VC Theory: Hoeffding Inequality

http://freemind.pluskid.org/slt/vc-theory-hoeffding-inequality

posted on Free Mind on July 27, 2012
generated with pandoc on December 3, 2015
category: Statistical Learning Theory

tags: Binary Classification, Probability Theory

之前提过的 Professor Yaser Abu-Mostafa 的机器学习课程在 Lecture 5、6、7 三课中讲到了 VC Theory 的一些内容，用来回答他在课程中提到的 “Can We Learn？” 这个问题。更具体地来说，他这里主要解决了 binary classification 问题中的 *Learnability* 的问题，结论就是：如果 hypothesis space $\mathcal{H}$ 的 VC dimension 是有限的，那么就是 Learnable 的。

建议感兴趣的同学仔细看一下这几课的内容（当然前面几课的内容作为背景铺垫也是非常推荐的），教授讲得非常好，生动形象，并且思路非常清晰，忽略掉了一些细节和不同的 variants，抓住了主要脉络让人理解 Statistical Learning Theory 背后的思想，而不是一下子就陷入一堆复杂的理论和推导中。而本文我希望将教授省略掉的一些细节补全出来，也是帮助自己整理一下学到的东西。

首先是问题设定，在这几课中一直讲的是 binary classification，记 input space 为 $\mathcal{X}$，output space 为 $\mathcal{Y}=\{+1,-1\}$，假设 $\mathcal{X}\times\mathcal{Y}$ 上有一未知的概率分布 P_{XY}，给定一个 hypothesis space $\mathcal{H}\subset\mathcal{Y}^{\mathcal{X}}$，一个 loss function $\ell:\mathcal{H}\times\mathcal{X}\times\mathcal{Y}\to\mathbb{R}^+$，和 N 个独立地从 P_{XY} 里采样出来的数据 sample $\{(x_i,y_i)\}_{i=1}^N$，目的是为了得到一个 final hypothesis $g\in\mathcal{H}$ 作为我们学习的结果，使得 *out-of-sample error*

$$E_{\text{out}}(g)=\mathbb{E}_{P_{XY}}\left[\ell(g,X,Y)\right]$$

足够小。这里有几点需要说明的地方，一是关于 P_{XY}，这是最 general 的 case，我们可以写为 $P_{XY}=P_XP_{Y|X}$，有一种简单的特殊情形是 $P_{Y|X}$ 只取值 0 或 1，也就是每个 x 确定地（更确切地说 almost surely）对应一个 y，此时 x 和 y 的关系可以用一个函数来刻画，通常称作是 target function。不过在课上教授提到即使是最 general 的 case，通常也不推荐写做联合分布 P_{XY} 的样子，因为我们所关注的（需要学习的）其实只是 $P_{Y|X}$，而 P_X 只是为了让我们得以在一个概率框架下来讨论问题而引入的，并不是我们很关心的对象[1]。不过有时候为了方便我们仍然会使用联合分布的记号，甚至有时候直接记 $\mathcal{Z}=\mathcal{X}\times\mathcal{Y}$、$z=(x,y)$ 以及 $P_Z=P_{XY}$。

此外，在 binary classification 中，我们将 loss function （在课中称为 error measure 或者 pointwise error measure）定为 binary error:

$$\ell(h,x,y)=\mathbf{1}\{h(x)\neq y\}$$

[1] 当然，对于 Generative Modeling 派系而言的话 P_X 也是需要去学习的，但那并不属于我们这里讨论的问题。

首先面临的问题是要优化的目标 E_{out} 是无法计算的，因为 P_{XY} 是未知的。所以我们采取一种迂回战术：在第 6 课 Theory of Generalization 中最后得到了这样形式的一个 generalization bound，根据 training sample $\{(x_i, y_i)\}_{i=1}^N$ 的随机性，以不低于 $1-\delta$ 的概率满足

$$E_{\text{out}} \leq E_{\text{in}} + \Omega \tag{1}$$

这里 E_{in} 称为 in-sample error，是在给定的训练数据上计算的 error，具体定义为

$$E_{\text{in}}(g) = \frac{1}{N}\sum_{i=1}^{N} \ell(g, x_i, y_i)$$

这一项是我们可以计算的而 Ω 项是这套理论中的一个重点，这里我们需要知道的是它是可以估计的就可以了。这样一来，我们就可以通过优化 E_{in} 的方式来间接地优化无法计算的 E_{out}。这就是这里所讲的理论的一个简单蓝图。

接下来让我们陷入到细节中去，首先是建立 E_{out} 和 E_{in} 之间的联系。首先，对于任意 $h \in \mathcal{H}$，我们记 $\ell(h, x, y)$ 为 $\ell_h(x, y)$，因此 ℓ_h 实际上是一个从 $\mathcal{X} \times \mathcal{Y}$ 到 $\mathbb{R}^+$ 的函数，将它在 P_{XY} 上求期望 $\mathbb{E}_{P_{XY}}[\ell_h(X, Y)]$ 实际上就是 out-of-sample error $E_{\text{out}}(h)$。现在**固定**一个 h，将 sample space $\mathcal{X} \times \mathcal{Y}$ 中的每一对点 (x, y) 带入进 ℓ_h 中，都可以得到 0 或 1 这两个值[2]，在课堂上教授用了花瓶里装的两种不同颜色的鹅卵石来做了一个很形象的比喻。

[2] 因为我们这里用的是 binary loss。

虽然由于 P_{XY} 未知我们没法直接求，但是我们有从 P_{XY} 中采样出来的一个 sample，也就是我们的训练数据，在概率论中根据 sample 来估算期望有一个标准的做法，就是计算 “经验期望” $1/N\sum_{i=1}^N \ell_h(x_i, y_i)$，可以看到这其实就是我们刚才定义的 E_{in}。评估这个估计的好坏的方法是使用大数定理，我之前也介绍过大数定理在机器学习中的作用。大数定理有非常多个不同的变种和形式，为了避免混乱，在课上教授只选用了 Hoeffding 不等式 [Hoeffding, 1963]，这里我们也集中在这个形式上。

定理 1 (Hoeffding Inequality) *设相互独立的随机变量 $\xi_1, \ldots, \xi_N$ 满足 $\xi_i \in [a, b]$, $i = 1, \ldots, N$，记 $\bar{\xi} = 1/N\sum_{i=1}^N \xi_i$，则对任意 $\epsilon > 0$，有*

$$P\left(\bar{\xi} - \mathbb{E}[\bar{\xi}] > \epsilon\right) \leq \exp\left(-2\frac{N\epsilon^2}{(b-a)^2}\right)$$

对于**固定的** h，我们可以将定理中的 ξ_i 与 $\ell_h(x_i, y_i)$ 对应起来，此时 $\bar{\xi}$ 就和 E_{in} 对应起来，而 $\mathbb{E}[\bar{\xi}]$ 则对应了 E_{out}，又注意到 binary loss 只取 0

和 1 两个值，因此定理中 $a=0, b=1$，于是我们可以得到如下的简单形式：

$$P\left(E_{\text{in}}-E_{\text{out}}>\epsilon\right)\leq \exp\left(-2N\epsilon^2\right) \tag{2}$$

注意到在课堂中使用的实际上是这样的形式：

$$P\left(|E_{\text{in}}-E_{\text{out}}|>\epsilon\right)\leq 2\exp\left(-2N\epsilon^2\right) \tag{3}$$

实际上只要注意到

$$\begin{aligned}P\left(|E_{\text{in}}-E_{\text{out}}|>\epsilon\right) &= P\left(E_{\text{in}}-E_{\text{out}}>\epsilon\right)+P\left(E_{\text{in}}-E_{\text{out}}<-\epsilon\right)\\ &= P\left(E_{\text{in}}-E_{\text{out}}>\epsilon\right)+P\left(-E_{\text{in}}+E_{\text{out}}>\epsilon\right)\end{aligned}$$

其中红色的部分用 $-\ell_h(x_i,y_i)$ 与 ξ_i 进行对应，就可以立即从 (2) 得到 (3) 了。

接下来，我们记 (2) 右边为 δ，反解出

$$\epsilon=\sqrt{\frac{\log\frac{1}{\delta}}{2N}}$$

于是，由 (2) 我们可以得到以不低于 $1-\delta$ 的概率，满足

$$E_{\text{out}}\leq E_{\text{in}}+\sqrt{\frac{\log\frac{1}{\delta}}{2N}} \tag{4}$$

这正是我们在 (1) 中所想要的形式。不过我们有一个问题，也是我一直在强调的，这里所有的 formulation 都只是针对**一个固定的** h 而言的，在机器学习问题中我们需要处理一个 hypothesis space $\mathcal{H}$ 里的所有 h，特别是我们并不事先知道我们的学习算法最终会选取 $\mathcal{H}$ 中的哪一个元素作为 final hypothesis，因此我们需要**同时**对所有的 $h\in\mathcal{H}$ 保证 (4) 都成立。如果 $\mathcal{H}$ 是个有限集，可以很容易通过 *Union Bound* 得到一个结果，这个我在大数定理军团中也介绍过用，不过如果 $\mathcal{H}$ 是无限集，就必须更精细地分析 $\mathcal{H}$ 的结构，这是下次要介绍的内容。本文余下的部分将会用来证明定理 1，不感兴趣的同学可以直接跳过。

进入证明之前，我们首先注意到定理 1 中式子左边的形式是 $P_{\mathbf{Z}}(f(\mathbf{Z}>\epsilon)$，我们可以用 indicator function 把它换一个形式：

$$P_{\mathbf{Z}}(f(\mathbf{Z})>\epsilon)=\mathbb{E}_{\mathbf{Z}}\left[\mathbf{1}\{f(\mathbf{Z})-\epsilon>0\}\right]$$

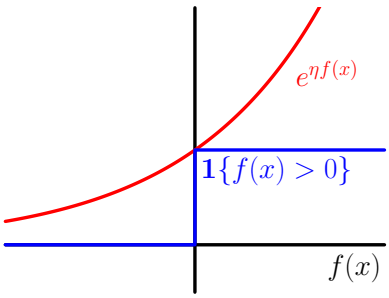

Figure 1: indicator function bounded above by the exponential function.

再注意到这个形式的 indicator function 可以被一个指数函数 bound 住，见图 1。于是

$$P_{\mathbf{Z}}(f(\mathbf{Z}) > \epsilon) \leq \mathbb{E}_{\mathbf{Z}}\left[\exp\left(\eta(f(\mathbf{Z}) - \epsilon)\right)\right]$$

这里 η 是任意一个正数。这个变化在证明包括 Hoeffding 不等式在内的几个类似的不等式里都有用到。直接套用这个式子，我们可以得到

$$\begin{aligned} P\left(\bar{\xi} - \mathbb{E}[\bar{\xi}] > \epsilon\right) &= P\left(N\bar{\xi} - N\mathbb{E}[\bar{\xi}] > N\epsilon\right) \\ &\leq \mathbb{E}\left[\exp\left(\eta(N\bar{\xi} - N\mathbb{E}[\bar{\xi}] - N\epsilon)\right)\right] \\ &= e^{-N\eta\epsilon}\prod_{i=1}^{N}\mathbb{E}\left[\exp\left(\eta(\xi_i - \mathbb{E}[\xi_i])\right)\right] \end{aligned} \tag{5}$$

上式中最后一步是由 $\xi_1, \ldots, \xi_N$ 的独立性得到的。接下来我们尝试去 bound 红色的部分。

引理 1 设 Z 是一个随机变量，且 $a \leq Z \leq b$，则对任意实数 η，我们有

$$\mathbb{E}\left[e^{\eta Z}\right] \leq \frac{b - \mathbb{E}[Z]}{b - a}e^{\eta a} + \frac{\mathbb{E}[Z] - a}{b - a}e^{\eta b}$$

证明很简单，注意到指数函数是凸函数，因此在区间 $[a, b]$ 中的函数图像会在连接 $(a, e^{\eta a})$ 和 $(b, e^{\eta b})$ 两点的直线之下，运用直线方程的两点式公式，然后两边同时求期望即证。

将刚才的红色部分套用这个引理，可以得到：

$$\begin{aligned} \mathbb{E}\left[\exp\left(\eta(\xi_i - \mathbb{E}[\xi_i])\right)\right] &= e^{-\eta\mathbb{E}\xi_i}\mathbb{E}\left[\exp\left(\eta\xi_i\right)\right] \\ &\leq e^{-\eta\mathbb{E}\xi_i}\left(\frac{b - \mathbb{E}\xi_i}{b - a}e^{\eta a} + \frac{\mathbb{E}\xi_i - a}{b - a}e^{\eta b}\right) \\ &= e^{-\eta(\mathbb{E}\xi_i - a)}\left(1 - \frac{\mathbb{E}\xi_i - a}{b - a} + \frac{\mathbb{E}\xi_i - a}{b - a}e^{\eta(b-a)}\right) \\ &= \exp\left(-\eta(b - a)p_i\right) \cdot \exp\log\left(1 - p_i + p_i e^{\eta(b-a)}\right) \\ &= \exp\left(-\eta_i p_i + \log\left(1 - p_i + p_i e^{\eta_i}\right)\right) \end{aligned} \tag{6}$$

这里 p_i 和 η_i 分别是为了简化记号而引入的符号，他们所代表的部分用对应的颜色标出。注意 p_i 是个常量，而 η_i 是依赖于 η 的量。现在我们记指数部分为 $L(\eta_i)$，对 η_i 求导得到：

$$L'(\eta_i) = -p_i + \frac{p_i e^{\eta_i}}{1 - p_i + p_i e^{\eta_i}} = -p_i + \frac{p_i}{(1-p_i)e^{-\eta_i} + p_i}$$
$$L''(\eta_i) = \frac{p_i(1-p_i)e^{-\eta_i}}{[(1-p_i)e^{-\eta_i} + p_i]^2} \leq \frac{\frac{1}{4}[(1-p_i)e^{-\eta_i} + p_i]^2}{[(1-p_i)e^{-\eta_i} + p_i]^2} = \frac{1}{4}$$

红色部分的不等式是根据均值不等式得到的。然后我们将 $L(\eta_i)$ 在 0 点处进行带余项 Taylor 展开：

$$\begin{aligned} L(\eta_i) &= L(0) + L'(0)\eta_i + \frac{1}{2}L''(\zeta)\eta_i^2 \\ &\leq L(0) + L'(0)\eta_i + \frac{1}{8}\eta_i^2 \\ &= \frac{1}{8}\eta^2(b-a)^2 \end{aligned}$$

再由指数函数的单调性加上式 (6) 立即得到：

$$\mathbb{E}\left[\exp\left(\eta(\xi_i - \mathbb{E}[\xi_i])\right)\right] \leq e^{L(\eta_i)} \leq e^{\frac{1}{8}\eta^2(b-a)^2}$$

最后，再回到式 (5)，我们得到

$$\begin{aligned} P\left(\bar{\xi} - \mathbb{E}[\bar{\xi}] > \epsilon\right) &\leq e^{-N\eta\epsilon} \prod_{i=1}^{N} e^{\frac{1}{8}\eta^2(b-a)^2} \\ &= \exp\left(\frac{N}{8}\eta^2(b-a)^2 - N\eta\epsilon\right) \end{aligned}$$

注意到这个式子对任意 $\eta > 0$ 都成立的，所以现在我们可以将不等式的右边对 η 进行最小化，以期望得到一个最好的 bound 。显然，右边的最小值在 $\eta = 4\epsilon/(b-a)^2 > 0$ 处取到，带入之后立即得到

$$P\left(\bar{\xi} - \mathbb{E}[\bar{\xi}] > \epsilon\right) \leq \exp\left(-2\frac{N\epsilon^2}{(b-a)^2}\right)$$

于是 Hoeffding 不等式就证完了。

References

[Hoeffding, 1963] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30.