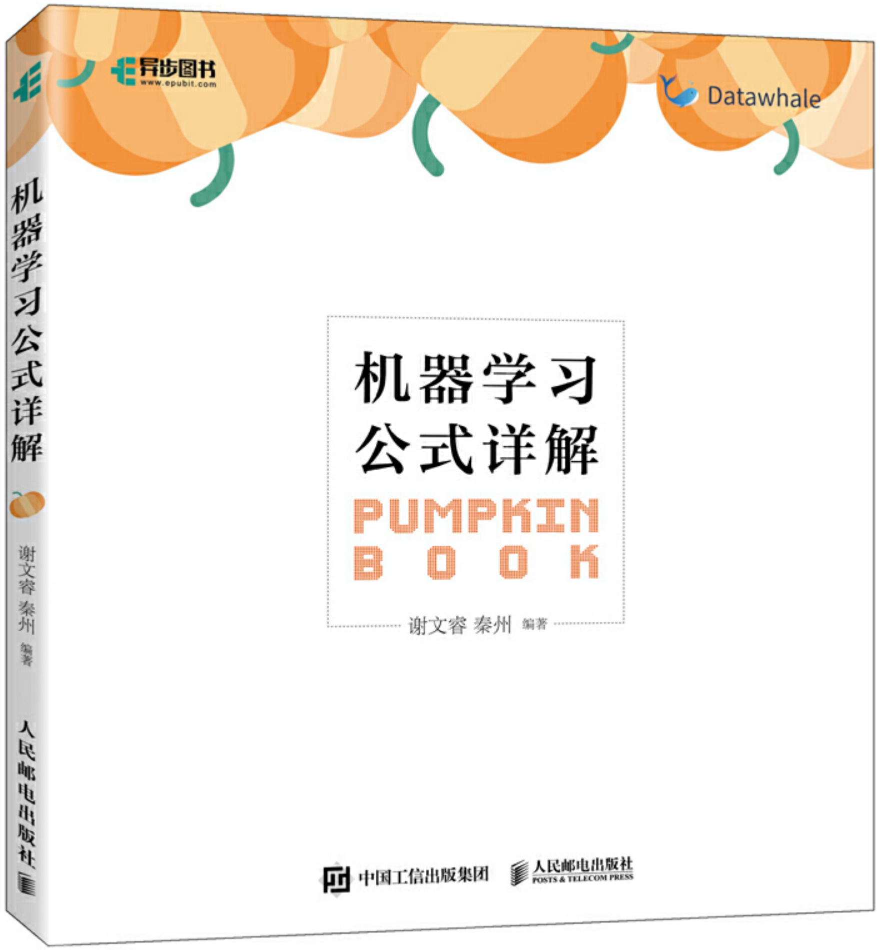

《机器学习公式详解》（南瓜书）

第12章 计算学习理论（上）

本节主讲：秦州

本节大纲

南瓜书对应章节：12.1 12.2 12.3

1. 计算学习理论概念
2. PAC学习
3. 有限空间假设

计算学习理论

何为“计算学习”

机器学习 -- 通过“计算”来进行学习的科学

计算学习理论 -- 为机器学习提供理论指导，分析学习任务的困难度，指导算法设计

符号：

给定样例集 $D=\{(x_1,y_1),(x_2,y_2),\ldots,(x_m,y_m)\}$, $x_i\in\mathcal{X}$, 假设 $\mathcal{X}$ 中的所有样本都是从一个隐含未知的分布 $\mathcal{D}$ 中独立同分布得采样得到的。

令 h 为从 $\mathcal{X}$ 到 $\mathcal{Y}$ 的一个映射, 其泛化误差为

$$E(h;\mathcal{D})=P_{\boldsymbol{x}\sim\mathcal{D}}(h(\boldsymbol{x})\neq y),$$

h 在 D 上的经验误差为

$$\widehat{E}(h;D)=\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\left(h\left(\boldsymbol{x}_i\right)\neq y_i\right)$$

计算学习理论-续

由于 D 是 $\mathcal{D}$ 的独立同分布采样, 因此 h 的经验误差的期望等于其泛化误差。我们将 $E(h;\mathcal{D})$ 和 $\widehat{E}(h;D)$ 分别简记为 $E(h)$ 和 $\widehat{E}(h)$. 令 ϵ 为 $E(h)$ 的上限, 即 $E(h) \leqslant \epsilon$; 我们通常用 ϵ 表示预先设定的学得模型所应满足的误差要求, 亦称 “误差参数”。

在计算学习理论中，我们将会研究经验误差与泛化误差之间的逼近程度. 若 h 在数据集 D 上的经验误差为 0 , 则称 h 与 D 一致, 否则称其与 D 不一致。 对任意两个映射 $h_1, h_2 \in \mathcal{X} \rightarrow \mathcal{Y}$, 可通过其 “不合” (disagreement)来度量它们之间的差别:

$$d\left(h_1, h_2\right) = P_{\boldsymbol{x} \sim \mathcal{D}}\left(h_1(\boldsymbol{x}) \neq h_2(\boldsymbol{x})\right).$$

PCA学习

PAC（概率近似正确理论）：

“概念” (concept)，记作c，表示从样本空间 $\mathcal{X}$ 到标记空间 $\mathcal{Y}$ 的映射。如果$c(\boldsymbol{x})=y$ 成立, 则称 c 为目标概念。

“概念类” (concept class)，记作$\mathcal{C}$，表示所有我们希望学得的目标概念所构成的集合。

“假设”（hypothesis），记作h，也是从样本空间 $\mathcal{X}$ 到标记空间 $\mathcal{Y}$ 的映射，h是学习算法学到的。

“假设空间” (hypothesis space), 记作 $\mathcal{H}$，表示给定学习算法 $\mathfrak{L}$所考虑的所有可能假设的集合。

“可分的” (separable), 又称 “一 致的” (consistent)：若目标概念 $c \in \mathcal{H}$, 则 $\mathcal{H}$ 中存在假设能将所有示例按与真实标记一致的方式完全分开, 我们称该问题对学习算法 $\mathfrak{L}$ 是可分的; 反之是“不可分的”或者叫“不一致的”。

PAC学习2

PAC 辨识 (PAC Identify): 对 $0 < \epsilon, \delta < 1$, 所有 $c \in \mathcal{C}$ 和分布 $\mathcal{D}$, 若存在学习算法 $\mathfrak{L}$, 其输出假设 $h \in \mathcal{H}$ 满足

$$P(E(h) \leqslant \epsilon) \geqslant 1 - \delta$$

则称学习算法 $\mathfrak{L}$ 能从假设空间 $\mathcal{H}$ 中 PAC 辨识概念类 $\mathcal{C}$.
通俗来讲，学习算法 $\mathfrak{L}$ 能以较大的概率 (至少 $1-\delta$) 学得目标概念 c 的近似 (误差最多为 ϵ)。

PAC 可学习 (PAC Learnable): 令 m 表示从分布 $\mathcal{D}$ 中独立同分布采样得到的样例数目, $0 < \epsilon, \delta < 1$, 对所有分布 $\mathcal{D}$, 若存在学习算法 $\mathfrak{L}$ 和多项式函数 $\operatorname{poly}(\cdot, \cdot, \cdot, \cdot)$, 使得对于任何 $m \geqslant \operatorname{poly}(1/\epsilon, 1/\delta, \operatorname{size}(\boldsymbol{x}), \operatorname{size}(c))$, $\mathfrak{L}$ 能从假设空间 $\mathcal{H}$ 中 PAC 辨识概念类 $\mathcal{C}$, 则称概念类 $\mathcal{C}$ 对假设空间 $\mathcal{H}$ 而言是 PAC 可学习的, 有时也简称概念类 $\mathcal{C}$ 是 PAC 可学习的。其中$\mathfrak{L}$称为**PAC学习算法**，最小的m称为**样本复杂度**。

PAC学习3

PAC学习的意义：给出了一个抽象地刻画机器学习能力的框架：比如至少需要多少样本才能训练得到较好的模型。

有限假设空间--可分情形

可分情形意味着目标概念 c 属于假设空间 $\mathcal{H}$, 即 $c \in \mathcal{H}$. 给定包含 m 个样例的训练集 D, 如何找出满足误差参数的假设呢?

一种简单的学习策略: 既然 D 中样例标记都是由目标概念 c 赋予 的, 并且 c 存在于假设空间 $\mathcal{H}$ 中, 那么, 任何在训练集 D 上出现标记错误的假设肯定不是目标概念 c. 于是, 我们只需保留与 D 一致的假设, 剔除与 D 不一致的假设即可。【问题，可能产生很多等效假设】

到底需多少样例才能学得目标概念 c 的有效近似呢? 对 PAC 学习来说, 只要训练集 D 的规模能使学习算法 $\mathfrak{L}$ 以概率 $1-\delta$ 找到目标假设的 ϵ 近似即可。

有限假设空间--可分情形2

我们先估计泛化误差大于 ϵ 但在训练集上仍表现完美的假设出现的概率. 假定 h 的泛化误差大于 ϵ, 对分布 $\mathcal{D}$ 上随机采样而得的任何样例 $(\boldsymbol{x}, y)$, 有

$$\begin{aligned} P(h(\boldsymbol{x}) = y) &= 1 - P(h(\boldsymbol{x}) \neq y) \\ &= 1 - E(h) \\ &< 1 - \epsilon. \end{aligned}$$

由于 D 包含 m 个从 $\mathcal{D}$ 独立同分布采样而得的样例, 因此, h 与 D 表现一致的概率为

$$\begin{aligned} P\left(\left(h\left(\boldsymbol{x}_1\right) = y_1\right) \wedge \ldots \wedge \left(h\left(\boldsymbol{x}_m\right) = y_m\right)\right) &= (1 - P(h(\boldsymbol{x}) \neq y))^m \\ &< (1 - \epsilon)^m. \end{aligned}$$

有限假设空间--可分情形3

我们事先并不知道学习算法 $\mathfrak{L}$ 会输出 $\mathcal{H}$ 中的哪个假设, 但仅需保证泛化误差大于 ϵ, 且在训练集上表现完美的所有假设出现概率之和不大于 δ 即可:

$$P(h \in \mathcal{H} : E(h) > \epsilon \wedge \widehat{E}(h) = 0) < |\mathcal{H}|(1-\epsilon)^m < |\mathcal{H}|e^{-m\epsilon}$$

令上式不大于 δ, 即

$$|\mathcal{H}|e^{-m\epsilon} \leqslant \delta$$

可得

$$m \geqslant \frac{1}{\epsilon}\left(\ln|\mathcal{H}| + \ln\frac{1}{\delta}\right)$$

由此可知, 有限假设空间 $\mathcal{H}$ 都是 PAC 可学习的。

有限假设空间--不可分情形

Datawhale 异步社区

显然, 当 $c \notin \mathcal{H}$ 时, 学习算法 $\mathfrak{L}$ 无法学得目标概念 c 的 ϵ 近似. 但是, 当 假设空间 $\mathcal{H}$ 给定时, 其中必存在一个泛化误差最小的假设, 找出此假设的 ϵ 近似也不失为一个较好的目标. $\mathcal{H}$ 中泛化误差最小的假设是 $\arg\min_{h\in\mathcal{H}} E(h)$, 于是, 以此为目标可将 PAC 学习推广到 $c \notin \mathcal{H}$ 的情况, 这称为 “不可知学 习" (agnostic learning). 相应的, 我们有

定义 12.5 不可知 PAC 可学习 (agnostic PAC learnable): 令 m 表 示从分布 $\mathcal{D}$ 中独立同分布采样得到的样例数目, $0<\epsilon,\delta<1$, 对所 有分布 $\mathcal{D}$, 若存在学习算法 $\mathfrak{L}$ 和多项式函数 $\operatorname{poly}(\cdot,\cdot,\cdot,\cdot)$, 使得对于任何 $m \geqslant \operatorname{poly}(1/\epsilon, 1/\delta, \operatorname{size}(\boldsymbol{x}), \operatorname{size}(c))$, $\mathfrak{L}$ 能从假设空间 $\mathcal{H}$ 中输出满足式 (12.20) 的 假设 h :

$$P\left(E(h)-\min_{h'\in\mathcal{H}} E\left(h'\right) \leqslant \epsilon\right) \geqslant 1-\delta,$$

则称假设空间 $\mathcal{H}$ 是不可知 PAC 可学习的.

预告

下一节：计算学习理论（下）

西瓜书对应章节：12.4 12.5 12.6

结束语

欢迎加入【南瓜书读者交流群】，我们将在群里进行答疑、勘误、本次直播回放、本次直播PPT发放、下次直播通知等最新资源发放和活动通知。

加入步骤：

1. 关注公众号【Datawhale】，发送【南瓜书】三个字获取机器人“小豚”的微信二维码
2. 添加“小豚”为微信好友，然后对“小豚”发送【南瓜书】三个字即可自动邀请进群