如何计算模型的偏差与方差 (Bias and Variance)

1．偏差与方差的定义

偏差 (Bias) 与方差 (bariance) 的概念来自于样本外误差的分解 (Decomposition of out-of-sample error)。样本外误差的定义如下所示，其中 $g^{(\mathcal{D})}$ 为基于数据集 $\mathcal{D}$ 得到的模型，f 为真实的目标函数（未知），$\mathbb{E}_{\mathbf{x}}$ 表示基于 $\mathbf{x}$ 的期望值（$\mathbf{x}$ 来自总体样本空间 $\mathcal{X}$）：

$$E_{out}(g^{(\mathcal{D})}) = \mathbb{E}_{\mathbf{x}}[(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x}))]$$

计算上式在数据集 $\mathcal{D}$ 上的期望可以得到一个更加广泛的结果：

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}[E_{out}(g^{(\mathcal{D})})] &= \mathbb{E}_{\mathcal{D}}[\mathbb{E}_{\mathbf{x}}[(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x}))]] \\ &= \mathbb{E}_{\mathbf{x}}[\mathbb{E}_{\mathcal{D}}[(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x}))]] \\ &= \mathbb{E}_{\mathbf{x}}[\mathbb{E}_{\mathcal{D}}[g^{\mathcal{D}}(\mathbf{x})^2] - 2\mathbb{E}_{\mathcal{D}}[g^{\mathcal{D}}(\mathbf{x})]f(\mathbf{x}) + f(\mathbf{x})^2]\end{aligned}$$

$\mathbb{E}_{\mathcal{D}}[g^{\mathcal{D}}(\mathbf{x})]$ 可视为“平均函数”，以 $\bar{g}(\mathbf{x})$ 表示。可以这样理解：假设有一系列数据集 $\mathcal{D}_1, ..., \mathcal{D}_{\mathcal{K}}$，通过给定的算法得到对应数据集的最优模型 $g_1, ..., g_K$，这样对于任意给定的 $\mathbf{x}$ 都可以计算出对应的平均函数的值：$\bar{g}(\mathbf{x}) \approx \frac{1}{K}\sum_{k=1}^{K} g_k(\mathbf{x})$.　　定义了 $\bar{g}(\mathbf{x})$ 的概念之后，可以重写上述公式：

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}[E_{out}(g^{(\mathcal{D})})] &= \mathbb{E}_{\mathbf{x}}[\mathbb{E}_{\mathcal{D}}[g^{\mathcal{D}}(\mathbf{x})^2] - 2\bar{g}(\mathbf{x})f(\mathbf{x}) + f(\mathbf{x})^2] \\ &= \mathbb{E}_{\mathbf{x}}[\mathbb{E}_{\mathcal{D}}[g^{\mathcal{D}}(\mathbf{x})^2] - \bar{g}(\mathbf{x})^2 + \bar{g}(\mathbf{x})^2 - 2\bar{g}(\mathbf{x})f(\mathbf{x}) + f(\mathbf{x})^2] \\ &= \mathbb{E}_{\mathbf{x}}[\mathbb{E}_{\mathcal{D}}[g^{\mathcal{D}}(\mathbf{x})^2] - 2\mathbb{E}_{\mathcal{D}}[g^{\mathcal{D}}(\mathbf{x})] \cdot \bar{g}(\mathbf{x}) + \bar{g}(\mathbf{x})^2 + (\bar{g}(\mathbf{x}) - f(\mathbf{x}))^2] \\ &= \mathbb{E}_{\mathbf{x}}[\mathbb{E}_{\mathcal{D}}[(g^{\mathcal{D}}(\mathbf{x}) - \bar{g}(\mathbf{x}))^2] + (\bar{g}(\mathbf{x}) - f(\mathbf{x}))^2]\end{aligned}$$

由于 $\bar{g}(\mathbf{x})$ 并不会因为数据集 $\mathcal{D}$ 的改变而改变，故 $\mathbb{E}_{\mathcal{D}}$ 中的 $\bar{g}(\mathbf{x})$ 可以视为常数（实际上是一个关于 $\mathbf{x}$ 的函数），因此上式中最后一步分变形是成立的。至此，我们便得到了偏差与方差的数学定义：

$$\begin{aligned}bias(\mathbf{x}) &= (\bar{g}(\mathbf{x}) - f(\mathbf{x}))^2 \\ var(\mathbf{x}) &= \mathbb{E}_{\mathcal{D}}[(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x}))^2]\end{aligned}$$

样本外误差可以使用偏差与方差来表示：

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}[E_{out}(g^{(\mathcal{D})})] &= \mathbb{E}_{\mathbf{x}}[var(\mathbf{x}) + bias(\mathbf{x})] \\ &= bias + var\end{aligned}$$

方差（var）可以视作衡量模型不稳定性的指标，因为它衡量的是不同数据集的情况下所得到的最优模型的之间的差异程度，若是方差较大，则表示换一批数据集时模型预测结果会与训练集有较大偏差，这一般意味着过拟合；偏差（$bias$）则衡量的是模型对真实目标函数的整体描述能力，因为它计算的是真实目标函数与平均函数的偏差。

2．偏差与方差的计算实例

假设真实的目标函数为 $f(x) = sin(\pi x)$，数据集的样本量为 $N = 2$。从 $[-1, 1]$ 随机均匀抽取 $x1, x2$ 用于生成数据集 $(x1, y1), (x2, y2)$，使用如下两个假设集拟合样本：

$$\begin{aligned}\mathcal{H}_0 &: h(x) = b \\ \mathcal{H}_1 &: h(x) = ax + b\end{aligned}$$

根据计算（详细的计算代码可参见 bias_and_variance_cal.py），$\mathcal{H}_0$ 的偏差与方差分别为 0.4964 与 0.2469，$\mathcal{H}_1$ 的偏差与方差分别为 0.2044 与 1.6647。尽管 $\mathcal{H}_1$ 的偏差不到原来的一半，但是方差增加了 7 倍，故整体的样本外误差明显高于 $\mathcal{H}_0$，因此在样本量为 2 的情况下 $\mathcal{H}_0$ 优于 $\mathcal{H}_1$。

3．总结

- 方差的计算公式：$var(\mathbf{x}) = \mathbb{E}_{\mathcal{D}}[(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x}))^2]$，关于数据集 $\mathcal{D}$ 的期望如何理解？
  - 可以将 $\mathbf{x}$ 看作常量，$\bar{g}(\mathbf{x})$ 为 $g_1, g_2, ..., g_K$ 在 $\mathbf{x}$ 处的平均值，$g^{(\mathcal{D})}(\mathbf{x})$ 则分别为 $g_1, g_2, ..., g_K$ 在 $\mathbf{x}$ 处的值，因此此公式可以看作不同 g 在 $\mathbf{x}$ 处值的方差，故 $\mathbb{E}_{\mathbf{x}}(var(\mathbf{x}))$ 为 $\mathbf{x}$ 上方差的期望。

- 通过模拟计算可知 $N = 2$ 时，$\mathcal{H}_0$ 优于 $\mathcal{H}_1$，但当数据集规模增加时，各个假设的方差都将减小，分别设想 $N = 10, 100, 1000, 10000$ 时，随着 N 的增加，$\mathcal{H}_0$, $\mathcal{H}_1$ 在不同数据集上得到的 g 都将固定在最优直线的附近（例如对假设 $\mathcal{H}_0$，当 $N = 10000$ 时，g 将在 $y = 0$ 附近波动，且波动范围很小），因此整体的方差也会随 N 的增加而减小，最终起决定作用的将是偏差，故 $\mathcal{H}_1$ 随着 N 的增加将优于 $\mathcal{H}_0$。
- 对于相同的假设集，使用不同的寻优算法可能得到不同的 $g^{(\mathcal{D})}$，如 perceptron 与 gradient descent，因此最终的偏差与方差可能也不同。