第一章　PAC 可学习性

1.1　从一个例子开始

假如你发现苹果的好吃程度和它的色泽和大小貌似有某种奇妙的联系，于是想用机器学习的方法建立一个模型，使它能够根据苹果色泽和大小，来预测对苹果是否好吃. 为了建立这个模型，我们简单将苹果的两种属性抽象为 $[0,1]$ 内的评分：

- x_1 表示苹果的色泽，x_1 越大表示苹果越红.
- x_2 表示苹果的大小，x_2 越大表示苹果也越大.

因此一个苹果的外观就可以通过向量 $\mathbf{x}=[x_1;x_2]\in\mathcal{X}=[0,1]^2$ 表示. 同时，我们只关注苹果好吃与否，意味着这是个二分类问题，它的标签是 $y\in\mathcal{Y}=\{-1,1\}$，其中 $y=-1$ 表示不好吃，$y=1$ 表示好吃.

我们现在有 n 个苹果作为样本，一个爱吃苹果的张先生已经给它们打上了标签，其组成样本集 $\mathcal{S}=\{(\mathbf{x}_1,y_1),(\mathbf{x}_2,y_2),\cdots,(x_n,y_n)\}$，在这里需要注意，所有的苹果样例 $\mathbf{x}_i$ 都是独立地取自一个相同分布 $\mathcal{D}_{\mathcal{X}}$ ($\mathcal{D}_{\mathcal{X}}$ 是定义在 $\mathcal{X}$ 上的分布)，所以它们是**独立同分布的 (independent & identically distributed, i.i.d)**.

我们现在的任务是确定一个预测函数，或者说**假设 (hypothesis)**：

$$h:\mathcal{X}\mapsto\mathcal{Y}, \tag{1.1.1}$$

其接受一个苹果的属性作为输入，并预测出它是否好吃. 这种函数有无穷无尽个，我们需要找到预测得最准确的那个假设. 为了简化问题，我们认为有一个非常厉害的函数 f，对于世界上任意一个苹果: (x,y) 都有

$$f(\mathbf{x})=y. \tag{1.1.2}$$

这就是所谓**苹果的绝对真理 (the absolute truth of apples)**，我们的目标是能找到这个函数.

当然，这个真理函数是十分难求得的，面对这个绝对真理，我们能做的就是让我们所提出的假设能够尽量少犯错误，这也就引出了**误差 (error)** 的概念，它是对一个假设所犯错程度的衡量. 我们用 $\mathcal{L}_{\mathcal{D}_{\mathcal{X}},f}(h)$ 来表示在绝对真理 f 下，当数据服从分布 $\mathcal{D}_{\mathcal{X}}$ 时，假设 h 所产生的误差，这称为**泛化误差 (generalization error)**. 例如在苹果的例子中，我们可以定义这样一个泛化误差：

$$\mathcal{L}_{\mathcal{D}_{\mathcal{X}},f}(h)=\Pr_{\mathbf{x}\sim\mathcal{D}_{\mathcal{X}}}\left(h(\mathbf{x})\neq f(\mathbf{x})\right)=\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathcal{X}}}\left(\mathbb{I}(h(\mathbf{x})\neq f(\mathbf{x})\right). \tag{1.1.3}$$

表示真理 f 和假设 h 之间发生出入的概率，其中 $\mathbb{I}(\cdot)$ 为示性函数. 泛化误差反映了一个假设在整个分布上的误差，真实地度量了真理 f 和假设 h 之间的差距，所以我们希望假设 h 的泛化误差越小越好. 但是有一个问题，我们其实不知道数据的真实分布 $\mathcal{D}$，也不知道真理

函数 f 是什么，我们唯一的资料就是苹果的样本集合 $\mathcal{S}$. 因此，我们只可以计算出下面这个基于样本定义的误差：

$$\mathcal{L}_{\mathcal{S}}(h) = \frac{1}{n} \sum_{(\mathbf{x},y)\in\mathcal{S}} \mathbb{I}(h(\mathbf{x}) \neq y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(h(\mathbf{x}_i) \neq y_i). \tag{1.1.4}$$

这被称为**经验误差 (empirical error/risk)**. 根据大数定律，当样本数量很大时，经验误差将逼近于泛化误差，即

$$\lim_{n\to\infty} \mathcal{L}_{\mathcal{S}}(h) = \mathcal{L}_{\mathcal{D}_{\mathcal{X}},f}(h). \tag{1.1.5}$$

这是一个很好的性质，它给出了一种思路：要想让泛化误差尽量小，我们让经验误差尽量小，这种策略就是**最小化经验误差 (empirical risk minimization, ERM)**. 这个并不难，我们很容易能够想到这样一个假设：

$$h'(\mathbf{x}) = \begin{cases} y_i, & \text{if } \exists\, i, \text{ such that } \mathbf{x} = \mathbf{x}_i, \\ -1, & \text{otherwise.} \end{cases} \tag{1.1.6}$$

注意到，这个假设 h' 的经验误差达到了 0. 按理说这个假设应该不错，但是我们会意识到一个问题，这个假设所做的事无非是把要预测的苹果和样本集里的对比，如果发现了一模一样的苹果，就打上相同的标签，否则就预测为 -1. 这是完全的经验主义，它不可能会达到很好的结果，因为样本集 n 总是有限的、离散的，但 $\mathcal{X}$ 是连续的、不可数的，因此样本 $\mathcal{S}$ 对于 $\mathcal{X}$ 是微不足道的，所以这样的 h' 也只能保证 n 个样本点一定预测正确，这会和 f 产生很大的出入，所以泛化误差也会很大. 反过来说，要想通过这样的 h' 达到准确的预测，就必须将 $\mathcal{X}$ 中对应的每一个苹果的标签都观察到，这是不可能的，因为遍历一个不可数的连续空间 $\mathcal{X}$ 是无法做到的.

我们似乎遇到了什么问题，上面这个 h' 的经验误差很小，但泛化误差却会很大，与我们的直觉是相悖，这种情况就是所谓的**过拟合 (overfitting)**.

1.2 有限假设集的 PAC 可学习性

过拟合是经验主义常会陷入的陷阱，因为这种假设完全依赖观测的样本，无法适用于更广泛的情况. 解决这种过拟合问题的一种重要方法，就是引入**先验知识 (prior knowledge)**.

TO BE CONTINUED...