Lecture 5　Training verus Testing

整理者: LobbyBoy*　2020年2月19日

1. Recap and Review

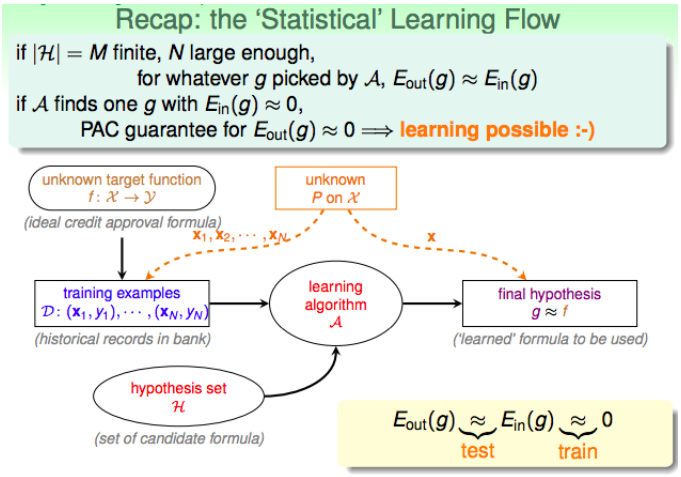

图 1: Learning Flow(学习流程)

在①假说集合的大小有限且②样本容量很大的前提下，透过multiple bins版本的Hoeffding不等式可知，对于任意的演算法A，我们依然可以得到较为“安全”的回传结果g，即g的in-sample error(经验误差)与其out-of-sample error(泛化误差)很接近。在此保证之下，如果某个演算法回传了一个经验误差接近于0的hypothesis，那么我们就可以在PAC的程度上说，我们拿到了一个泛化误差也接近于0的hypothesis——这正是我们想要的结果。

仔细考虑上述过程，我们其实可以将learning拆分为两个关键的问题：①我们能够保证回传结果的经验误差约等于泛化误差吗？②我们能够真的找到经验误差接近于0的hypothesis吗？如果对于上述两个问题的回答都是Yes，那么就形成了一个完美的learning的过程。然而我们会看到，由于M(假说集合的大小)在其中扮演着微妙的角色，使得我们不得不在这两个问题中进行trade-off，也就是鱼和熊掌很难兼得。

*本笔记根据台湾大学林轩田教授于线上教育平台Coursera开设的“机器学习基石”课程整理而成(课程内容见: https://www.coursera.org/learn/ntumlone-mathematicalfoundations/home/welcome)。笔记内的大多数图片来自于林老师的课程slides。感谢林老师能够将如此精彩的课程通过线上平台同所有人分享，thanks!

先考虑M较小。

M较小时，由multiple bins版本的Hoeffding不等式可知，在很大程度上，对于任何演算法而言回传结果的经验误差都会是其泛化误差的较准确反映，这是我们希望看到的。然而，M较小也说明我们的候选者比较少，因为可能较难选到经验误差很小的hypothesis。

再考虑M较大。

M较大时，候选者变多了，因此我们很可能可以选到经验误差很小的hypothesis。然而，由multiple bins版本的Hoeffding不等式可知，回传结果的经验误差不再是其泛化误差的较准确反映，这样即使拿到经验误差很小的回传结果，我们也不能很确定其真的是一个好的hypothesis。

更甚者，在例如Perceptron的机器学习模型中，hypothesis是空间中的一个个超平面，有无数个，即M为$+\infty$。再回到multiple bins版本的Hoeffding不等式，我们可以看到不等式右边也跟着趋于无穷大，这说明回传结果的经验误差并不再是其泛化误差的良好反映，learning好像又变得不可行了。

然而在实践中我们看到，PLA算法的的确确具有学习的能力，因此直接将M作为正无穷带入Hoeffding不等式的方法在某种程度上是错误的，或者说是应该被修正的。一种很直接的修正思路是：PLA算法的优良表现说明M并不是正无穷，那么可能存在其他的某种度量hypothesis数量的方式，我们将这种度量方式作用于假说集合$\mathcal{H}$的结果记作$m_{\mathcal{H}}$，并修正我们的multiple bins版本的Hoeffding不等式：

$$P\Big[|E_{in}(g) - E_{out}(g)| > \epsilon\Big] \leq 2m_{\mathcal{H}}e^{-2\epsilon^2 N}$$

在接下来的几节中，我们将讨论如果建构这个新的假说集合大小的度量方式。

2. Effective Number of Lines

首先我们回忆一下multiple bins版本的Hoeffding不等式的产生过程(将事件$|E_{in}(h_i) - E_{out}(h_i)| > \epsilon$简写为$B_i$)：

$$\begin{aligned} P\Big[|E_{in}(g) - E_{out}(g)| > \epsilon\Big] &\leq P\Big[B_1 \text{ or } B_2 \text{ or } ... \text{ or } B_M\Big] \\ &\overset{union\ bound}{\leq} P[B_1] + P[B_2] + ... + P[B_M] \\ &\leq M(2e^{-2\epsilon^2 N}) \\ &= 2Me^{-2\epsilon^2 N} \end{aligned}$$

从上述过程中我们看出，M的出现是由于我们使用了union bound去放松上界。容易想到的一点是，如果某个假说集合中的hypothesis都非常类似，那么这种使用union bound的粗暴的放松方式，会损失掉许多精确性。一种极端的情况是，假说集合中的所

有hypothesis都完全一样，那么$P\left[B_1 \ or \ B_2 \ or \ ... \ or \ B_M\right]$就应该等于$P[B_1]$，而使用union bound得到的上界为$M \cdot P[B_1]$这样的上界。可以看到这种情况下，用union bound得到的上界并不是一个很精确的上界，甚至是没有意义的上界(和大于1)。

简而言之，由于假说集合中可能存在相似的hypothesis，而直接使用union bound放松上界会完全忽略掉这些相似性，导致上界被过度放松。因此，我们可能可以尝试对假说集合中的hypothesis进行“分类”，归为一类的hypotheses算作“一整个”，然后对“整个整个”的类别组进行union bound的放松，也许可以得到较之前更精确的上界。下面，我们开始探索如何度量hypothesis的相似性，并以perceptron为例。

以二维平面的perceptron为例。此前我们认为，由于平面中存在无数条直线，因此对于PLA，其假说集合的容量为无穷大。现在假设我们手头有一个样本点$\mathbf{x_1}$。如果我们从这个样本点的视角出发来看整个空间中的直线，会有多少类呢？2类：一类直线将其分类为+1，另一类直线将其分类为-1，不会再存在第三类直线(不考虑直线过该点)。

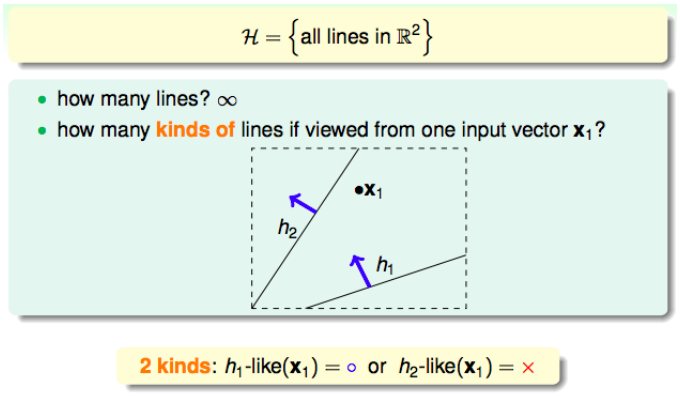

图 2: 其实只有“两类”直线

当手头有两个样本点时，依照上面的思路，我们可以很快画出，平面内的直线有4类，可以记作：{o,o},{x,x},{o,x},{x,o}。然而，当手头有三个样本点时，我们可能会遇到一些麻烦：当三个样本点不在一条直线上时，直线有8类；但当三个样本点共线时，只有6类，因为{o,x,o}与{x,o,x}这两种情形并不能被平面上任何直线做出来。此时我们会说，当手头有三个样本点时，平面中的直线“至多”被分为8类。

当手头有四个样本点时，我们可能会立刻想到平面内的直线可以被分为$2^4 = 16$类，或者说“至多”被分为16类。然而，下图中的两种情形是**无论样本点如何排列都无法被平面中的直线所分割形成的**。因此，当样本点数目为4时，平面中的直线“至多”被分为14类，并非16类。

综上，如果我们从样本点的视角出现，给定N个样本点$\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_N}$，平面中的直线数量并不再是无限多条，而是至多有限多组，例如：1个样本点时至多为2组，2个样本点时至多为4组，3个样本点时至多为8组，4个样本点时至多为14组。我们将上面“2组”、“4组”、“8组”、“14组”称为“有效直线数目”，即effective number of lines，记

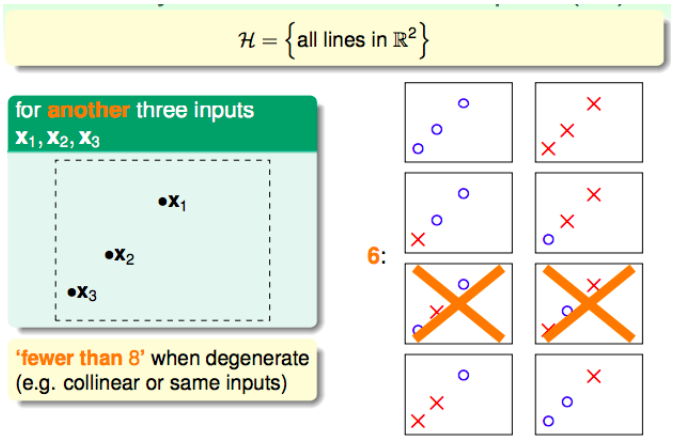

图 3: 三点共线，只有6类

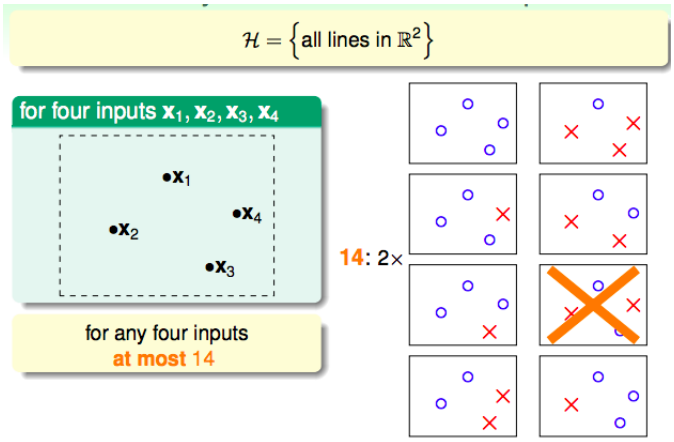

图 4: 三点共线，只有6类

作$effective(N)$，因为很明显它是样本点数目的函数。从上面的数据中我们不难看出两个规律：第一，$effective(N) \leq 2^N$；第二，可能存在某个“断点”，在此断点前，$effective(N) = 2^N$，在此断点及之后，$effective(N) < 2^N$。在PLA中，断点为$N = 4$。

最后，我们将multiple bins版本的Hoeffding不等式右侧的M用$effective(N)$替换，得到：

$$P\Big[|E_{in}(g) - E_{out}(g)| > \epsilon\Big] \leq 2 \cdot effective(N) \cdot e^{-2\epsilon^2 N}$$

这样替换的好处是：我们已经发现了在样本数量大于4之后，$effective(N)$的增长不再是指数型增长，而与其相乘的$e^{-2\epsilon^2 N}$则是伴随着N的增长以指数型速度下降。因此，当N足够大时，我们可以期望不等式的右侧将趋于0。因此，我们便可以解释为什么PLA能够在实践中取得很好的效果，而这是用此前的M所无法解释的。

3. Effective Number of Hypothesis

在上一节中，我们以Perceptron为例，探讨了二维平面内作为hypothesis的直线的有效数目，而本节我们将把effective number of lines的概念和思想拓展到其他假说集合中，计算某个假说集合中effective hypotheses的数量。

1) Dichotomies：对分

考虑某二分类问题的假说集合：

$$\mathcal{H} = \left\{ hypothesis\ h : \mathcal{X} \to \{x, o\} \right\}$$

对于给定的N个样本点$\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_N}$，我们将一个hypothesis作用于其所有成员而得到的$\{o, x\}^N$结果称为一个dichotomy，即：

$$h(\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_N}) = (h(\mathbf{x_1}), h(\mathbf{x_2}), ..., h(\mathbf{x_N})) \in \{o, x\}^N$$

而将假说集合中所有hypotheses作用于这批N个样本点所得到的dichotomies的集合记作$\mathcal{H}(\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_N})$，其大小记作$\left|\mathcal{H}(\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_N})\right|$。易知，$\left|\mathcal{H}(\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_N})\right|$并不是无限大的，而是最多为$2^N$。

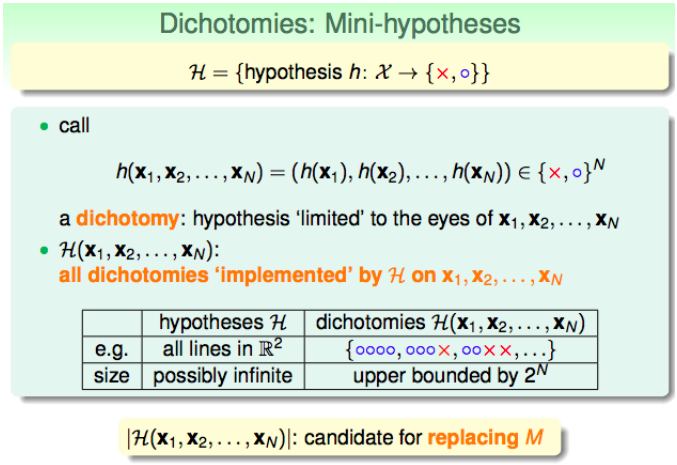

图 5: Dichotomies: 对分

2) Growth Function：增长函数

$\left|\mathcal{H}(\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_N})\right|$并不是一个很好的替换M的指标，因为该指标依赖于具体的Input set $(\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_N})$，就像我们在PLA中所看到的那样，有些输入集能够使得假说集合的对分达到2^N，有些输入集则不行。为了摆脱指标对于具体输入集的依赖，使之只取决于输入

集的大小N，我们可以对$\left|\mathcal{H}(\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_N})\right|$求最大值：

$$m_{\mathcal{H}}(N) = \max_{(\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_N}) \in \mathcal{X}} \left|\mathcal{H}(\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_N})\right|$$

上述过程得到的$m_{\mathcal{H}}(N)$，只取决于样本容量，而不取决于具体的样本组合，相当于PLA例子中“N=1时至多为2个，N=2时至多为4个，N=3时至多为8个，N=4时至多为14个……”。我们将这个指标称为“增长函数(growth function)”。二维平面中直线的增长函数如下：

lines in 2D

N	$m_{\mathcal{H}}(N)$
1	2
2	4
3	$\max(\ldots, 6, 8) = 8$
4	$14 < 2^N$

图 6: 二维平面中直线的增长函数

3) Some examples：增长函数的一些算例

第一种假说集合，Positive Rays。背景依然是二分类问题，输入空间是1维直线，假说集合中的每一个hypothesis均为数轴上某一点向数轴正向引出的射线，位于该射线上的input全部被该hypothesis认为是+1，其余点为-1，如下图所示。很容易算出，这样的假说集合，增长函数为$m_{\mathcal{H}}(N) = N + 1$。

(Hint: 考虑最general的情况，即不考虑所有点重合，因为增长函数是max的情形；N个样本点把数轴分成了N+1段，每一段中的所有hypotheses是一类，因此有N+1种hypotheses，能够将N个样本划成N+1种不同的对分。)

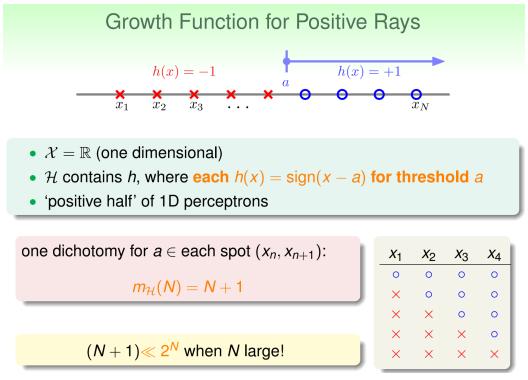

图 7: Positive Rays

第二种假说集合，Positive Intervals。输入空间依然是1维直线，但假说集合中的每一个hypothesis变为一个个线段，位于该线段上的input被该hypothesis认为是+1，其余点为-1，如下图所示。这样的假说集合，增长函数为$m_{\mathcal{H}}(N) = \frac{1}{2}N^2 + \frac{1}{2}N + 1$。

(Hint: 从N个点中任取两点作为positive interval的两端，可以产生各不相同的对分方式，故共有C_N^2种，另外还有一种是将所有点预测为-1，因此共$C_N^2 + 1$种。)

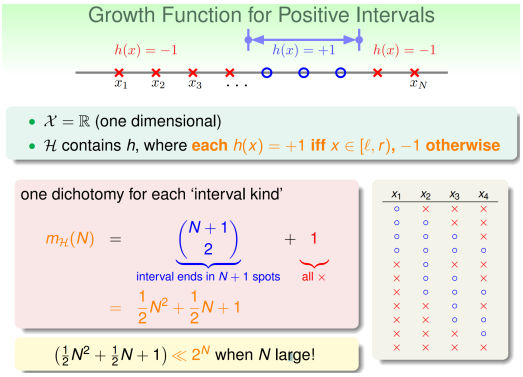

图 8: Positive Intervals

第三种假说集合，Convex Set。输入空间依然是2维平面，假说集合为平面上所有凸集。凸集上及其内的点被预测为+1，其余点被预测为-1。可以证明，该假说集合的增长函数为$m_{\mathcal{H}}(N) = 2^N$，即永远以最快的指数型增长增长下去。

(Hint: 假设N个点在一个圆周上，那么对于2^N种对分，每一种对分都可以通过连结所有+1点的方式，构造出相应的凸集，完成该种对分的实现。)

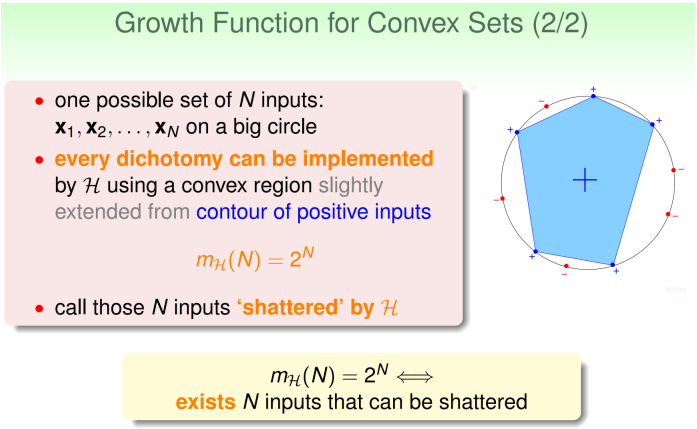

图 9: Convex Sets

在第三个例子中，作为假说集合的所有凸集，其增长函数为$m_{\mathcal{H}}(N) = 2^N$。这里引入“shatter(打散)”的概念。如果对于某N个样本点，假说集合能够做出其所有(2^N)种对分，那么我们说该假说集合能够shatter掉这N个样本点。2维平面中的直线能够shatter掉任何1个样本点、任何2个样本点、某些3个样本点。但从4个样本点开始，不论这4个点如何排

列，共线或不共线，都无法被shatter。但对于2维平面中的凸集，它可以shatter掉任意N个样本点。

补充(Decision Stump): Consider positive and negative rays as $\mathcal{H}$, which is equivalent to the perceptron hypothesis set in 1D. The hypothesis set is often called "decision stump" to describe the shape of its hypotheses. What is the growth function $m_{\mathcal{H}}(N)$? (2(N-1)+2=2N)

4. Break Point

在上一节中，我们计算得到了三类假说集合的增长函数的具体形式。但对于2D的感知机，我们只知道其增长函数在$N = 1, 2, 3, 4$时的取值，并且看出在$N = 4$时假说集合开始无论如何也无法打散全部的样本点，仍然没有得到其具体形式，但我们有理由期望其增长函数是polynomial型的，而非exponential型的，理由是：①我们已经发现在N=4时2D perceptron无法shatter；②如果是exponential型的话，multiple bins版本的Hoeffding不等式在将M换成增长函数后，依然无法提供有效的保证，因为增长函数是指数型增长，后面的exp项是指数型下降，但实践证明PLA的效果很好，所以其增长函数一定不是exponential型的。

对于2D感知机，我们知道它能够shatter 某3个样本点(exists shatter for 3 inputs)，但无法shatter掉任何4个样本点(no shatter for all 4 inputs)。我们将4称为2D perceptrons的最小“断点(break point)”。这是因为，既然无法shatter掉any四个样本点，那么肯定也无法shatter掉any五个、六个以及更多的样本点，所以5、6……都是它的断点，而4是最小的那一个。简而言之，断点即假说集合无论如何也shatter不掉的样本数量。

再次注意，如果k是假说集合的一个断点，这表明对于所有(任意)k个样本点的组合，假说集合都不能shatter掉。例如2D感知机，N=3时，虽然三点共线时不能被shatter，但是存在三点不共线的情况能够被shatter，因此k=3并不是2D感知机的一个断点。

最后，对于某个假说集合，我们可以初步归纳一下其增长函数的增长速度与其(最小)断点之间的关系：

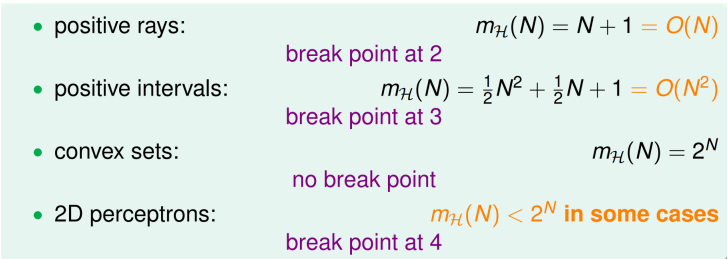

图 10: Growth function & Break point

初步可以猜想：当不存在断点时，$m_{\mathcal{H}}(N) = 2^N$；当(最小)断点为k时，$m_{\mathcal{H}}(N) = O(N^{k-1})$。其推导将在下一章中介绍。