2.4 Hoeffding's Inequality

一个理解模型复杂性与泛化之间权衡的理论框架

Very General!

- The following discussion is **model-agnostic.** 与模型无关
- In the following discussion, we don't have assumption about **data distribution.** 与数据分布无关
- In the following discussion, we can use any **loss function.** 与loss-function无关

$$P(\mathcal{D}_{train} \text{ is } \textbf{bad}) = \bigcup_{h \in \mathcal{H}} P(\mathcal{D}_{train} \text{ is } \textbf{bad} \text{ due to } h)$$

$$\leq \sum_{h \in \mathcal{H}} P(\mathcal{D}_{train} \text{ is } \textbf{bad} \text{ due to } h)$$

$$\leq \sum_{h \in \mathcal{H}} 2exp(-2N\varepsilon^2)$$

$$= |\mathcal{H}| \cdot 2exp(-2N\varepsilon^2)$$

ε是自己决定的

How to make $P(\mathcal{D}_{train} \text{ is } bad)$ smaller?

Larger N and smaller $|\mathcal{H}|$

更大的训练资料＋更少的参数（模型复杂度）可以减少P(D_train is bad)

Hoeffding's Inequality:

$$P(\mathcal{D}_{train} \text{ is bad due to } h) \leq 2exp(-2N\varepsilon^2)$$

- The range of loss L is $[0,1]$
- N is the number of examples in $\mathcal{D}_{train}$

N是训练数据数目

$P(D_{\text{train}} \text{ is bad}) \leq |H| \times 2\exp(-2N\varepsilon^2)$

其中：

- $P(D_{\text{train}} \text{ is bad})$ 是训练数据得出不良结果的概率。
- $|H|$ 是假设空间的大小。
- N 是训练数据的数量。
- ε 是实际误差与经验误差之间的差值。

从这个不等式可以看出：

1. **当假设空间 $|H|$ 变小时**：训练数据得出不良结果的概率也会变小。这意味着，如果我们有一个简化的模型（较小的假设空间），则该模型在训练数据上表现不佳的概率会减小。但这也可能导致模型过于简单，不能捕获数据的所有复杂性（即可能欠拟合）。
2. **当训练数据的数量 N 增加时**：训练数据得出不良结果的概率也会变小。这是直观的，因为当我们有更多的数据时，我们的估计通常会更准确。

这个不等式给我们一个理论上的上界，告诉我们在最坏的情况下，坏结果发生的概率是多少。在实际应用中，这个概率通常会远小于这个上界。但这确实提供了一种理解假设空间大小和训练样本数量如何影响模型性能的方式。

假设空间就是 Model complexity

Model Complexity

$$P(\mathcal{D}_{train} \text{ is } \textbf{bad}) \leq |\mathcal{H}| \cdot 2exp(-2N\varepsilon^2)$$

Why don't we simply use a very small $|\mathcal{H}|$?

理想崩壞

"$\mathcal{D}_{train}$ is **good**" means ...

$$\forall h \in \mathcal{H}, |L(h, \mathcal{D}_{train}) - L(h, \mathcal{D}_{all})| \leq \varepsilon$$

$$L\left(h^{train}, \mathcal{D}_{all}\right) - L\left(h^{all}, \mathcal{D}_{all}\right) \leq \delta \quad \varepsilon = \delta/2$$

Larger loss

h^{all} 是所有 H 中的 h 中，可以让 loss 最小的那个。

当 |H| 很大时，可以选的 h 有限 ⇒ 找不到一个好的 h^{all}

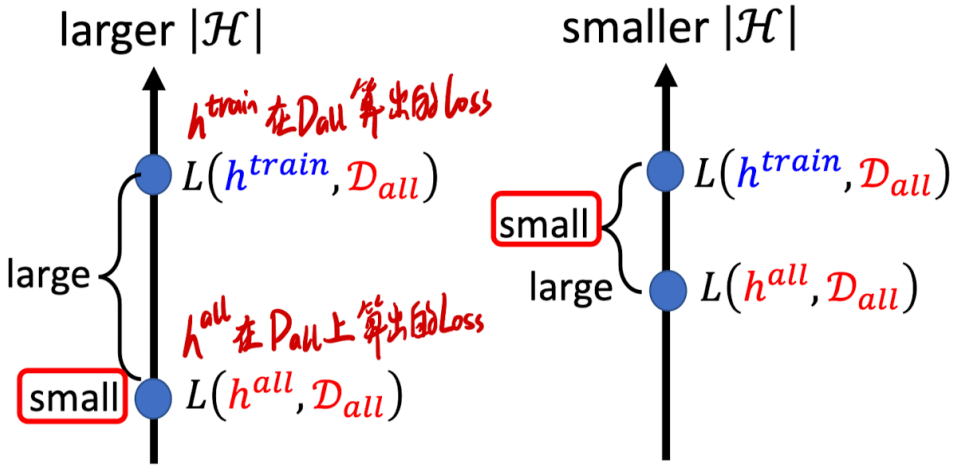
$$h^{all} = arg \min_{h \in \mathcal{H}} L(h, \mathcal{D}_{all})$$

fewer candidates

Tradeoff of Model Complexity 两难

Larger N and smaller $|\mathcal{H}|$ ➡ $L\left(h^{train}, \mathcal{D}_{all}\right) - L\left(h^{all}, \mathcal{D}_{all}\right) \leq \delta$

Larger $|\mathcal{H}|$ ➡ Larger $L\left(h^{all}, \mathcal{D}_{all}\right)$



魚與熊掌可以兼得嗎？ Yes, **Deep Learning.**