Part VI Learning Theory

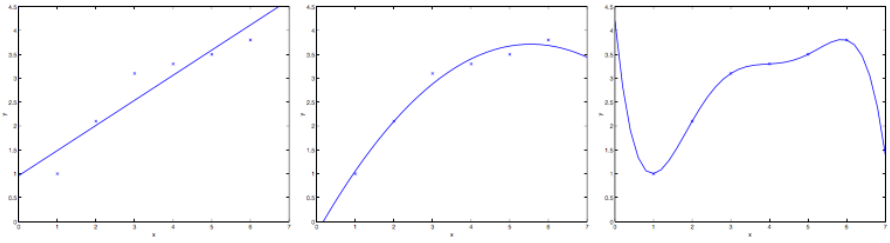
笔记

2017-06-28

目录

1 Bias/variance tradeoff

当我们提到线性回归时，我们讨论的问题其实是是否可以适应一个简单的模型，例如说线性的“$y = \theta_0 + \theta_1 x$”，或者说是一个更复杂的模型，例如多项式“$y = \theta_0 + \theta_1 x + ...\theta_5 x^5$”。我们从讲义上的图就能看得出，最右边的那幅经过了五次多项式拟合但结果并不是一个好的模型。详细地说就是，即使 5 次多项式拟合后的模型在训练集中，能根据 x 得到很好的 y 的预测值，但我们并不能很好地预测出不在训练集中的房屋的价格。换句话说，从训练集中学习到的模型不能够泛化出其他房屋的特点。一个假设的泛化误差（generalization error）指它在不属于训练集上的样例上的预测误差。



最左边的模型和最右边的模型都有着大泛化误差，不过它们产生大泛化误差大原因并不一样，如果说 y 与 x 的关系不是线性的，那么即使说我们的线性模型能够拟合绝大多数的训练数据，该模型依旧不能够捕捉到训练集的数据结构。于是我们定义一个模型的偏倚（bias）作为其泛化误差，即使模型能够拟合绝大多数的训练数据。因此，线性模型就会有比较明显的偏倚，也就是欠拟合。

除了偏倚，泛化误差还有另一个部分就是一个模型拟合过程的方差（variance）。具体来说就是，经过 5 次多项式拟合，最后边的模型会有很大的风险恰巧在小的有限的训练集中能够拟合它的数据，但是并不能广泛地表现出 x 与 y 的关系。这就是说，我嘛在训练集中碰巧得到了一个微微大于平均值的房屋，和一个微微小于平均价格的房屋。为了拟合这样的虚假的训练集，我们再次得到了一个大泛化误差的模型。这种情况下，我们称该模型有大的方差。

更多的情况是，我们需要在偏倚（bias）与方差（variance）中进行折中。如果我们的模型太简单，并且参数少，那就很容易会产生大偏倚（小方差）；如果模型复杂且有许多参数，那么则会有大的方差（小偏倚）。在上面的例子中，拟合一个二次方程要比它们两个的效果更好。

2 Preliminaries

在这一系列讲义中，我们将要开始进入学习理论。这个话题除了其本身的趣味性外，还能帮助我们能够对不同的设定中最好地应用学习算法。我们需要寻求一些问题的答案：首先，我们能够形式化 bias 和 variance 的折中吗? 这个问题最终会带领我们讨论关于模型选择的方法，例如自动地决定在一个训练集上拟合几次。第二个问题是，在机器学习中我们关注泛化误差，但是大多数的学习算法在训练集上是适合它们的模型的。为何在训练集上表现得好就反映了泛化误差? 具体得说就是，我们是否能够将在训练集上误差和泛化误差联系起来? 第三个问题就是，是否存在一些条件，在这些条件下就能证明出学习算法能够有很好的效果?

以下，先引入两个虽然简单但是十分有用的辅助定理:

Lemma 2.1（*The union bound)* 令 $A_1, A_2, ..., A_k$ 是 k 个不同的事件（它们可能不是互相独立的），那么

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

在概率论中 union bound 通常被视作是一个公理，它表达的其实就是说 k 个事件中任何一个事件发生的概率最多是这 k 个不同事件的概率之和。

Lemma 2.1（*Hoeffding* 不等式）令 $Z_1, ..., Z_m$ 是从一个伯努利分布 ϕ 中的 m 个独立同分布 *iid* 的随机变量，例如说 $P(Z_i = 1) = \phi$，并且 $P(Z_i = 0) = 1 - \phi$。令 $\hat{\phi} = (1/m)\sum_{i=1}^{m} Z_i$ 是这些随机变量的平均值，并且令任意的 $\gamma > 0$。那么就有

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

该定理（在计算理论中也被称为 Chernoff bound）说，如果我们用 m 个伯努利 ϕ 的随机变量的均值 $\hat{(\phi)}$ 来表示我们估计的 ϕ，那么随着 m 的增大，我们估计的参数值离真实值误差很大的概率会变小。另一种说法就是，如果你有一个有偏差的硬币，正面向上的概率是 ϕ，那么如果你投掷 m 次的话，用分数计算其正面向上的次数，那么会得到一个对 ϕ 的高概率的估计值 (如果 m 是很大的话)。

利用以上的两个辅助定理，我们就能证明出计算理论中更深层次和重要的结论。简单起见，我们考虑二分类问题，也就是标签值 $y \in \{0,1\}$，当然它们也概括了其他的诸如回归问题，多分类问题。

我们假设给定一个 m 大小的训练集 $S = (x^{(i),y^{(i)}}, ; i = 1, ..., m$，并且这里的训练样本 $(x^{(i)}, y^{(i)})$ 是对概率分布 D 的独立同分布 iid。对于一个假设 h，我们定义其训练误差（training error，同时也叫做 empirical risk 或者 empirical error）为

$$\hat{\varepsilon}(h) = \frac{1}{m}\sum_{i=1}^{m} 1\{h(x^{(i)}) \neq y^{(i)}\}$$

这其实就是 h 的错误分类的样例比上总的训练集。若是要在训练集 S 上加上 $\hat{\varepsilon}(h)$ 的依赖的话，就可以写成 $\hat{\varepsilon}_S(h)$。我们同样可以定义出泛化误差为

$$\varepsilon(h) = P_{(x,y)\sim D}(h(x) \neq y)$$

这是，如果我们从一个分布 D 中得到一个样本 (x, y)，h 将会错误分类它的概率。

需要注意的是我们假设训练数据和评价我们假设的数据都是相同的分布 D（这在泛化误差的定义中）。这与 PAC 假设也有关。

考虑线性分类的设定，使得 $h_\theta(x) = 1\{\theta^T x > 0\}$。什么合理的方法可以拟合参数 θ 呢？其中一个方法就是尝试去最小化训练误差然后选择 $\hat{\theta} = arg\min_{\theta} \varepsilon(\hat{h}_\theta)$

我们将这个过程称作是经验风险最小化（empirical risk minimization），然后学习算法输出的假设结果是 $\hat{h} = h_{\hat{\theta}}$。我们会在这个讲义中多关注我们认为的最基本的学习算法 ERM 算法。（类似于逻辑回归的算法也可以视作是 ERM 相近的算法）。

学习理论对概括假设的具体的参数或者我们是否要用线性分类都是有用的。我们定义假设类 ***hypothesis class*** $\mathcal{H}$, 对于线性分类来说，$\mathcal{H} = \{h_\theta : h_\theta(x) = 1\{\theta^T x \geq 0\}, \theta \in \mathbb{R}^{n+1}\}$ 因此是带有线性决策边界 $\mathcal{X}$ 上的所有分类器的集合。更广泛地说就是，如果我们在学习神经网络，那么我们就能够使得 $\mathcal{H}$ 是由神经网络架构代表的分类器的集合。

经验风险最小化被认为是在分类函数 $\mathcal{H}$ 上的最小化，因此学习算法挑

选这样的假设:

$$\hat{h} = arg\min_{h\in\mathcal{H}} \hat{\epsilon}(h).$$

3 The case of finite $\mathcal{H}$

有这样的一个学习问题，我们有一个有限的由 k 个假设组成的假设类 $\mathcal{H} = \{h_1, ..., h_k\}$。因此，$\mathcal{H}$ 其实就是从 $\mathcal{X}$ 映射到 $\{0, 1\}$ 的 k 个函数，并且经验风险误差则是选择这 k 个函数中的某个 $\hat{h}$ 使得有最小的训练误差。

我们想要保证 h 的泛化误差。一般我们的策略将分为 2 步：首先，我们需要证明 $\hat{\epsilon}(h)$ 是对任意 h 的 $\epsilon(h)$ 的可靠估计。其次，我们将证明，它会得到在 $\hat{h}$ 上的泛化误差的上界。

我们令 $h_i \in \mathcal{H}$。考虑一个如下定义的伯努利分布的随机变量 Z，样本由分布 D 生成 $(x, y) \sim \mathcal{D}$。然后我们设定 $Z = 1\{h_i(x) \neq y\}$，令 Z 用来指示是否该样本被 h_i 错误分类。同样地，我们要定义 $Z_j = 1\{h_i(x^{(j)}) \neq y^{(j)}\}$。由于我们的训练集都是来自独立同分布 $\mathcal{D}, \mathcal{Z}$, 因此 Z_j 也有相同的分布。

我们知道在一个随机的样本上的错误分类的概率为 $\epsilon(h)$, 它也就是随机变量 $\mathcal{Z}$ 的期望值。那么训练误差就可以写成

$$\hat{\epsilon}(h_i) = \frac{1}{m}\sum_{m}^{j=1} Z_j$$

。因此 $\hat{\epsilon}(h_i)$ 就是 m 个随机变量 $\mathcal{Z}_j$ 的平均值。而 $\mathcal{Z}_j$ 是来自平均值为 $\epsilon(h_i)$ 的伯努利分布的独立同分布。因此就有了 Hoeffding 不等式，

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

这就说明了假设 m 很大的话，对于一个 h_i, 训练误差将会有很大的概率接近泛化误差。然后我们并不简单得只保证让一个 h_i 的 $\epsilon(h_i)$ 有很大概率接近 $\hat{\epsilon}(h_i)$, 而是想证明对于任意一个 h 该结论都争取。令 A_i 代表着 $|\epsilon(h_i - \hat{\epsilon}(h_i))| > \gamma$ 的事件。由于我们已经得到了对于任意一个 A_i, $P(A_i) \leq 2exp(-2\gamma^2 m)$ 成立了。因此利用上面的 The union bound 定理，

我们有:

$$\begin{aligned} P(\exists h \in \mathcal{H}.|\epsilon(h_i) - \hat{\epsilon}(h_i)| > \gamma) &= P(A_1 \cup ... \cup A_k) \\ &\leq \sum_{i=1}^{k} P(A_i) \\ &\leq \sum_{i=1}^{k} 2exp(-2\gamma^2 m) \\ &2k\ exp(-2\gamma^2 m) \end{aligned}$$

两边用 1 减，就能得到

$$\begin{aligned} P(\neg\exists h \in \mathcal{H}.|\epsilon(h_i) - \hat{\epsilon}(h_i)| > \gamma) &= P(\forall h \in \mathcal{H}.|\epsilon(h_i) - \hat{\epsilon}(h_i)| \leq \gamma)) \\ &\geq 1 - 2k\ exp(-2\gamma^2 m) \end{aligned}$$

那么对于所有的 $h \in \mathcal{H}$，$\epsilon(h)$ 与 $\hat{\epsilon}(h)$ 的间距在 γ 以内的概率至少是 $1 - 2k\ exp(-2\gamma^2 m)$，也就是说假设集内任意假设函数的训练误差和泛化误差的的接近程度小于一个常数的概率那么对于所有的 $h \in \mathcal{H}$，$\epsilon(h)$ 与 $\hat{\epsilon}(h)$ 的间距在 γ 以内的概率，这就叫做一致收敛性 (uniform convergence)，因为这是一个同时满足所有属于 $\mathcal{H}$ 的 h。

在上面的不等式中有三个参数，样本量 m、误差阈值 γ 和误差概率。我们可以通过任意两个得到第三个。

例如说这样的一个问题: 给定 γ 和 $\delta > 0$，多大的 m 能够保证训练误差和泛化误差相差不超过 γ 的概率至少为 $1-\delta$? 通过使得 $\delta = 2k\ exp(-2\gamma^2 m$，解出 m，我们发现如果

$$m \geq \frac{1}{2\gamma^2} log \frac{2k}{\delta}$$

，那么对于所有的 $h \in \mathcal{H}$，$|\epsilon(h_i) - \hat{\epsilon}(h_i)| \leq \gamma$ 的概率至少为 $1 - \delta$。

这样一个界就告诉我们为了能够确保一个理想的效果，我们需要多少训练样本。那么为了能够得到理想级别的性能，算法需要的训练集的大小 m 也称作为算法的样本复杂度（algorithm's sample complexity）。

上面这个界的一个关键的性质是所需要的训练样本的容量跟 k 的对数相关，而 k 是在 $\mathcal{H}$ 中的假设的数量。在之后会提到它的重要性。

相似地，如果保持 m 和 δ 不变，利用前面的公式就能计算出 γ，那么对于所有的 $h \in \mathcal{H}$，且概率为 $1 - \delta$，则有

$$|\hat{\epsilon}(h_i) - \epsilon(h_i)| \leq \sqrt{\frac{1}{2m} log \frac{2k}{\delta}}.$$

我们现在假设，一致收敛性成立，即对于所有的 $h \in \mathcal{H}$，有 $|\epsilon(h_i) - \hat{\epsilon}(h_i)| \leq \gamma$，那么我们该如何证明挑选 $\hat{h}$ 的学习算法的泛化误差?

定义 $h^* = arg\ min_{h \in \mathcal{H}} \hat{\epsilon}(h)$ 为 $\mathcal{H}$ 最有可能的假设。注意 h^* 表示在集合中使得泛化误差最小的那个假设函数，我们有:

$$\begin{aligned} \epsilon(\hat{h}) &\leq \hat{\epsilon}(\hat{h}) + \gamma \\ &\leq \hat{\epsilon}(h^*) + \gamma \\ &\leq \epsilon(h^*) + 2\gamma \end{aligned}$$

不等式第一行用的是一致收敛性，即 $|\epsilon(\hat{h}) - \hat{\epsilon}(\hat{h})| \leq \gamma$。第二行则是因为 $\hat{h}$ 被用于最小化 $\hat{\epsilon}(h)$，因此对于所有的 h，$\hat{\epsilon}(\hat{h}) \leq \hat{\epsilon}(h)$，那么特殊地自然有 $\hat{\epsilon}(\hat{h}) \leq \hat{\epsilon}(h^*)$。第三行再次利用了一致性收敛，即 $\hat{\epsilon}(h^*) \leq \hat{\epsilon}(h^*) + \gamma$。因此，也就是说，如果一致性收敛成立的话，$\hat{h}$ 的泛化误差最多比在 $\mathcal{H}$ 中最好的假设要差 2γ!

因此，就有了下面的定理:

Theorem 3.1 令 $|\mathcal{H} = k|$，任意的 m, δ 不变，由于概率至少为 $1 - \delta$，我们有

$$\epsilon(\hat{h}) \leq (\min_{h \in \mathcal{H}} \epsilon(h)) + 2\sqrt{\frac{1}{2m} log \frac{2k}{\delta}}$$

证明的话，通过令 γ 等于那个根号项，利用之前的一致性收敛概率至少为 $1-\delta$，以及之前提到的，一致性收敛导出的 $\epsilon(h)$ 最多比 $\epsilon h^* = min_{h \in \mathcal{H}} \epsilon h$ 大 2γ。

这就量化了之前我们提到的，在模型选择时，偏倚与方差之间的折中。具体地说就是，假设我们有一个假设类 $\mathcal{H}$，并且考虑着想要切换到某个更大的假设集 $\mathcal{H}' \supseteq \mathcal{H}$。如果我们切换到了 $\mathcal{H}'$，那么第一项 $min_h \epsilon(h)$ 只会减少（因为我们会在一个更大的集合中寻找最小的 min）。因此，通过在一个更大的假设集中学习，我们的偏倚 bias 也只会降低。然而，如果 k 增加了，那么第二个带有根号的项就会增加。这就对应了当我们利用一个更大的假设集的时候，我们的方差 variance 会增加。

通过使得 γ 和 δ 不变，像之前一样解出 m，我们同样能够得到下面的样本复杂边界：

Corollary 3.1 令 $|\mathcal{H} = k|$，任意的 δ, γ 不变。然后，由于 $\epsilon(\hat{h}) \leq min_{h \in \mathcal{H}} \epsilon(h) + 2\gamma$ 使得概率至少是 $1 - \delta$，那么它满足，

$$\begin{aligned} m &\geq \frac{1}{2\gamma^2} log \frac{2k}{\delta} \\ &= \mathcal{O}(\frac{1}{\gamma^2 log \frac{k}{\delta}}) \end{aligned}$$

4 The case of infinite $\mathcal{H}$

我们已经证明了在有限假设集下的有用的定理，但是（正如在线性分类中）包括了实数数据的假设集实际上包含了无限数量的函数。我们能得到类似的结果吗?

我们会从一个不太对的论断开始。

假设有一个 $\mathcal{H}$，并且有 d 个实数参数。计算机中通常用 IEEE 双精度浮点数（C 语言中的 double 类型），也就是说我们的学习算法使用 64 位来表示一个浮点数的。因此，我们的假设集最多由 2^{64d} 个不同的假设组成。那么由上面的一个 Corolary 定理，我们发现，确保 $\epsilon(\hat{h}) \leq \epsilon(h^*) + 2\gamma$，以及概率至少是 $1 - \delta$，则有 $m \geq \mathcal{O}(\frac{d}{r^2} log \frac{1}{\delta}) = \mathcal{O}_{\gamma,\delta}(d)$。（下表 γ, δ 表示的是大 O 依赖这两个参数）。因此，所需要的训练样本的数量与模型参数成线性关系。

由于事实上的 64 位浮点数使得结果不能满足，但是结论还是大致正确的：如果我们想要做的是最小化训练误差，那么为了能够用带有 d 个参数的假设集学习出好的结果，我们需要在 d 上顺序的线性数量的训练集。

（需要注意的是，这个结论是由使用了经验风险最小化的算法证明出的。这里需要知道其实在许多非 ERM 学习算法建立好的理论依旧是一个活跃的研究领域）

另一部分就是由于它依赖 $\mathcal{H}$ 的参数，无限集会不会不满足? 直觉上它并不满足，我们之前已经表过，

$h_\theta(x) = 1\{\theta_0 + \theta_1 x_1 + ... + \theta_n x_n \geq 0\}$，

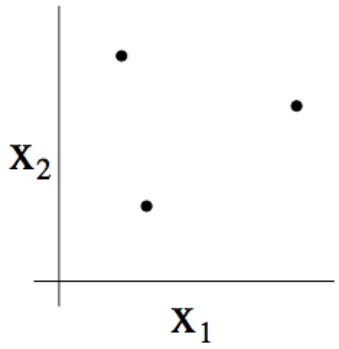
有 n+1 个参数，但是它可以写成 $h_{u,v} = 1\{(u_0^2 - v_0^2) + (u_1^2 - v_1^2)x_1 + ...(u_n^2 - v_n^2)x_n \geq 0\}$，这里有 2n+2 个参数 u_i, v_i。这两个等式都定义了同样的 $\mathcal{H}$：n 维的线性分类集。

为了得出一更满意的结论，让我们再定义一些东西。

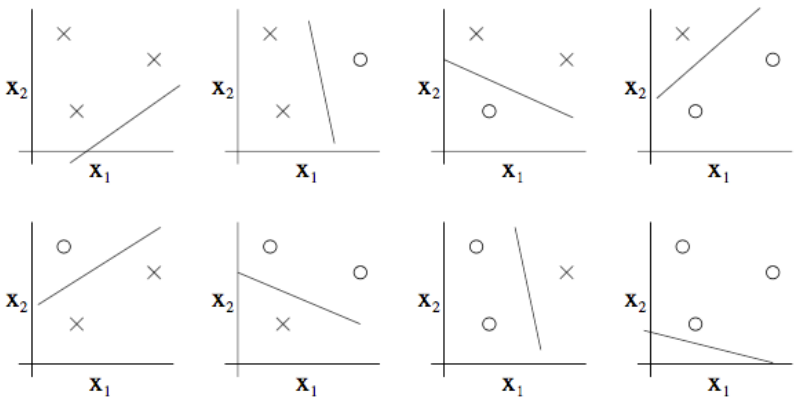
给定一个集合 $S = \{x^{(i)}, ..., x_{(d)}\}$（与训练集无关），其中 $x_{(i)} \in \mathcal{X}$，那么，如果 $\mathcal{H}$ 能够识别 S 上的任何标签，我们就说 $\mathcal{H} shatters S$，也就是 $\mathcal{H}$ 散列了 S。例如，如果对于任意的标签集合 $\{y^{(1)}, ..., y^{(d)}\}$，存在一些 $h \in \mathcal{H}$ 以致于对于所有的 $i = 1, ..., d, h(x^{(i)}) = y^{(i)}$。

给定一个假设集 $\mathcal{H}$，我们定义它的 VC 维（Vapnik-Chervonenkis dimension），$VC(\mathcal{H})$ 表示被 $\mathcal{H}$ 散列的最大的集合的大小。（如果 $\mathcal{H}$ 能够散列绝对大小的集合，那么 $VC(\mathcal{H}) = \infty$）。

例如下面的三个点集：



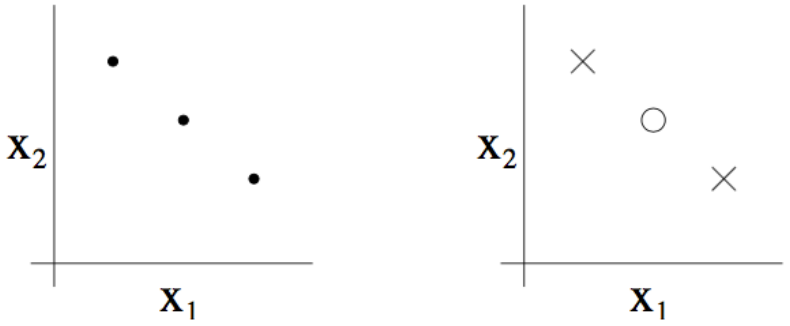
在两维（$h(x) = 1\{\theta_0 + \theta_1 x_1 + \theta_2 x_2 \geq 0\}$）中的线性分类的集合 $\mathcal{H}$ 是否能够散列上面的集合？那么答案是能够的。具体地说，我们能偶看到这些点标签的所有的八种可能，我们可以找到一个线性分类在它们中包含了“零训练误差”：



还能看出，4 个点的集合就不能被该假设散列了。那么 $\mathcal{H}$ 能散列的大

小为 3 的集合，因此，$VC(\mathcal{H}) = 3$。需要注意的是即使是 3 也不能保证它能散列。

例如，如果我们的三个点是在一条直线上的，那么就无法找到一个线性的分离器来分离它们的标签类别：



换句话说，在 VC 维的定义下，为了证明 $VC(\mathcal{H})$ 至少为 d，我们需要至少有一个能被 $\mathcal{H}$ 散列的发小为 d 的集合。

Theorem 4.1 *令 $\mathcal{H}$ $d = VC(\mathcal{H})$，至少为 $1-\delta$ 的概率，对于所有的 $h \in \mathcal{H}$，有*

$$|\epsilon(h) - \hat{\epsilon}(h) \leq O(\sqrt{\frac{d}{m}log\frac{m}{d} + \frac{1}{m}log1\delta})$$

因此，由于至少概率为 $1-\delta$，则我们有

$$\epsilon(\hat{h}) \leq \epsilon(h^*) + O(\sqrt{\frac{d}{m}log\frac{m}{d} + \frac{1}{m}log1\delta})$$

换句话说，如果一个分类集有有限的 VC 维，那么当 m 变大时会发生一致性收敛。这就跟之前一样，就 $\epsilon(h^*)$ 而言给定了 $\epsilon(\hat{h})$ 一个界。我们就有以下的必然结果：

Corollary 4.1 *当对所有 $h \in \mathcal{H}$，$|\epsilon(h) - \hat{\epsilon}(h) \leq \gamma$，最小概率值为 $1-\delta$ 成立，那么它就满足 $m = \mathcal{O}_{\gamma,\delta}(d)$。*

换句话说，就是想要在 $\mathcal{H}$ 上学习好的训练样本的数量在 $\mathcal{H}$ 的 VC 维上是线性的。这就意味着对于大多数的假设集而言，VC 维（假设是一个合理的参数化）在参数数量上大致是呈线形的。

总而言之，我们有这样的一个结论：（对于一个想要最小化训练误差的算法而言）所需要的训练集的数量与 $\mathcal{H}$ 的参数的数量大致是成线性关系的。