深度学习理论与实践

偏差方差分解 *

为了避免过拟合，我们经常会在模型的拟合能力和复杂度之间进行权衡．拟合能力强的模型一般复杂度会比较高，容易导致过拟合．相反，如果限制模型的复杂度，降低其拟合能力，又可能会导致欠拟合．因此，如何在模型的拟合能力和复杂度之间取得一个较好的平衡，对一个机器学习算法来讲十分重要．偏差-方差分解（Bias-Variance Decomposition）为我们提供了一个很好的分析和指导工具．

以回归问题为例，假设数据的真实分布为 $p_r(x,y)$，并采用平方损失函数，模型 $f(x)$ 的期望错误 (期望风险) 为:

$$\mathcal{R}(f) = \mathbb{E}_{(x,y)\sim p_r(x,y)}\left[(y - f(x))^2\right] \tag{1}$$

这里，我们把期望风险分解一下，加入 $\mathbb{E}_{y\sim p_r(y|x)}[y]$，即数据的真实条件分布，即已知 x 时 y 的期望，那么:

$$\mathcal{R}(f) = \mathbb{E}_{(x,y)\sim p_r(x,y)}\left[(y - \mathbb{E}_{y\sim p_r(y|x)}[y] + \mathbb{E}_{y\sim p_r(y|x)}[y] + f(x))^2\right] \tag{2}$$

将平方项进行展开:

$$\begin{aligned}\mathcal{R}(f) = \mathbb{E}_{(x,y)\sim p_r(x,y)}\Big[&\big(y - \mathbb{E}_{y\sim p_r(y|x)}[y]\big)^2 + \big(\mathbb{E}_{y\sim p_r(y|x)}[y] - f(\boldsymbol{x})\big)^2 \\ &+ 2\big(y - \mathbb{E}_{y\sim p_r(y|x)}[y]\big)\big(\mathbb{E}_{y\sim p_r(y|x)}[y] - f(\boldsymbol{x})\big)\Big]\end{aligned} \tag{3}$$

在等式 3 中，y 是不变的，$\mathbb{E}_{y\sim p_r(y|x)}[y]$ 也是不变的。只有 $f(x)$ 是要学习的模型，是可变的。根据期望风险最小化原则，最小化 $\mathcal{R}(f)$ 即可以获得最优模型 $f^*(x)$，则使的等式 3 中的第二项和第三项变为 0。则:

$$f^*(x) = \mathbb{E}_{y\sim p_r(y|x)}[y] \tag{4}$$

那么这个时候等式 3 就只剩下了第一项:

$$\mathcal{R}(f) = \mathbb{E}_{(x,y)\sim p_r(x,y)}\left[\big(y - \mathbb{E}_{y\sim p_r(y|x)}[y]\big)^2\right] = \mathbb{E}_{(x,y)\sim p_r(x,y)}\left[(y - f^*(x))^2\right] \tag{5}$$

此时，等式 5 就是期望风险最小化以后的值，这个是通过学习 $f(x)$ 不能再减小的值，相当于优化的下限，将这一损失称为损失 ϵ。这一部分损失通常是由于样本分布以及噪声引起的。

$$\epsilon = \mathbb{E}_{(x,y)\sim p_r(x,y)}\left[(y - f^*(x))^2\right] \tag{6}$$

*来自于互联网公开资料，所有权归原作者所有

则，最小化平方损失期望风险得到的最优模型为 $f^*(x)=\mathbb{E}_{y\sim p_r(y|x)}[y]$，即 $f^*(x)$ 的预测值应该是在已知 x 的前提下从数据分布中得到的 y 的值的期望。则有：

$$\begin{aligned}&\mathbb{E}_{(\boldsymbol{x},y)\sim p_r(\boldsymbol{x},y)}\left[(y-f^*(\boldsymbol{x}))\right]=0\\&\mathbb{E}_{(\boldsymbol{x},y)\sim p_r(\boldsymbol{x},y)}\left[(y-f^*(\boldsymbol{x}))^2\right]=\varepsilon\neq 0\end{aligned}\tag{7}$$

现在我们已经知道了 $f^*(x)$，那么对于任意的模型 $f(x)$，它的期望误差是：

$$\begin{aligned}\mathcal{R}(f)&=\mathbb{E}_{(x,y)\sim p_r(x,y)}\left[(y-f(\boldsymbol{x}))^2\right]\\&=\mathbb{E}_{(\boldsymbol{x},y)\sim p_r(\boldsymbol{x},y)}\left[(y-f^*(\boldsymbol{x})+f^*(\boldsymbol{x})-f(\boldsymbol{x}))^2\right]\\&=\mathbb{E}_{(\boldsymbol{x},y)\sim p_r(\boldsymbol{x},y)}\left[(y-f^*(x))^2+(f(\boldsymbol{x})-f^*(\boldsymbol{x}))^2+2\,(y-f^*(x))\,(f^*(x)-f(x))\right]\\&=\mathbb{E}_{\boldsymbol{x}\sim p_r(\boldsymbol{x})}\left[(f(\boldsymbol{x})-f^*(\boldsymbol{x}))^2\right]+\varepsilon\end{aligned}\tag{8}$$

其中，第一项就是当前模型与最优模型之间的差距，是机器学习算法可以优化的真实目标。

在实际训练一个模型 $f(x)$ 时，训练集 D 是从真实分布 $p_r(x,y)$ 上独立同分布地采样出来的有限样本集合。不同的训练集会得到不同的模型。令 $f_D(x)$ 表示在训练集 D 上学习到的模型，一个机器学习算法（包括模型以及优化算法）的能力可以用不同训练集上的模型的平均性能来评价。

现在我们有实际的训练集 D，我们在该训练集上训练得到的模型是 $f_D(x)$。在和 D 相同规模（相同样本数量）的不同训练集上可以训练出不同参数的模型 $f_{D_1}(x)$、$f_{D_2}(x)$...，这些模型有各自的预测结果。我们用这些模型的平均性能（期望）来表示该模型在规模为 D 的训练集上的表现：$\mathbb{E}_D\left[f_D(x)\right]$。

对于单个样本 x，不同训练集 D 得到模型 $f_D(x)$ 和最优模型 $f^*(x)$ 的期望差距为：

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}}\left[(f_{\mathcal{D}}(\boldsymbol{x})-f^*(\boldsymbol{x}))^2\right]\\&=\mathbb{E}_{\mathcal{D}}\left[(f_{\mathcal{D}}(\boldsymbol{x})-\mathbb{E}_{\mathcal{D}}\left[f_{\mathcal{D}}(\boldsymbol{x})\right]+\mathbb{E}_{\mathcal{D}}\left[f_{\mathcal{D}}(\boldsymbol{x})\right]-f^*(\boldsymbol{x}))^2\right]\\&=(\mathbb{E}_{\mathcal{D}}\left[f_{\mathcal{D}}(\boldsymbol{x})\right]-f^*(\boldsymbol{x}))^2+\mathbb{E}_{\mathcal{D}}\left[(f_{\mathcal{D}}(\boldsymbol{x})-\mathbb{E}_{\mathcal{D}}\left[f_{\mathcal{D}}(\boldsymbol{x})\right])^2\right]\end{aligned}\tag{9}$$

现在来证明等式 9 第二行到第三行：

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}}\left[(f_{\mathcal{D}}(\boldsymbol{x})-\mathbb{E}_{\mathcal{D}}\left[f_{\mathcal{D}}(\boldsymbol{x})\right]+\mathbb{E}_{\mathcal{D}}\left[f_{\mathcal{D}}(\boldsymbol{x})\right]-f^*(\boldsymbol{x}))^2\right]\\&=\mathbb{E}_{\mathcal{D}}\left[(f_{\mathcal{D}}(\boldsymbol{x})-\mathbb{E}_{\mathcal{D}}\left[f_{\mathcal{D}}(\boldsymbol{x})\right])^2+(\mathbb{E}_{\mathcal{D}}\left[f_{\mathcal{D}}(\boldsymbol{x})\right]-f^*(x))^2\right.\\&\qquad\left.+2\,(f_{\mathcal{D}}(\boldsymbol{x})-\mathbb{E}_{\mathcal{D}}\left[f_{\mathcal{D}}(\boldsymbol{x})\right])\,(\mathbb{E}_{\mathcal{D}}\left[f_{\mathcal{D}}(\boldsymbol{x})\right]-f^*(x))\right]\end{aligned}\tag{10}$$

第一项是方差：

$$\mathbb{E}_{\mathcal{D}}\left[(f_{\mathcal{D}}(\boldsymbol{x})-\mathbb{E}_{\mathcal{D}}\left[f_{\mathcal{D}}(\boldsymbol{x})\right])^2\right]\tag{11}$$

第二项是偏差的平方：

$$\mathbb{E}_{\mathcal{D}}\left[(\mathbb{E}_{\mathcal{D}}\left[f_{\mathcal{D}}(\boldsymbol{x})\right]-f^*(x))^2\right]\tag{12}$$

其中，$\mathbb{E}_{\mathcal{D}}\left[f_{\mathbb{D}(x)}\right]$ 是模型的平均表现，与某一个训练集合 D 的分布无关，所以：

$$\mathbb{E}_{\mathcal{D}}\left[\left(\mathbb{E}_{\mathcal{D}}\left[f_{\mathcal{D}}(\boldsymbol{x})\right]-f^{*}(x)\right)^{2}\right]=\left(\mathbb{E}_{\mathcal{D}}\left[f_{\mathcal{D}}(\boldsymbol{x})\right]-f^{*}(x)\right)^{2} \tag{13}$$

第三项，

$$\mathbb{E}_{\mathcal{D}}\left[2\left(f_{\mathcal{D}}(\boldsymbol{x})-\mathbb{E}_{\mathcal{D}}\left[f_{\mathcal{D}}(\boldsymbol{x})\right]\right)\left(\mathbb{E}_{\mathcal{D}}\left[f_{\mathcal{D}}(\boldsymbol{x})\right]-f^{*}(x)\right)\right] \tag{14}$$

其中右边的大括号和上面第二项中已经分析了，与 D 的分布无关，所以就等价于：

$$\mathbb{E}_{\mathcal{D}}\left[\left(f_{\mathcal{D}}(\boldsymbol{x})-\mathbb{E}_{\mathcal{D}}\left[f_{\mathcal{D}}(\boldsymbol{x})\right]\right)\right]\times 2\left(\mathbb{E}_{\mathcal{D}}\left[f_{\mathcal{D}}(\boldsymbol{x})\right]-f^{*}(x)\right) \tag{15}$$

右边一项中 $f_{\mathcal{D}}(x)$ 的期望是 $\mathbb{E}_{\mathcal{D}}\left[f_{\mathcal{D}}(x)\right]$， $\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\mathcal{D}}\left[f_{\mathcal{D}}(x)\right]\right]=\mathbb{E}_{\mathcal{D}}\left[f_{\mathcal{D}}(x)\right]$ 期望它是本身，所以相减为 0，该项消除。因此：

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}}\left[\left(f_{\mathcal{D}}(\boldsymbol{x})-f^{*}(\boldsymbol{x})\right)^{2}\right]\\&=\underbrace{\left(\mathbb{E}_{\mathcal{D}}\left[f_{\mathcal{D}}(\boldsymbol{x})\right]-f^{*}(\boldsymbol{x})\right)^{2}}_{\text{(偏差. x)}^2}+\underbrace{\mathbb{E}_{\mathcal{D}}\left[\left(f_{\mathcal{D}}(\boldsymbol{x})-\mathbb{E}_{\mathcal{D}}\left[f_{\mathcal{D}}(\boldsymbol{x})\right]\right)^{2}\right]}_{\text{方差. x}}\end{aligned} \tag{16}$$

其中第一项为偏差（Bias），是指一个模型在不同训练集上的平均性能和最优模型的差异，可以用来衡量一个模型的拟合能力. 第二项是方差（Variance），是指一个模型在不同训练集上的差异，可以用来衡量一个模型是否容易过拟合.

偏差越小，模型和最优模型之间的差距就越小，即模型的拟合能力越强。

方差越大，代表模型对不同数据集的分布学习的越好（过拟合），导致不同数据集训练出来的模型与模型的平均期望的差距越大。

用 $\mathbb{E}_{\mathcal{D}}\left[\left(f_{\mathcal{D}}(\boldsymbol{x})-f^{*}(\boldsymbol{x})\right)^{2}\right]$ 来代替公式 8 中的 $\left(f(x)-f^{*}(x)\right)^{2}$，期望错误可以进一步写为：

$$\begin{aligned}\mathcal{R}(f)&=\mathbb{E}_{x\sim p_r(x)}\left[\mathbb{E}_{\mathcal{D}}\left[\left(f_{\mathcal{D}}(\boldsymbol{x})-f^{*}(\boldsymbol{x})\right)^{2}\right]\right]+\epsilon\\&=(\text{ bias })^{2}+\text{ variance }+\epsilon\\(\text{ bias })^{2}&=\mathbb{E}_{x}\left[\left(\mathbb{E}_{\mathcal{D}}\left[f_{\mathcal{D}}(\boldsymbol{x})\right]-f^{*}(\boldsymbol{x})\right)^{2}\right]\\\text{variance }&=\mathbb{E}_{x}\left[\mathbb{E}_{\mathcal{D}}\left[\left(f_{\mathcal{D}}(\boldsymbol{x})-\mathbb{E}_{\mathcal{D}}\left[f_{\mathcal{D}}(\boldsymbol{x})\right]\right)^{2}\right]\right]\end{aligned} \tag{17}$$