# 泛化误差定理的推广形式

## 泛化误差定理回顾

### 泛化能力

- 泛化能力指的是模型对未观测数据的预测能力
  - 可以通过泛化误差来评估，定义如下：

$$R(f) = \mathbb{E}\left[\mathcal{L}\left(Y, f(X)\right)\right] = \int_{X\times Y} \mathcal{L}\left(y, f(x)\right) p(x,y)\,dxdy$$

  - $p(x,y)$是潜在的（可能是未知的）联合数据分布
- 在训练数据集上对泛化能力的经验估计为：

$$\hat{R}(f) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}\left(y_i, f(x_i)\right)$$

### 泛化误差

- 有限假设集 $\mathcal{F} = \{f_1, f_2, \ldots, f_d\}$
- 泛化误差约束定理：

对任意函数$f \in \mathcal{F}$，以不小于$1-\delta$的概率满足下式：

$$R(f) \leq \hat{R}(f) + \epsilon(d, N, \delta)$$

其中，

$$\epsilon(d, N, \delta) = \sqrt{\frac{1}{2N}\left(\log d + \log\frac{1}{\delta}\right)}$$

- $N$：训练实例个数
- $d$：假设集的函数个数

## 引理： Hoeffding不等式

- 令$X_1, X_2, \ldots, X_n$为独立同分布的随机变量，其中$X_i \in [a, b]$，那么平均变量$Z$为：

$$Z = \frac{1}{n}\sum_{i=1}^{n} X_i$$

那么下述不等式成立：

$$P(Z - \mathbb{E}[Z] \geq t) \leq \exp\left(\frac{-2nt^2}{(b-a)^2}\right)$$

$$P(\mathbb{E}[Z] - Z \geq t) \leq \exp\left(\frac{-2nt^2}{(b-a)^2}\right)$$

## 泛化误差约束定理证明

- 假设约束的损失函数$L\big(y, f(x)\big) \in [0,1]$
- 基于Hoeffding不等式，对$\epsilon > 0$，有

$$P\big(R(f) - \hat{R}(f) \geq \epsilon\big) \leq \exp(-2N\epsilon^2)$$

- 由于$\mathcal{F} = \{f_1, f_2, \ldots, f_d\}$为有限集，满足

$$\begin{aligned} P(\exists f \in \mathcal{F}: R(f) - \hat{R}(f) \geq \epsilon) &= P\Big(\bigcup_{f\in\mathcal{F}}\{R(f) - \hat{R}(f) \geq \epsilon\}\Big) \\ &\leq \sum_{f\in\mathcal{F}} P(R(f) - \hat{R}(f) \geq \epsilon) \\ &\leq d\exp(-2N\epsilon^2) \end{aligned}$$

# 泛化误差约束定理证明

- 下述不等式等价：

$$P(\exists f \in \mathcal{F}: R(f) - \hat{R}(f) \geq \epsilon) \leq d\exp(-2N\epsilon^2)$$
$$\Updownarrow$$
$$P(\forall f \in \mathcal{F}: R(f) - \hat{R}(f) < \epsilon) \geq 1 - d\exp(-2N\epsilon^2)$$

- $\delta$参数化方式如下：

$$\delta = d\exp(-2N\epsilon^2) \Leftrightarrow \epsilon = \sqrt{\frac{1}{2N}\log\frac{d}{\delta}}$$

那么泛化误差以$1-\delta$概率被约束

$$P(R(f) < \hat{R}(f) + \epsilon) \geq 1 - \delta$$

## 泛化误差定理的推广形式

**引入**

这个优化的insight在于

1. 希望使用$f^L$的相对连续的结构进行优化。
2. $\cup$转$\Sigma$ 这一步为了尽量不浪费，使用误差稍大的有限覆盖来取代无穷求和。
3. 计算机有限位数表示的非本质性——难以相信无穷精确度的情况下有什么实质的区别导致定理约束不住了。

如果$\mathcal{F}$为无限假设集或为一个$d$很大的有限假设集，但是有较好的连续性，那么就可以考虑使用如下推广以达到更好的不等式限制：

**定义**

对函数$f, g$定义

$$d(f,g) := max_{x,y}(|\mathcal{L}(y, f(x)) - \mathcal{L}(y, g(x))|)$$

另记

$$f^L(x,y) := L(y, f(x))$$
$$||f^L|| := max_{x,y}(f^L) = d(f, 0)$$

实际上这就是无穷范数。

可以看到

$$d(f,g) = ||f^L - g^L||$$

**条件**

如果针对$\mathcal{F}$,可以找到另一个大小为$d'$的有限假设集$\mathcal{F}'$ 以及一个小常数$\alpha$,使得$\forall f \in \mathcal{F}, \exists f' \in \mathcal{F}'$ 使得$d(f, f') \leq \alpha$ 。

注：大多数情况下，可以选取$\mathcal{F}'$ 为$\mathcal{F}$一个合适的子集，使得$d'$和$\alpha$尽可能小。

**证明**

那么就有

$$|R(f) - R(f')| \leq \alpha$$
$$|\hat{R}(f) - \hat{R}(f')| \leq \alpha$$

而对于有限假设集$\mathcal{F}'$，套用经典的泛化误差公式，有

$$R(f') \leq \hat{R}(f') + \epsilon(d', N, \delta)$$

结合这三个不等式，有**结论**

$$|R(f) - \hat{R}(f)| \leq \epsilon(d', N, \delta) + 2\alpha \tag{1}$$

即为泛化误差定理的推广形式。

# 带参情况下的泛化误差定理的推广形式

上面的式子有一个问题，就是$d'$与$\alpha$和$f^L$都有关。我们希望能使用$f^L$的一些信息消除这种相关性。

如果$f$可以用参数确定，即可写成$f_\theta$的形式，且$\theta \in [a, b]$,

（注：这里只考虑$\theta$是实数标量的情况。如果$\theta$是向量，使用雅克比行列式应能得到类似结论。）

那么直接推出：

若存在一列$\theta_1 = a < \theta_2 < \cdots < \theta_{d'} = b$，使得对任意$\theta_i \leq \theta \leq \theta_{i+1}$，有$d(f_\theta, f_{\theta_i}) \leq \alpha$或$d(f_\theta, f_{\theta_{i+1}}) \leq \alpha$，那么$(1)$式仍然成立。

然后我们再看看针对$f^L$的微分与范数的一些结论

如果$f^L$针对$\theta$有较好的连续性，且$\Delta\theta$是正小量，那么有$||f^L(\theta + \Delta\theta) - f^L|| \approx ||\frac{\partial f^L}{\partial \theta}||\Delta\theta$

（注：$\frac{\partial f^L}{\partial \theta}$ 指 $\frac{\partial (f^L)}{\partial \theta}$。类似地，下面的$\mathrm{d}f^L$指$\mathrm{d}(f^L)$。

令$\Delta\theta$趋于$0$，上式就是说

$$||\mathrm{d}f^L|| = ||\frac{\partial f^L}{\partial \theta}||\mathrm{d}\theta$$

其中$\frac{\partial f^L}{\partial \theta}$是$f^L$对$\theta$的变分。

我们现在想让$d'$尽可能小以得到好的结果。

考虑如下取法：对任意$\theta_i \leq \theta \leq \theta_{i+1}$，不仅有$d(f_\theta, f_{\theta_i}) \leq \alpha$或$d(f_\theta, f_{\theta_{i+1}}) \leq \alpha$，且存在$\theta_i \leq \theta_m \leq \theta_{i+1}$ 使得$d(f_{\theta_m}, f_{\theta_i}) = \alpha$且$d(f_{\theta_m}, f_{\theta_{i+1}}) = \alpha$

（注：这是为了在平缓的地方少取点，在陡峭的地方多取点，尽量拉大$f_{\theta_i}$与$f_{\theta_{i+1}}$之间的距离，使得在某个点间距被拉满了）

对这种取法，因为

$$\alpha = d(f_{\theta_m}, f_{\theta_i}) = ||f^L_{\theta_m} - f^L_{\theta_i}|| = ||\int_{\theta_i}^{\theta_m} \mathrm{d}f^L|| \leq \int_{\theta_i}^{\theta_m} ||\mathrm{d}f^L|| = \int_{\theta_i}^{\theta_m} ||\frac{\partial f^L}{\partial \theta}||\mathrm{d}\theta$$

（注：这个积分与范数交换的不等号使用了三角不等式）

与类似的

$$\alpha \leq \int_{\theta_m}^{\theta_{i+1}} ||\frac{\partial f^L}{\partial \theta}||\mathrm{d}\theta$$

相加得

$$2\alpha \leq \int_{\theta_i}^{\theta_{i+1}} ||\frac{\partial f^L}{\partial \theta}||\mathrm{d}\theta$$

对所有的$i$得到的上述式子进行相加，得到

$$2d'\alpha \leq \int_{\theta_i}^{\theta_{i+1}} ||\frac{\partial f^L}{\partial \theta}||\mathrm{d}\theta$$

就得到，存在一种取法，使得对$(1)$式中$d'$至少有如下这么小：

$$d' \leq \frac{\int_a^b ||\frac{\partial f^L}{\partial \theta}||\mathrm{d}\theta}{2\alpha}$$

代入$(1)$式有**结论**：

$$|R(f) - \hat{R}(f)| \leq \epsilon(\frac{\int_a^b ||\frac{\partial f^L}{\partial \theta}||\mathrm{d}\theta}{2\alpha}, N, \delta) + 2\alpha \tag{1.1}$$

我们还可以对右侧取极小值，即对$\alpha$求导为$0$，可以知道右侧取到最小值时$\alpha$满足
$\epsilon(\frac{\int_a^b ||\frac{\partial f^L}{\partial \theta}||\mathrm{d}\theta}{2\alpha}, N, \delta) = 8N\alpha$

于是我们仅仅使用$f^L(\theta)$与$N$也可以得到一个很好的概率成立误差界！