机器学习 - 周志华 - 绪论

基本概念

这里我们做一个术语上的对照，换为一些相对更数学专业的表达.

首先先说明机器学习的定义.

> **定义：机器学习**
>
> 机器学习致力于研究如何通过计算的手段，利用经验来改善系统自身的性能.在计算机系统中，“经验”通 常 以 “数据”形式存在，因此，机器学习所研究的主要内容，是关于在计算机上从数据中产生“模型 " （model）的算法，即 “学习算法” （learning algorithm）.

教材上也给出了更为形式化的定义：

> 假设用 P 来评估计算机程序在某任务类 T 上的性能, 若一个程序通过利用经验 E 在 T 中任务上获得了性能改善，则我们就说关于 T 和 P，该程序对 E 进行了学习.

实际上很简单的话来说就是：如果我们能让一个程序通过“经验”来在一些任务上表现更好，那么我们就让这个程序“学习”了.

而通常学习的经验 E 实际上就是数据集 (Dataset)，这也是为什么机器学习常常也称为**统计学习**. 数据集的载体有很多，例如表格数据：

背包品牌	背包容量	价格
adidas	8kg	300 ￥

还有图像数据等. 根据上面的表格可以有如下概念：

> **定义：特征、特征向量、特征空间**
>
> - 背包品牌、背包容量、价格等都称为**特征** (feature) 或 **属性** (attribute).
> - 特征的取值称**特征值** （注意不要和线性代数中的内容混淆）.
> - n 个特征取值结合起来成为 n 维向量后称为**特征向量**，而所有可能的特征向量张成的空间称为**特征空间**.

我们不难发现，实际上上述概念就是数理统计中样本、样本点、样本空间概念的改名而已. 因此，下面也引入数理统计当中的随机向量符号来说明：

- $\mathcal{X}$： 样本空间
- $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{id})$： 随机向量
- d： 维数

从数据中学得模型的过程称为“学习”（learning）或“练”（training），这个过程通过执行某个学习算法来完成.

对于特定的学习任务，我们通常会有输出.

> **定义：输出空间**
>
> 一般而言，我们关心的输出称为**标记**，记为 y_i，所有可能的标记张成的空间为**标记空间**或**输出空间**，记为 $\mathcal{Y}$.

如果 $\mathcal{Y}$ 是有限集（即输出仅有有限种），则任务为**分类任务**，若输出为连续值，则任务为**回归任务**.

没有免费午餐定理 (NFL)

我们接下来考虑一个问题：存不存在一个机器学习算法，使得它在任何一个任务上都是最优的？这个问题的答案是：不存在，甚至可以说，对于任意的两个学习算法，我们总可以构造出数据集，使得其中一个表现更好，另一个更差.

这个定理实际上是一个最优化理论的定理，在此我们给出定理内容.

> **定理：No Free Lunch Theorem**
>
> 对于一个学习算法 $\mathfrak{L}_a$，如果它在某些任务上表现优于 $\mathfrak{L}_b$，则一定存在某些任务，使得 $\mathfrak{L}_b$ 优于 $\mathfrak{L}_a$.

我们仅在样本空间 $\mathcal{X}$ 和假设空间 $\mathcal{H}$ 都有限的情形下的二分类问题进行证明. 那么可以计算 $\mathfrak{L}_a$ 的**训练集外误差**，我们定义训练集为 X，令 $P(h \mid X, \mathfrak{L}_a)$ 代表算法 $\mathfrak{L}_a$ 基于训练数据 X 产生假设 h 的概率，再令 f 代表我们希望学习的目标函数，记训练集外误差为 $E_{ote}(\mathfrak{L}_a \mid X, f)$.

它实际是怎么计算的？这个误差实际上就是 $h \neq f$ 的一个总概率，于是有

$$E_{ote}(\mathfrak{L}_a \mid X, f) = \sum_h \sum_{\boldsymbol{x} \in \mathcal{X} - X} P(\boldsymbol{x}) \mathbb{1}_{(h(\boldsymbol{x}) \neq f(\boldsymbol{x}))} P(h \mid X, \mathfrak{L}_a)$$

> **对公式的理解**
>
> 这个公式可以理解为全概率公式的展开，只不过展开的样本空间为 $\mathcal{X} - X$.

对二分类问题，其真实目标函数可能是任何 $\mathcal{X} \to \{0, 1\}$ 的函数，函数空间 $\{0, 1\}^{|\mathcal{X}|}$，对所有可能的 f 按均匀分布对误差求和. 考虑

$$\sum_f E_{ote}(\mathfrak{L}_a \mid X, f) = \sum_f \sum_h \sum_{\boldsymbol{x} \in \mathcal{X} - X} P(\boldsymbol{x}) \mathbb{1}_{(h(\boldsymbol{x}) \neq f(\boldsymbol{x}))} P(h \mid X, \mathfrak{L}_a)$$

我们考虑化简这个式子，首先仅有示性函数含有 f，因此可以把求和号放在内层，h 齐次，$\boldsymbol{x}$ 几乎每个项都有，因此放在最外层：

$$= \sum_{\boldsymbol{x} \in \mathcal{X} - X} P(\boldsymbol{x}) \sum_h P(h \mid X, \mathfrak{L}_a) \sum_f \mathbb{1}_{(h(\boldsymbol{x}) \neq f(\boldsymbol{x}))}$$

由于任务是二分类的，我们考虑只有一半的 f 映射结果与 h 不同，因此最内部的求和是常数，提出有

$$= \frac{1}{2} 2^{|\mathcal{X}|} \sum_{\boldsymbol{x} \in \mathcal{X} - X} P(\boldsymbol{x}) \sum_h P(h \mid X, \mathfrak{L}_a)$$

最后，遍历所有的 h，概率自然为 1.

$$= \frac{1}{2} 2^{|\mathcal{X}|} \sum_{\boldsymbol{x} \in \mathcal{X} - X} P(\boldsymbol{x})$$

但 X 是不变的！因此期望误差（条件期望）是相等的，换言之和学习方法无关. 因此对于 $\mathfrak{L}_a$ ，其表现不可能处处优于 $\mathfrak{L}_b$. □

这个定理最重要的结论是：要谈论算法的相对优劣，必须要针对具体的学习问题，因此每个学习问题都必须尝试不同的学习方法.

习题

> **T1.1**
>
> 表 1.1 中若只包含编号为 1 和 4 的两个样例，试给出相应的版本空间.

即数据集表格应该为

编号	色泽	根蒂	敲声	好瓜
1	青绿	蜷缩	浊响	是
4	乌黑	稍蜷	沉闷	否

列举出所有的假设，然后删去和样本不符的假设即可：

1. （色泽 = 青绿）∧（根蒂 = ∗）∧（敲声 = ∗）
2. （色泽 = ∗）∧（根蒂 = 蜷缩）∧（敲声 = ∗）
3. （色泽 = ∗）∧（根蒂 = ∗）∧（敲声 = 浊响）
4. （色泽 = 青绿）∧（根蒂 = 蜷缩）∧（敲声 = ∗）
5. （色泽 = 青绿）∧（根蒂 = ∗）∧（敲声 = 浊响）
6. （色泽 = ∗）∧（根蒂 = 蜷缩）∧（敲声 = 浊响）
7. （色泽 = 青绿）∧（根蒂 = 蜷缩）∧（敲声 = 浊响）

□

T1.2

与使用单个合取式来进行假设表示相比，使用“析合范式”（实际上应该在数理逻辑当中称**合取范式**和**析取范式**）将使得假设空间具有更强的表示能力，例如：

$$\begin{aligned}\text{好瓜} \longleftrightarrow &((\text{色泽} = *) \wedge (\text{根蒂} = \text{蜷缩}) \wedge (\text{敲声} = *)) \\ &\vee ((\text{色泽} = \text{乌黑}) \wedge (\text{根蒂} = *) \wedge (\text{敲声} = \text{沉闷}))\end{aligned}$$

会把青绿、蜷缩、清脆的瓜以及乌黑、硬挺、沉闷的瓜分类为好瓜，若使用最多包含 k 个合取式的析合范式来表达表 1.1 西瓜分类问题的假设空间，试估算共有多少种可能的假设.

T1.3

若数据包含噪声，则假设空间中有可能不存在与所有训练样本都一致的假设。在此情形下，设计一种归纳偏好用于假设选择.

选择可以满足最多训练样本的假设. □

T1.4

本章 1.4 节在论述没有免费的午餐定理时，默认使用了“分类错误率”作为性能度量来对分类器进行评估，若使用其他性能度量 ℓ，则式 (1.1) 将改为：

$$E_{ote}(\mathfrak{L}_a \mid X, f) = \sum_h \sum_{\boldsymbol{x} \in \mathcal{X}-X} P(\boldsymbol{x}) \ell(h(\boldsymbol{x}), f(\boldsymbol{x})) P(h \mid X, \mathfrak{L}_a)$$

试证明 NFL 仍成立.

在原证明当中，换用其他度量后我们仍能顺利达到如下步骤：

$$= \sum_{\boldsymbol{x} \in \mathcal{X}-X} P(\boldsymbol{x}) \sum_h P(h \mid X, \mathfrak{L}_a) \sum_f \ell(h(\boldsymbol{x}), f(\boldsymbol{x}))$$

由于我们还是在一个分类问题，因此 $\ell(h(\boldsymbol{x}), f(\boldsymbol{x}))$ 还是一些离散的值，这些值和学习方法无关，因此可记为常数 A 有

$$= A \sum_{\boldsymbol{x} \in \mathcal{X}-X} P(\boldsymbol{x})$$

后面的结论是一样的. □

> **T1.5**
>
> 试说明机器学习能在互联网搜索的哪些环节起什么作用.

从搜索过程进行分析

- 在向搜索引擎提交信息阶段，通过 NLP, CV 等技术提高系统对提交信息中的关键信息提取性能
- 在搜索引擎进行信息匹配阶段，提高信息匹配程度
- 在向用户进行信息展示阶段，提高信息展示顺序与用户兴趣的匹配程度.

□