

Learning Theory

@author Jimmy

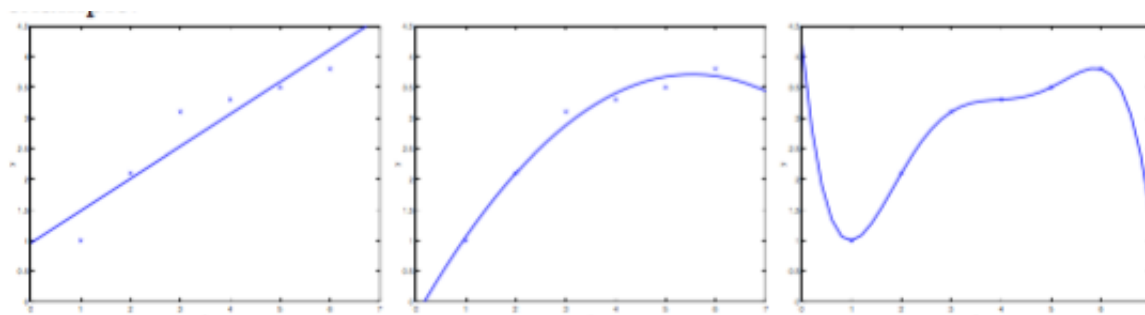
2016年3月6日

简介

截止目前，我们已经讨论了几种常用的机器学习算法，了解了它们学习的步骤。但机器学习的应用背景是多种多样的，做实际工程必须学会如何根据具体的问题评估一个学习模型的好坏，了解如何合理地选择模型、提取特征和对参数调优。通过学习learning theory，我们能够获得一些指导性的结论。

1 偏差\方差权衡

当我们在讨论线性回归时，用来拟合数据的模型可能简单如 $y = \theta_0 + \theta_1 x$ ，或许复杂点如 $y = \theta_0 + \theta_1 x + \dots + \theta_5 x^5$ 。在以下例子中，我们可以看到不同的模型对相同数据的拟合图形。



最左边的图是一次模型，中间的图是二次模型，最右边的图是五次模型。在这些模型中，我们希望得到第二个，因为它既能够描述训练数据的规律，又具有很好的泛化能力（指的是学习到的模型对位置数据预测的能力，是学习方法本质上的重要性质）。一次模型对大多数数据都不能够正确拟合，这种情况称为欠拟合（underfitting），这样的模型偏差（bias）很大；五次模型虽然100%拟合了训练数据，但是模型过于灵活，并不能很好的预测未知数据，这种情况称为过拟合（overfitting），这样的模型方差（variance）很大。在这里我无法给出偏差和方差的准确定义，不过总体而言，简单模型偏差较大，复杂模型方差较大，为了得到较小的泛化误差，我们需要对模型的复杂程度有所把控，对偏差和方差做出权衡。

2 基础知识

learning theory是用来帮助我们在面对不同场景时更好的应用学习算法的理论，它研究的是一些我们在实际应用中经常面对的问题：

1. 我们该如何对偏差和方差如何用公式统一起来，以方便我们定量的权衡？
2. 在机器学习中我们关注的是模型的泛化误差（generalization error），它衡量的是模型预测表现的好坏，但是大部分的学习算法都是在训练集上进行的，那么我们应该如何将训练集和泛化误差联系起来？
3. 我们能否证明在满足某些条件时，学习算法的表现最好？

为了给出这些问题的答案，让我们先从两个简单但非常有用的引理说起。

引理2.1（联合边界）：令 A_1, \dots, A_k 是 k 个不同的事件（不一定相互独立），那么有

$$P(A_1 \cup \dots \cup A_k) \leq p(A_1) + \dots + p(A_k)$$

这个引理是概率论的内容，无需多说。

引理2.2（Hoeffding不等式）：令 Z_1, \dots, Z_m 是 m 个独立同分布（independent and identically distributed, iid）的随机变量，且服从参数为 ϕ 的Bernoulli分布。令 $\hat{\phi} = (1/m) \sum_{i=1}^m Z_i$ ， $\gamma > 0$ 为一个常数，那么有

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

这个引理在learning theory中也称为Chernoff边界，它给出了用Bernoulli随机变量均值来估计参数 ϕ 的误差上界。这句话好拗口，意思就是Bernoulli分布的参数 ϕ 与其估计值 $\hat{\phi}$ （所有变量的平均值，这些变量服从以 ϕ 为参数的Bernoulli分布）之间的误差有上界。可以看到，当 m 取值很大时误差大于 γ 的概率接近于0，这也是误差上界的意义所在。

利用这两个引理我们可以证明learning theory领域里的许多重要结论。接下来我们将介绍几个概念，为了描述上的方便，我们将研究的范围限制为二分类问题，其中目标变量 $y \in \{0, 1\}$ 。

给定训练集 $S = \{(x^{(i)}, y^{(i)}) ; i = 1, \dots, m\}$ ，假设样本 $(x^{(i)}, y^{(i)})$ 独立同分布，且服从于分布 D 。对于一个模型假设 h ，我们定义训练误差（在learning theory中也称为经验误差）为

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

当我们想明确指出训练误差对训练集 S 的依赖关系时，我们也会把它写成 $\hat{\epsilon}_S(h)$ 。我们定义泛化误差为

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

它的含义是服从 D 分布的新样本 (x, y) 被错误分类的概率。假设我们采用了线性分类模型，且假设 $h_\theta(x) = 1\{\theta^T x \geq 0\}$ ，我们学习参数 θ 的一种方式是最小化训练误差，然后选择

$$\hat{\theta} = \arg \min_{\theta} \hat{\epsilon}(h_\theta)$$

作为学习的结果，我们称这个过程为经验风险最小化（empirical risk minimization, ERM）。接下来，让我们从具体的参数和问题中抽离出来，将学习参数的过程转变成学习模型的过程。定义假设类 H 是被用于学习的分类器的集合，那么ERM可以视为在一系列函数集 H 上进行最小化的过程，这时我们取

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

作为训练集上的最优假设，因为 \hat{h} 的训练误差最小。

3 有限假设类

令假设类 $\mathcal{H} = \{h_1, \dots, h_k\}$ 中包含 k (k 不等于无穷) 个假设, 其中

$h_i: \mathcal{X} \mapsto \{0, 1\}; i = 1, \dots, k$, 且我们通过ERM学习到假设 \hat{h} 。本节接下来的内容是按以下两个部分进行的:

1. 证明训练误差 $\hat{\epsilon}(h)$ 是泛化误差 $\epsilon(h)$ 的一个可靠估计
2. 证明 \hat{h} 的泛化误差 $\epsilon(\hat{h})$ 有上界

取任意确定的假设 $h_i \in \mathcal{H}$, 并且定义 $Z_j = 1\{h_i(x^{(j)}) \neq y^{(j)}\} \in \{0, 1\}$ 。 Z_j 表达的是 h_i 是否对样本 $(x^{(j)}, y^{(j)})$ 分类错误, 且分类错误的概率为 $p(Z_j = 1) = \epsilon(h_i)$ 。(注: 由于随机变量 Z 满足 $Z = 1\{h_i(x) \neq y\}$, 故 $E(Z) = P_{(x,y) \sim D}(h_i(x) \neq y) = \epsilon(h_i)$, 又显然有 $Z \sim \text{Bernoulli}$, 故 $p(Z_j = 1) = \epsilon(h_i)$ 。) 此时, 训练误差为

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

应用Hoeffding不等式得

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

对于确定的 h_i , 当 m 很大时, 训练误差和泛化误差非常接近 (准确来讲是两者差值的超过 γ 的概率很小), 对任意 $h \in \mathcal{H}$ 会得到什么结论呢? 令 A_i 表示事件 $|\epsilon(h_i) - \hat{\epsilon}(h_i)| > \gamma$, 我们已经证明了对任意确定的 A_i , $P(A_i) \leq 2 \exp(-2\gamma^2 m)$ 结果为真, 由联合边界引理得

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

也就是说

$$P(\text{not } h \in \mathcal{H}. |\epsilon(h_i) - \hat{\epsilon}(h_i)| > \gamma) \geq 1 - 2k \exp(-2\gamma^2 m)$$

这个结论称为一致收敛 (uniform convergence), 因为它对任意 $h \in \mathcal{H}$ 都成立。在上述的讨论中, 给定 m, γ 可得概率 $P(h \in \mathcal{H}, |\epsilon(h) - \hat{\epsilon}(h)| > \gamma)$ 的关系式。我们可得到此关系式的变形形式。

例如, 给定 γ 和 $\delta > 0$, 我们需要多大的样本才有

$P(\text{any } h \in \mathcal{H}. |\epsilon(h_i) - \hat{\epsilon}(h_i)| \leq \gamma) \geq 1 - \delta$ (关系式*)? 要让关系式*成立满足 $\delta \geq 2k \exp(-2\gamma^2 m)$ 即可, 解得

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

这个结论告诉我们不论我们想让训练误差和泛化误差多接近（ γ 足够小），我们都可以通过控制训练集的大小或者说为样本复杂度（sample complexity）来达到这一点，本节开头的part1得证。

再例如，给定 m, δ ，要使关系式*成立， γ 需满足

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

记 $h^* = \operatorname{argmin}_{h \in H} \epsilon(h)$ 为假设类中的最优假设，那么

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

下面让我们将目前得到的所有结论综合到我们的定理3.1中。

定理3.1： 令 $|H| = k$ ，且 m, δ 为任意常数，那么关系式*成立的充分条件是

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

这个定理给出了 $\epsilon(\hat{h})$ 的上界，本节part2得证。现在让我们回过头讨论在模型选择中的偏差/方差权衡。假设我们已有一个假设类 H ，并考虑选择一个更大的假设类 $H' \supseteq H$ （更大意味着假设中包含次数更高的项，假设空间的维度更高）。如果我们使用假设类 H' ，那么上式第一项会减小（在更大的假设函数集中寻求最小化），也就是说通过学习更大的假设类，偏差会减小；然而，这时候由于假设数目 k 增大，上式的第二项会增大，也就是说学习更大的假设类会使方差增大。

推论： 令 $|H| = k$ ，且 δ, γ 为任意常数，那么 $P\left(\epsilon(\hat{h}) \leq \min_{h \in H} \epsilon(h) + 2\gamma\right) \geq 1 - \delta$ 的充分条件是

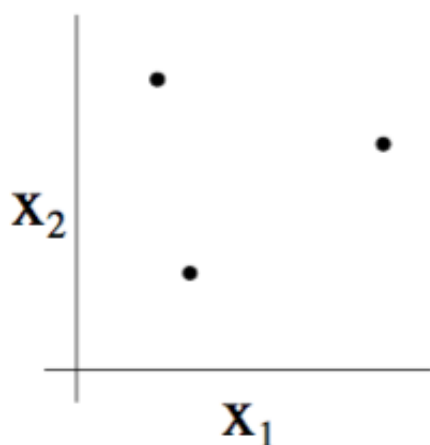
$$\begin{aligned} m &\geq \frac{1}{2\gamma^2} \log \frac{2k}{\delta} \\ &= O\left(\frac{1}{\gamma^2} \log \frac{k}{\delta}\right) \end{aligned}$$

4 无限假设类

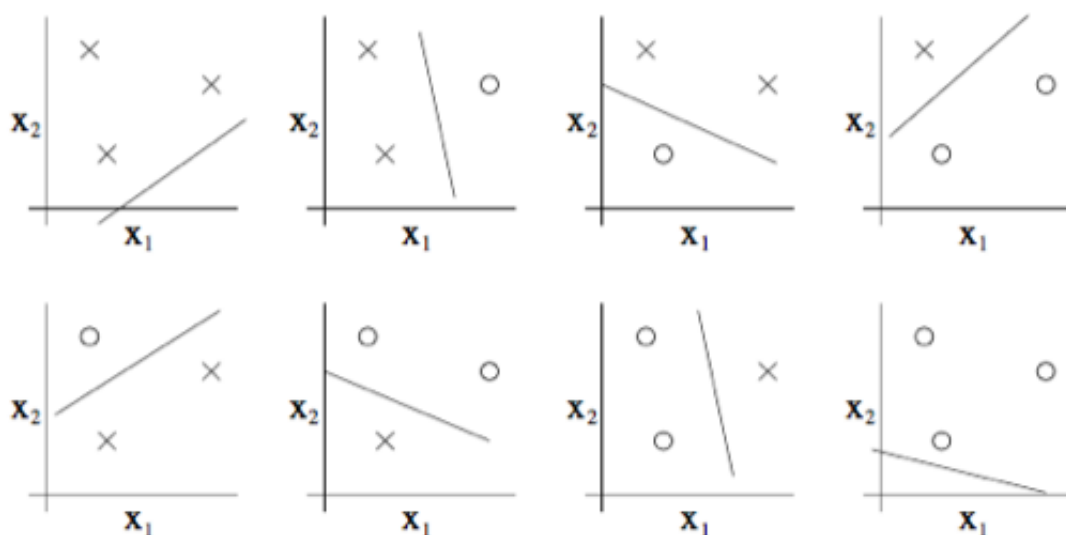
在真实情况下，假设类里包含的假设是无穷的，在这种情况下我们能不能得到与有限假设类相似的结论？在开始介绍之前，让我们给出一些定义。

1. 给定集合 $S = \{x^{(i)}, \dots, x^{(d)}\}$ （与训练集无关），其中点 $x^{(i)} \in \mathcal{X}$ ，如果 H 能够划分 S 的任意一种分布，我们就说 H 划分了 S 。

2. 给定一个假设类 \mathcal{H} ，我们定义它的VC维（Vapnik-Chervonenkis dimension）为它所能划分的最大的集合的大小（即集合所包含点的数目），写作 $VC(\mathcal{H})$ 。



对上图所示的大小为3的集合 S ，二维线性分类器 ($h(x) = 1\{\theta_0 + \theta_1 x_1 + \theta_2 x_2 \geq 0\}$) 集合 \mathcal{H} 能够划分吗？答案是可以，以下展示的就是 \mathcal{H} 如何对 S 进行划分的。



事实上，在二维空间上的假设类 \mathcal{H} 所能划分的最大集合大小为3，也就是说 $VC(\mathcal{H}) = 3$ （并非任何大小为3的集合都能被二维假设类）；更一般地，对于 n 维假设类 \mathcal{H} 有， $VC(\mathcal{H}) = n + 1$ 。

下面给出在learning theory领域可能最重要的定理。

定理4.1: 给定 \mathcal{H} ，且 $d = VC(\mathcal{H})$ ，那么关系式*成立需满足对所有 $h \in \mathcal{H}$ 有

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

并且

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

这说明，当一个假设类的VC维有限时，如果m很大也会得到一致收敛结论。

推论：关系式*成立的充分条件是 $m = O_{\gamma, \delta}(d)$ 。

这里大O表示对常数 γ, δ 依赖的时间复杂度。