Lecture 3 Excess Risk Decomposition

Alex 2019 年 7 月 23 日

一 误差分解

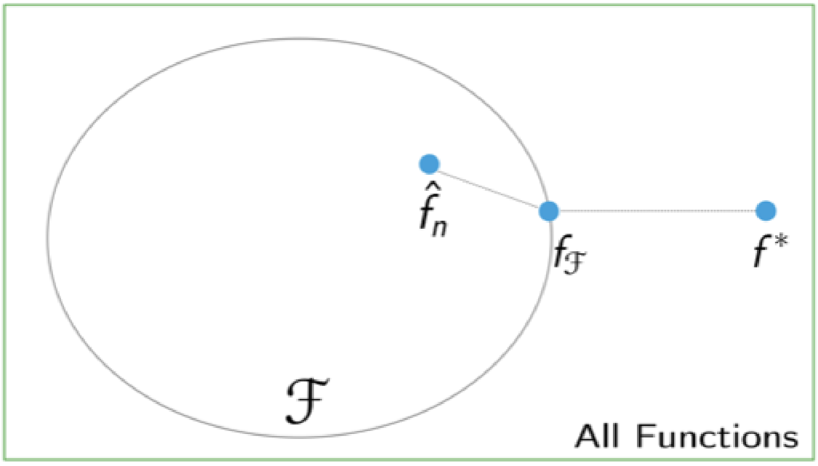

图 1: 误差分解示意图

上图中的矩形表示整个函数空间，其中的椭圆表示我们所考虑的假说空间，即 hypothesis space，也就是我们做经验误差最小化 (ERM) 时考虑的函数集合。我们将整个函数空间中 risk 最小的函数称为贝叶斯决策函数，记做：

$$f^* = \arg\min_{f} E\Big[l(f(X), Y)\Big]$$

将 hypothesis space 中 risk 最小的函数记做 $f_{\mathcal{F}}$：

$$f_{\mathcal{F}} = \arg\min_{f\in\mathcal{F}} E\Big[l(f(X), Y)\Big]$$

将 hypothesis space 中经验误差最小的函数记做 $\hat{f}_n$：

$$\hat{f}_n = \arg\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\Big[l(f(x_i), y_i)\Big]$$

由于完备的函数空间非常庞大，加之我们对于 target function 的了解甚少，因此 f^* 往往不在我们所选择的 hypothesis space 中——我们将其标记在椭圆之外的某处。$f_{\mathcal{F}}$ 是

hypothesis space 中”最好”的那一个 hypothesis，它往往位于 hypothesis space 的边界，趋向于贝叶斯决策函数。这里，我们将 $f_{\mathcal{F}}$ 与 f^* 之间的 risk 之差定义为**近似误差——Approximation Error**：

$$R(f_{\mathcal{F}}) - R(f^*)$$

在 hypothesis space 中根据训练样本做 ERM 得到的 $\hat{f}_n$ 可用椭圆内一点表示。我们将 $\hat{f}_n$ 与 $f_{\mathcal{F}}$ 之间的 risk 之差——注意不是 empirical risk 之差！——定义为**估计误差 (Estimation Error)**：

$$R(\hat{f}_n) - R(f_{\mathcal{F}})$$

最后,我们将任意一个函数 f 的 risk 与贝叶斯决策函数的 risk 之差记为**总误差 (Excess Risk)**：

$$R(f) - R(f^*)$$

误差分解是指将 excess risk 分解成 approximation error 与 estimation error 之和。对 ERM 的结果 $\hat{f}_n$ 的 risk 进行分解，得到：

$$\begin{aligned} Excess\ Risk(\hat{f}_n) &= R(\hat{f}_n) - R(f^*) \\ &= \left(R(\hat{f}_n) - R(f_{\mathcal{F}})\right) + \left(R(f_{\mathcal{F}}) - R(f^*)\right) \\ &= estimation\ error + approximation\ error \end{aligned}$$

二　近似误差：approximation error

当损失函数确定时，贝叶斯决策函数也确定。当损失函数与 hypothesis space 确定时，hypothesis space 中 risk 最小的那个 hypothesis，即 $f_{\mathcal{F}}$ 也确定。我们一般均假定损失函数给定，除非我们的研究对象为损失函数。因此，我们可以将 approximation error 看做是 hypothesis space 的函数——hypothesis space 确定 $\to f_{\mathcal{F}}$ 确定 $\to$approximation error 确定。

近似误差表现的是我们所选择的 hypothesis space 中最优秀的那个 hypothesis 与最佳决策函数之间的差距，可以看做是对我们使用 hypothesis space 进行复杂度限制的一种”惩罚”——如果 hypothesis space 为全函数空间，根据定义，近似误差达到最小值，为 0。因此，越大、越复杂的 hypothesis space，approximation error 越小。

三　估计误差：estimation error

estimation error 是指运用特定算法在某个训练集上的训练结果与 $f_{\mathcal{F}}$ 的差距。当算法、训练样本固定时，”越大”的 hypothesis space 会带来越大的 estimation error——这也就是 overfitting 的现象。在统计学习理论中，有 VC 维、拉德马赫尔复杂度等方式衡量一个 hypothesis space 的复杂度。

注意，在这里我们更倾向于将 estimation error 看做是随机变量——对于训练集随机(算法固定)——不同的训练集 $D^{(1)}, D^{(2)}, \cdots$ 会得到不同的 $\hat{f}_1, \hat{f}_2, \cdots$，它们的 estimation error 会有差别。最后需要注意，样本量越大，$\hat{f}$ 则越能迫近 $f_{\mathcal{F}}$，estimation error。这是因为——以 ERM 算法为例，考虑无限样本量，用无限样本量，或者说所有样本来进行经验误差最小化，这时“经验误差”=risk，因此也就是在 hypothesis space 中找到了 risk 最小的那一个 hypothesis，即 $f_{\mathcal{F}}$。

如何将 estimation error 限制在一个较低水平是统计学习理论研究的主要方向。对 estimation error 进行限制，也就是找到 estimation error 的一个上界，使得对于任何固定数量的训练集，算法训练出的 hypothesis 的 estimation error 大概率在上界之下。

四 优化误差：optimization error

回顾 ERM 的流程：

- 确定损失函数：$l : \mathcal{A} \times \mathcal{Y} \to R$；
- 选定 hypothesis space $\mathcal{F}$；
- 用某种优化方法找到经验误差最小的 hypothesis $\hat{f}_n \in \mathcal{F}$：

$$\hat{f}_n = \underset{f \in \mathcal{F}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} l(f(x_i), y_i)$$

在此过程中，需要对 hypothesis space 进行谨慎的选择。在数据量一定时，hypothesis space 越复杂，approximation error 越小，但 estimation error 会很大；hypothesis space 越简单，approximation error 越大，但 estimation error 会很小。因此，我们需要 trade-off，选择一个合适的 hypothesis space。当然，如果数据量充足，那么我们可以放心地选择更加复杂的 hypothesis space——因为大数据量能够帮助我们减小 estimation error，防止过拟合。

然而，在实务中，囿于优化手段本身的误差，我们其实并没有拿到真正的经验误差最小的 $\hat{f}_n$，而是拿到经验误差”差不多最小”的 $\tilde{f}_n$。例如，用梯度下降法，在没有收敛前就停止，我们拿到的 hypothesis 肯定不是我们在公式 $\arg\min$ 中所列的那个经验误差最小的 $\hat{f}_n$。因此，我们将我们希望所得 $\hat{f}_n$ 与实际所得 $\tilde{f}_n$ 的 risk 的差距，定义为**优化误差 (Optimization Error)**：

$$R(\tilde{f}_n) - R(\hat{f}_n)$$

注意，optimization error 可正可负。因此，我们可以进一步进行误差分解：

$$\begin{aligned}
Excess\ Risk(\tilde{f}_n) &= R(\tilde{f}_n) - R(f^*) \\
&= \left(R(\tilde{f}_n) - R(\hat{f}_n)\right) + \left(R(\hat{f}_n) - R(f_{\mathcal{F}})\right) + \left(R(f_{\mathcal{F}}) - R(f^*)\right) \\
&= optimization\ error + estimation\ error + approximation\ error
\end{aligned}$$