \{\boldsymbol{x}_m\}$ is shattered by $\mathcal{G}$ because each $g \in \mathcal{G}_2$ has a twin $\tilde{g} \in \mathcal{G}_1$ that is identical except on $\boldsymbol{x}_m$
- So, $VC(\mathcal{G}_2) \leqslant VC(\mathcal{G}) - 1 \leqslant VC(\mathcal{H}) - 1$ and by induction we have $\mathcal{N}(\mathcal{G}_2, \mathcal{X} \backslash \{\boldsymbol{x}_m\}) \leqslant \sum_{k=0}^{VC(\mathcal{H})-1} \binom{m-1}{k}$

Proof (5)

- Combining the above we have
  $\mathcal{N}(\mathcal{H},\mathcal{X}) = \mathcal{N}(\mathcal{G},\mathcal{X}) = \mathcal{N}(\mathcal{G}_1,\mathcal{X}) + \mathcal{N}(\mathcal{G}_2,\mathcal{X}) = \mathcal{N}(\mathcal{G}_1,\mathcal{X}\backslash\{\boldsymbol{x}_m\}) + \mathcal{N}(\mathcal{G}_2,\mathcal{X}\backslash\{\boldsymbol{x}_m\}) \leqslant \sum_{k=0}^{VC(\mathcal{H})} \begin{pmatrix} m-1 \\ k \end{pmatrix} + \sum_{k=0}^{VC(\mathcal{H})-1} \begin{pmatrix} m-1 \\ k \end{pmatrix} = \begin{pmatrix} m \\ 0 \end{pmatrix} + \sum_{k=1}^{VC(\mathcal{H})} \begin{pmatrix} m-1 \\ k \end{pmatrix} + \sum_{k=1}^{VC(\mathcal{H})} \begin{pmatrix} m-1 \\ k-1 \end{pmatrix} = \sum_{k=0}^{VC(\mathcal{H})} \begin{pmatrix} m \\ k \end{pmatrix}$

Outline

1. Why Learning Theory?
   - When Does Learning Work?
   - How Well Could We Learn?
2. Preliminaries
3. When Does Learning Work?
4. The Consistency Bound
   - Complexity Measure Revised
   - From Consistency Bound to VC Theorem
   - $P(\sup_{g \in \mathcal{H}} |R[g] - R_{emp}[g]| > \epsilon) < \delta$ for Finite Cases
   - $P(\sup_{g \in \mathcal{H}} |R[g] - R_{emp}[g]| > \epsilon) < \delta$ for Infinite Cases
5. Generalization Error
6. Proofs*
   - Proof of Hoeffding's Inequality
   - Proof of Sauer's Lemma
   - Proof of Ghost Sample Bound

Ghost Sample Bound

Theorem (Ghost Sample Bound)

For $N\epsilon^2 \geqslant 2$, we have

$$P(\sup_{g\in\mathcal{H}} |R[g] - R_{emp}[g]| > \epsilon) \leqslant 2P(\sup_{g\in\mathcal{H}} |R_{emp}[g] - R'_{emp}[g]| > \epsilon/2),$$

*where $R'_{emp}[g]$ is the empirical risk of g over another dataset consisting of N i.i.d. **ghost samples**.*

- Before going into the proof, read Appendix on Probability for Chebyshev's Inequality: $P(|X-\mu_X| \geqslant t) \leqslant \frac{\sigma_X^2}{t^2}$ for any $t > 0$

Proof (1)

- Denote $\mathcal{X}'$ the ghost sample set
- For simplicity, we assume that $\sup_{g\in\mathcal{H}}|R[g]-R_{emp}[g]|$ is attained at g^*, i.e., $P(\sup_{g\in\mathcal{H}}|R[g]-R_{emp}[g]|>\epsilon)=P(|R[g^*]-R_{emp}[g^*]|>\epsilon)$
- At r.h.s., $P(\sup_{g\in\mathcal{H}}|R_{emp}[g]-R'_{emp}[g]|>\epsilon/2)$
  $\geqslant P(|R_{emp}[g^*]-R'_{emp}[g^*]|>\epsilon/2)$
  $=\int_{\mathcal{X},\mathcal{X}'}1(|R_{emp}[g^*]-R'_{emp}[g^*]|>\epsilon/2)p(\mathcal{X}'|\mathcal{X})p(\mathcal{X})d\mathcal{X}'d\mathcal{X}$
  $\geqslant$
  $\int_{\mathcal{X},\mathcal{X}'}1(|R[g^*]-R_{emp}[g^*]|>\epsilon\wedge|R[g^*]-R'_{emp}[g^*]|<\epsilon/2)p(\mathcal{X}'|\mathcal{X})p(\mathcal{X})d\mathcal{X}'d\mathcal{X}$
  $=\int_{\mathcal{X}}1(|R[g^*]-R_{emp}[g^*]|>\epsilon)\int_{\mathcal{X}'}1(|R[g^*]-R'_{emp}[g^*]|<\epsilon/2)p(\mathcal{X}'|\mathcal{X})d\mathcal{X}'p(\mathcal{X})d\mathcal{X}$
  $=\int_{\mathcal{X}}1(|R[g^*]-R_{emp}[g^*]|>\epsilon)P_{\mathcal{X}'|\mathcal{X}}(|R[g^*]-R'_{emp}[g^*]|<\epsilon/2)p(\mathcal{X})d\mathcal{X}$
  $=\int_{\mathcal{X}}1(|R[g^*]-R_{emp}[g^*]|>\epsilon)P_{\mathcal{X}'}(|R[g^*]-R'_{emp}[g^*]|<\epsilon/2)p(\mathcal{X})d\mathcal{X}$

Proof (2)

- By Chebyshev's Inequality, we have
  $P_{\mathcal{X}'}(|R[g^*]-R'_{emp}[g^*]| \geqslant \epsilon/2) \leqslant \frac{4Var_{\mathcal{X}'}[R'_{emp}[g^*]]}{\epsilon^2} = \frac{4Var_{\mathcal{X}'}[1(g^*(\boldsymbol{x}) \neq r)]}{N\epsilon^2} \leqslant \frac{1}{N\epsilon^2}$
  - Note that $Var[1(g^*(\boldsymbol{x}) \neq r)] \leqslant \frac{1}{4}$ [Homework]
  - This amounts to $P_{\mathcal{X}'}(|R[g^*]-R'_{emp}[g^*]| < \epsilon/2) \geqslant 1 - \frac{1}{N\epsilon^2}$
- For $N\epsilon^2 \geqslant 2$, we have $P_{\mathcal{X}'}(|R[g^*]-R'_{emp}[g^*]| < \epsilon/2) \geqslant \frac{1}{2}$, implying that
  $\int_{\mathcal{X}} 1(|R[g^*]-R_{emp}[g^*]| > \epsilon) P_{\mathcal{X}'}(|R[g^*]-R'_{emp}[g^*]| < \epsilon/2) p(\mathcal{X}) d\mathcal{X} \geqslant$
  $\frac{1}{2}\int_{\mathcal{X}} 1(|R[g^*]-R_{emp}[g^*]| > \epsilon) p(\mathcal{X}) d\mathcal{X} = \frac{1}{2} P(|R[g^*]-R_{emp}[g^*]| > \epsilon)$