VC Theory: Symmetrization

http://freemind.pluskid.org/slt/vc-theory-symmetrization

posted on Free Mind on July 29, 2012
generated with pandoc on December 3, 2015
category: Statistical Learning Theory

tags: Binary Classification, Empirical Process

我们上一次介绍了 Hoeffding 不等式，结论是对任意**固定的** hypothesis $h \in \mathcal{H}$，我们有

$$P(E_{\text{out}}(h) - E_{\text{in}}(h) > \epsilon) \leq e^{-2N\epsilon^2}$$

但是正如教授在课上讲的一样，仅仅在一个固定的 hypothesis 上做出这样的保证并不足以构成“学习”，最多只是“验证”。为了保证我们的学习算法从 $\mathcal{H}$ 中选中任何一个 h 都是可行的，我们需要得到这样形式的结论：

$$P\left(\sup_{h\in\mathcal{H}}(E_{\text{out}}(h) - E_{\text{in}}(h)) > \epsilon\right) \leq \text{something}$$

换句话说，我们希望得到的界对于所有 $h \in \mathcal{H}$ 能够**一致**成立。所以接下来我们就来尝试得到这样的结论。具体来讲，本文中我们将会介绍一种叫做 Symmetrization 的技术。方便起见，我们记 $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$，并记 $z = (x, y)$。于是，给定的 N 个训练数据也记为 $\{z_i\}_{i=1}^N$。此外，我们简单地将 E_{out} 记为 E，而 E_{in} 记为 E_N（因为它依赖于 N 个训练数据嘛）。

接下来我们假设除了 $\{z_i\}_{i=1}^N$ 之外还有另一组数据 $\{z_i^*\}_{i=1}^N$，也是 IID 地从 P_{XY} 中采样得到的。这组数据通常称作 *ghost sample*，它仅在理论分析中出现，并不是说我们在实际学习中需要额外的一份数据。

为什么要引入 ghost sample 呢？为了回答这个问题，我们先定义一些辅助的符号，首先，给定了 loss 函数 ℓ 之后，我们可以定义一个 *Loss Class*，它是一个集合，其元素和 Hypothesis Space $\mathcal{H}$ 里的元素一一对应[1]：

$$\mathcal{F} = \{f_h | h \in \mathcal{H}\}$$

这里每个 $f_h : \mathcal{Z} \to \mathbb{R}^+$ 是这样定义的一个函数：

$$f_h(z) = \ell(h, z)$$

看着有点像同意反复，其实就是这么回事。由于 $\mathcal{F}$ 和 $\mathcal{H}$ 的元素一一对应，所以在 $\mathcal{H}$ 里学习也可以等价地认为是在 $\mathcal{F}$ 里学习。接下来我们定义 $\mathcal{F}$ 到 $\{z_i\}_{i=1}^N$ 上的投影为：

[1] 实际上，有可能存在 $h_1 \neq h_2$ 但是 $f_{h_1} = f_{h_2}$，但是由于在任何数据 z 上都有 $\ell(h_1, z) = f_{h_1}(z) = f_{h_2}(z) = \ell(h_2, z)$，所以我们的学习算法或者说我们的 problem formulation 是无法区分这样的 h_1 和 h_2 的，所以如果需要严格一点的话，可以将这样的 h 集合起来构成等价类，这样就能保证 $\mathcal{H}/\sim$ 和 $\mathcal{F}$ 确实是一一对应的了。

$$\mathcal{F}(z_1,\dots,z_N)=\{(f(z_1),\dots,f(z_N)|f\in\mathcal{F}\}$$

它是由一些 N 维向量所构成的集合。接下来到两件事情：第一，由于我们使用的是 binary loss，因此所有 f 实际上只取 0 和 1 两个值，所以 $\mathcal{F}(z_1,\dots,z_N)$ 这个由 N 维 binary vector 所组成的集合的元素个数是有限的，最多不超过 2^N 个——不论原来的集合 $\mathcal{F}$ 是否有限；第二，这个有限的集合完全决定了任意 $f\in\mathcal{F}$ 的 in-sample error[2]，因为

[2] 因为 Loss Class $\mathcal{F}$ 中的函数 f 和 Hypothesis Space $\mathcal{H}$ 中的函数 h 一一对应，所以我们在谈论 f 的 error 的时候，可以认为实际上是在谈论它所对应的那个 h 的 error，以下我们将混用这样的概念。

$$E_N(f)=\frac{1}{N}\sum_{i=1}^{N}f(z_i)$$

我之前在大数定理军团中曾经介绍过一种简单的利用 Union Bound 将上一次讲的 Hoeffding 不等式推广到一致成立的方法，但是那种方法只对有限的 $\mathcal{F}$ 适用，对于无限的 $\mathcal{F}$，我们将会得到一个 $\leq\infty$ 这样的毫无意义的结果。然而这里似乎看到了一个好兆头：$\mathcal{F}$ 里的元素的 error 将由一个有限的集合来完全决定，不论原来的集合 $\mathcal{F}$ 是有限还是无限。不过显然还有一个问题就是这里我们只能刻画 in-sample error，而 out-of-sample error 则不只是依赖于有限的 N 个数据点，而需要在所有 $z\in\mathcal{Z}$ 上求值。于是 Symmetrization 就出场了。

引理 1 (Symmetrization) *对任意的 $\epsilon>0$，且 $N\epsilon^2\geq 2$，我们有*

$$P\left(\sup_{f\in\mathcal{F}}(E(f)-E_N(f))>\epsilon\right)\leq 2P\left(\sup_{f\in\mathcal{F}}(E_N^*(f)-E_N(f))>\frac{\epsilon}{2}\right)$$

这里 E_N^* 是定义在 ghost sample $\{z_1^*,\dots,z_N^*\}$ 上的 in-sample error。

这样一来，不等式的右边就可以用两个有限的集合 $\mathcal{F}(z_1,\dots,z_N)$ 和 $\mathcal{F}(z_1^*,\dots,z_N^*)$ 来进行刻画了，从而避开了无限集的问题。而 ghost sample 的引入也正是为了这个目的。至于究竟如何来进行刻画并得到最终的一致 Hoeffding 界，我们将在下一次进行介绍，本文余下来的部分将用来证明引理 1，不感兴趣的同学可以直接跳过。

简单起见，我们假设不等式左边的上确界可以达到，并在 $f_N\in\mathcal{F}$ 处达到。注意 f_N 是依赖于 E_N 的定义的，因此依赖于 $\{z_i\}_{i=1}^N$。注意到

$$\begin{aligned}&\mathbf{1}\{E(f_N)-E_N(f_N)>\epsilon\}\mathbf{1}\{E(f_N)-E_N^*(f_N)<\epsilon/2\}\\=&\mathbf{1}\left\{(E(f_N)-E_N(f_N)>\epsilon)\wedge(E(f_N)-E_N^*(f_N)<\epsilon/2)\right\}\\\leq&\mathbf{1}\left\{E_N^*(f_N)-E_N(f_N)>\epsilon/2\right\}\end{aligned}\tag{1}$$

在不等式两边先对 ghost sample 求期望，得到

$$\begin{aligned}&\mathbf{1}\{E(f_N)-E_N(f_N)>\epsilon\}P\left(E(f_N)-E_N^*(f_N)<\epsilon/2\right)\\ \leq &P\left(E_N^*(f_N)-E_N(f_N)>\epsilon/2\right)\end{aligned} \tag{2}$$

我们看红色的部分，由 Chebyshev's Inequality[3]，我们有

$$P(E(f_N)-E_N^*(f_N)\geq\epsilon/2)\leq\frac{4\mathrm{var}(f_N)}{N\epsilon^2}$$

再次注意到 f_N 只取值 0 和 1，因此 $\mathrm{var}(f_N)\leq 1/4$[4]，所以

$$\begin{aligned}P(E(f_N)-E_N^*(f_N)\geq\epsilon/2)&\leq\frac{1}{N\epsilon^2}\\ \Rightarrow 1-\frac{1}{N\epsilon^2}&\leq P\left(E(f_N)-E_N^*(f_N)<\epsilon/2\right)\end{aligned}$$

再注意到引理中有 $N\epsilon^2\geq 2$ 这个奇怪的条件，所以 $1-\frac{1}{N\epsilon^2}\geq 1/2$，带入 (2) 的红色部分，得到：

$$\frac{1}{2}\mathbf{1}\{E(f_N)-E_N(f_N)>\epsilon\}\leq P\left(E_N^*(f_N)-E_N(f_N)>\epsilon/2\right)$$

最后两边同时再对真正的 training sample 求期望，即得到：

$$\begin{aligned}P\left(\sup_{f\in\mathcal{F}}(E(f)-E_N(f))\right)&=P\left(E(f_N)-E_N(f_N)>\epsilon\right)\\&\leq 2P\left(E_N^*(f_N)-E_N(f_N)>\epsilon/2\right)\\&\leq 2P\left(\sup_{f\in\mathcal{F}}(E_N^*(f)-E_N(f))>\frac{\epsilon}{2}\right)\end{aligned}$$

这样一来，引理就得证了。

[3] 注意这里 $E(f_N)$ 是常量，而 f_N 虽然依赖于 $\{z_i\}_{i=1}^N$，但是与 ghost sample 无关，否则这里不能直接套用 Chebyshev's Inequality。

[4] **简单证明**：记 $p_0=P(f_N=0)$，$p_1=P(f_N=1)$，则 $\mathbb{E}[f_N]=p_1$，而 $\mathrm{var}(f_N)=\mathbb{E}[(f_N-p_1)^2]=p_1p_0^2+p_0p_1^2$，注意到 $p_0+p_1=1$，由均值不等式立即得到 $\mathrm{var}(f_N)\leq 1/4$。