# computational learning theory

计算学习理论是机器学习的理论基础，假设有数据集$D=\{x_1,x_2,...,x_m\}$，对其中所有$x_i$都是从样本空间$\mathscr{D}$进行独立同分布采样(independent and identically distributed)得到。考虑二分类问题$y_i\in\{1,-1\}$，假设空间$h\in\mathscr{H}$，对于$h\in h_\theta$，其泛化误差定义为

$$E(h;\mathscr{D})=P_{x-\mathscr{D}}(h(x)\neq y)$$

h再数据集D上的经验误差定义为

$$\hat{E}(h;D)=\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}(h(x_i)\neq y_i)$$

## 经验误差最小化(ERM)

经验误差最小化可以表示为，$\hat{h}$表示对和估计

$$\hat{h}=\arg\min_{h\in\mathscr{H}}\hat{E}(h;D)$$

**Union Band**

$$P(A_1\cup...\cup A_k)\leq P(A_1)+...+P(A_k)$$

**Hoeffding不等式**

对于一组变量$z_1,z_2,...,z_m$满足伯努利分布，并且$P(z_i=1)=\phi$，则对$\phi$的估计

$$\hat{\phi}=\frac{1}{m}\sum_{i=1}^{m}z_i$$

估计误差存在一个上限

$$P(|\phi-\hat{\phi}|>\epsilon)\leq 2\exp(-2m\epsilon^2)$$

**有限的假设空间**

假设空间$\mathscr{H}=\{h_1,h_2,...,h_k\}$，ERM：$\hat{h}=\arg\min_{h\in\mathscr{H}}\hat{E}(h;D)$，我们要证明的是在经验误差最小化下得到的假设$\hat{h}$的一般泛化误差$E(\hat{h};\mathscr{D})$是存在上限。我们定义$z_i=\mathbb{I}(h_j(x_i)\neq y_i)$，显然$z_i$满足伯努利分布，则$P(z_i=1)=\phi=\epsilon(h_j)$，使用Hoeffding不等式，

$$P(|\epsilon(h_j)-\hat{\epsilon}(h_j)|>\epsilon)\leq 2\exp(-2m\epsilon^2)$$

对所有的h

$$P(h_j\in\mathscr{H}|\epsilon(h_j)-\hat{\epsilon}(h_j)|>\epsilon)\leq\sum_j P(\epsilon(h_j)-\hat{\epsilon}(h_j)|>\epsilon)\leq 2k\exp(-2m\epsilon^2)$$

对上述结论取非，即存在h，

$$P(h_j\in\mathscr{H}|\epsilon(h_j)-\hat{\epsilon}(h_j)|<\epsilon)\geq 1-\delta$$

这称之为一致收敛(uniform converges)，$1-\delta=1-2k\exp(-2m\epsilon^2)$表示一致收敛的概率。

上述结论存在其他的等价变形描述，固定$\delta,\epsilon$，即我们希望以大于$1-\delta$的概率得到$|\epsilon(h_j)-\hat{\epsilon}(h_j)|<\epsilon$，那么样本数m满足

$$m \geq \frac{1}{2\epsilon^2}\ln\frac{2k}{\delta}$$

若固定$\delta, m$，即我们希望以m个样本，以$1-\delta$的概率可以得到，

$$|\epsilon(h_j)-\hat{\epsilon}(h_j)| \leq \sqrt{\frac{1}{2m}\ln\frac{2k}{\delta}}$$

在假设空间能得到的最好的假设是

$$h^* = \arg\min_{h\in\mathscr{H}} E(h;D)$$

可以得到，由于$|\epsilon(h)-\hat{\epsilon}(h)| > \epsilon$

$$E(\hat{h}) \leq \hat{E}(\hat{h}) + \epsilon \leq \hat{E}(h^*) + \epsilon \leq E(h^*) + 2\epsilon$$

即我们估计得到最好的$\hat{h}$的泛化误差假设空间能得到的最好的假设$h^*$泛化误差之差存在上限。

同样的，若固定$\delta, m$

$$E(\hat{h}) \leq E(h^*) + 2\sqrt{\frac{1}{2m}\ln\frac{2k}{\delta}}$$

右边第一项表示偏差，第二项表示方差

**无限的假设空间**

尽管假设空间$\mathscr{H}$可能包含无数个假设，但是对数据集D上有限的数据的可能结果数表示是有限的。对于二分类问题，有m个数据，则最大有$2^m$种可能的表示，若假设空间$\mathscr{H}$可以实现对数据集D上的所有表示，则称D能被$\mathscr{H}$打散。

**VC维**

假设空间$\mathscr{H}$是能被$\mathscr{H}$打散的最大D的大小

$$VC(\mathscr{H}) = \max\{m : \Pi_{\mathscr{H}} = 2^m\}$$

假设$VC(\mathscr{H}) = d$，我们有

$$P(|\epsilon(h)-\hat{\epsilon}(h)| < \epsilon) \geq 1-\delta, \quad \epsilon = \sqrt{\frac{8d\ln\frac{2em}{d} + 8\ln\frac{4}{\delta}}{m}}$$

若固定$\epsilon, \delta$

$$m = O(d)$$