Lecture 1　Statistical Learning Theory

Alex　2019 年 7 月 21 日

一　决策理论：宏观视角

统计学习理论 (Statistic Learning Theory) 可以看做是决策理论 (Decision Theory) 的子类或衍生。因此在进入统计学习理论的学习前，有必要对决策理论进行一些阐述。那么，什么是决策理论？

> **Decision theory is about finding "optimal" actions, under various definitions of optimality.**

即：决策理论的主要内容是，在确定“什么是最优”的前提下，寻找“最优的”actions。那么，什么是 actions?

1.1　什么是 action?

> **An action is the generic term for what is produced by our system.**

一个 action 是我们所构建的系统的产出。例如，对于一个预测未来 3 小时候后台风位置的系统，它的 action 可能是一个经纬度坐标 (当然也可能是一个关于经纬度的概率分布等等)。

对于系统的 actions，我们需要一个能够对其进行打分的评价准则 (Evaluation Criteria)。只有这样，我们才能评价我们所构建的系统的优劣。例如，对于刚刚提到的台风位置预测系统，假设 action 是一个经纬度坐标，我们则可以用 3 小时后台风的真实位置与 action 的距离作为评价准则——距离越大，则说明 action 越不好，那么也可以间接反映我们的预测系统其实不太准确。综上，评价准则对于决策理论十分重要。

系统所有可能的 actions 的集合称为 Action Space。在模型化现实问题时，确定 Action Space 与　确定评价准则是最先进行的两个步骤。之后，我们则需要明确输入 (input)、结果 (outcome) 和对应的输入空间 (input space)、结果空间 (outcome space)。

1.2 输入与结果

输入，即 input，指传入系统的数据。在统计学中，输入又被称为 covariate。结果，即 outcome，或称 label，指输入所对应的“正确答案”，是 what actually happen。特别注意，outcome 虽然也可以称为 output，但它并不是系统的 output——系统的 output 是上文提到的 action！

例如，对于银行的反欺诈系统，input 可能是一位客户的信息，如：年龄、性别、历史贷款情况、历史还款情况等。outcome 是该客户的真实身份判定，假设该客户是事实上的“坏人”，那么 outcome 即为“坏人”，这是确定的、唯一的。action 是系统基于 input 的输出，即预测。预测可能出错，即 action 可能为“好人”。并且，不同的系统对于同一个人的预测也可能不同。因此，我们说 outcome like “right answer” is what actually happen，但 action 则是系统的输出，并不等同于客观事实。

1.3 继续建模

综上，我们可以将问题模型化为 (假定 y 与 a 独立)：

- 观察，得到 input x；
- 系统根据 input 得到一个 action a；
- 观察，得到 outcome y；
- 根据评价准则，结合 y，对 a 进行评价：$l(a, y)$。

上述问题对应三个空间：

- Input space，记做 $\mathcal{X}$，表示由所有可能输入所构成的集合；
- Action space，记做 $\mathcal{A}$，表示由系统所有可能输出所构成的集合；
- Outcome space，记做 $\mathcal{Y}$，表示由所有可能结果所构成的集合。

例如，对于线性回归问题，input space 是 R^d，action space 是 R，outcome space 也是 R。对于 Logistic 回归问题，input space 是 R^d，action space 是 $[0, 1]$，outcome space 是 $\{0, 1\}$。

定义了三大空间后，我们接着定义决策函数 (decision function) 与损失函数 (loss function)：

A decision function gets input $x \in \mathcal{X}$ and produces an action $a \in \mathcal{A}$, that is:

$$f : \mathcal{X} \to \mathcal{A};$$

A loss function evaluates an action in the context of the outcome y, that is:

$$l : \mathcal{A} \times \mathcal{Y} \to \mathcal{R}。$$

决策函数即系统，它根据输入 x，确定 action a 作为预测。损失函数以 action a 与 outcome y 作为它的输出，得到一个数字作为对该 action a 的评价。但损失函数仅仅是对一个 action 的评价——如何评价一个决策函数？这就是统计学习理论所关注的核心问题。

二 统计学习理论

2.1 Risk：泛化误差

统计学习理论有两个基本假设：第一，action a 与 outcome y 独立；第二，存在联合分布 $P_{\mathcal{X}\times\mathcal{Y}}$，所有的数据 (x,y) 都是从该联合分布中 i.i.d 产生的，即样本满足" 独立同分布" 的特点。

如果一个决策函数 $f(x)$ 是“好的”，那么说明该决策函数在联合分布上的平均表现还不错：

$f(x)$ **does well on average, or** $l(f(x), a)$ **is usually small.**

在统计学习理论中，我们用决策函数的“risk”表征它的平均表现，等同于泛化误差：

The risk of a decision function $f : \mathcal{X}\to\mathcal{A}$ **is:**

$$R(f) = E\Big[l(f(x), y)\Big]$$

In words, it's the expected loss of f **on a new example (x,y) drawn randomly from distribution** $P_{\mathcal{X}\times\mathcal{Y}}$**.**

在所有可能的决策函数中，一定存在一个 f 的 risk 最小。该 risk 最小的决策函数被称为贝叶斯决策函数：

A Bayes decision function $f^* : \mathcal{X}\to\mathcal{A}$ **is a function that achieves the minimal risk among all possible functions:**

$$f^* = \arg\min_f R(f)$$

where the minimum is taken over all functions from $\mathcal{X}$ **to** $\mathcal{A}$**.**

贝叶斯决策函数的 risk 被称为 Bayes risk，它代表决策函数能够达到的最好的情况，是一种“标杆”。有时我们也把贝叶斯决策函数称为 target function。下面是两种情景下的贝叶斯函数的形式：

- 最小二乘回归：评价准则是平方误差，贝叶斯函数为 $f^*(x) = E[y|x]$；
- 多元分类问题：评价准则是 0-1 误差，贝叶斯函数为 $f^*(x) = \arg\max_{1\le k\le K} P(y = k|x)$。

然而，我们无法真正解出贝叶斯决策函数，因为我们没有办法计算 risk——由于我们仅仅知道有这样一个产生资料的联合分布，但并不知道该联合分布的具体信息，因此 risk 无法计算。所以我们需要“估计”!

2.2 Empirical Risk：经验误差

我们现有的数据是从未知分布中产生的样本：$D_n = \{(x_1\ y_1), \cdots, (x_n, y_N)\}$。定义决策函数 f 的经验风险 (empirical risk)：

The empirical risk of $f : \mathcal{X} \to \mathcal{A}$ with respect to D_n is:

$$\hat{R}_n(f) = \frac{1}{n} \sum_{i=1}^{n} l(f(x_i), y_i)$$

根据大数定理，当 n 趋于无穷大时，$\hat{R}_n(f)$ 依概率收敛于 $R(f)$。因此，经验风险是 risk 的一个相合估计量。同时，很容易证明，经验风险也是 risk 的无偏估计量。因此，用经验风险去估计 risk 是一个可行的方法。既然经验风险可以计算，那么我们就可以进行经验风险最小化 (empirical risk minimization)：

A function $\hat{f}$ is an empirical risk minimizer if:

$$\hat{f} = \arg\min_{f}\ \hat{R}_n(f)$$

where the minimum is taken over all functions.

经验风险最小化的坏处是：它使得最后得到的 $\hat{f}$ 好像仅仅“记住”了数据——在训练数据上不犯任何错误，但这样往往说明它的泛化能力 (generalize) 很差——对于新的、没有见过的数据的预测能力很差。一种修改方法是，加入一些平滑的因素——smooth flavor，使得决策函数的变化不要特别剧烈。具体而言，我们可以使用 Constrained ERM：在所有可行的决策函数中，拿出一个子集，即拿出一部分决策函数出来进行选择：

Instead of minimizing empirical risk over all decision functions.

这“一部分决策函数”，构成假说空间 (hypothesis space)：

A hypothesis space $\mathcal{F}$ is a set of functions mapping $\mathcal{X} \to \mathcal{A}$. It is the collection of decision functions we are considering.

在 hypothesis space 中，我们有 risk 最小的 hypothesis，记作：

$$\hat{f}_n = \arg\min_{f \in \mathcal{F}}\ \hat{R}_n(f)$$

也有经验风险最小的 hypothesis，记作：

$$f_{\mathcal{F}}^* = \arg\min_{f \in \mathcal{F}}\ R(f)$$