

1 偏差-方差分解

前面我们多次提到过如何评估一个模型的优劣, 其关键在于评估其泛化性能。第七讲中也提到, 模型在训练集与测试集上的 MSE 与“偏差”、“方差”密切相关, 本节我们将详细讨论偏差与方差的细节。

1.1 预测误差期望

考虑模型 $y = f(x) + \varepsilon$, 其中 $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ 。假设我们在训练集上训练得拟合函数 \hat{f} , 那么对样本 x 的预测值 $\hat{f}(x)$ 与真实值 y 之误差期望可写作 $E[(y - \hat{f}(x))^2]$ 。由期望的性质, 有

$$\begin{aligned} E[(y - \hat{f}(x))^2] &= E[(f - \hat{f} + \varepsilon)^2] \\ &= E[(f - \hat{f})^2] + 2E[\varepsilon(f - \hat{f})] + E[\varepsilon^2] \\ &= E[(f - \bar{f} + \bar{f} - \hat{f})^2] + \sigma^2 \\ &= E[(f - \bar{f})^2] + 2E[(f - \bar{f})(\bar{f} - \hat{f})] + E[(\bar{f} - \hat{f})^2] + \sigma^2 \\ &= (f - \bar{f})^2 + D(\hat{f}) + \sigma^2 \end{aligned} \quad (1)$$

其中 $\bar{f} = E\hat{f}$ 。若称 $f - \bar{f}$ 为偏差 (Bias), 称 $D(\hat{f})$ 为方差 (Variance), 则

$$E[(y - \hat{f}(x))^2] = \text{Bias}^2 + \text{Variance} + \sigma^2 \quad (2)$$

式 (2) 即为预测误差期望的分解式。可以看出, 我们训练得到的模型所给出的预测值与真实值的误差来源于 3 处:

- Bias: 偏差项, 即为预测值的期望与真实值的差。若该项不为 0, 则我们训练所得模型是有偏估计; 反之, 则是无偏估计。
- Variance: 方差项, 即为预测值的方差。该项指示了我们训练所得模型预测值的离散程度。此处的“离散程度”是指, 我们在不同的数据集上会训练出含不同参数的模型, 他们对同一 x 预测值的离散程度。
- σ^2 : 扰动项, 即为随机扰动的方差。该项是源于测量真实值时的噪音或是观察样本时的扰动, 一般我们认为是一服从正态分布的随机扰动, 则该项为常数。

1.2 参数误差期望

也可以通过另一个角度来看待偏差、方差。考虑真实值 y 服从参数为 θ 的模型 $f(x; \theta)$, 我们在训练集上训练所得模型参数为 $\hat{\theta}$, 那么 $\hat{\theta}$ 与 θ 的误差期望可以同样地推导:

$$\begin{aligned} E[(\theta - \hat{\theta})^2] &= E[(\theta - \bar{\theta})^2] + E[(\bar{\theta} - \hat{\theta})^2] + 2E[(\theta - \bar{\theta})(\bar{\theta} - \hat{\theta})] \\ &= (\theta - \bar{\theta})^2 + D(\hat{\theta}) \\ &= \text{Bias}^2 + \text{Variance} \end{aligned} \quad (3)$$

其中 $\bar{\theta} = E\hat{\theta}$ 。式 (3) 说明, 训练得到的参数与真实参数间的误差来源于 2 处:

- Bias: 偏差项, 训练所得参数与真实参数之差的期望。显然, 该项越小, 参数估计就越准确。
- Variance: 方差项, 训练所得参数的离散程度。同样此处是指在不同训练集上所得参数的离散程度。该项越小, 参数估计的波动就越小。

此时我们再回顾第七讲中展示过的图1。

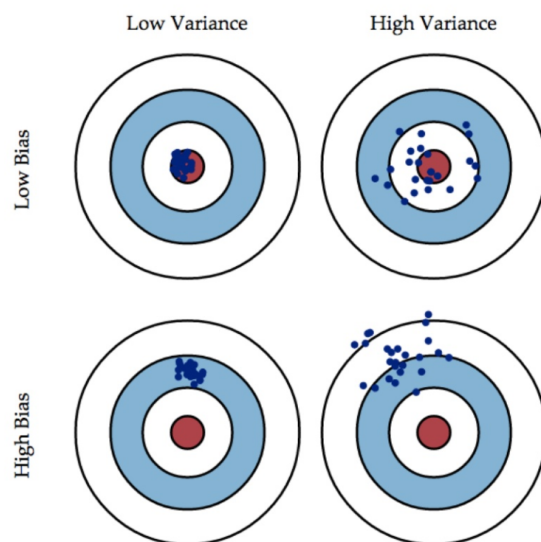


Figure 1: 偏差-方差示意图

中心红点可看作是模型真实的参数，蓝点则是在某个数据集上训练出的参数。偏差体现在蓝点平均位置与红点的远近上，方差体现在蓝点自身的散布疏密上。很显然，偏差与方差同时较低是我们追求的理想目标，可惜的是，通常在实际应用中需要牺牲一方以期望另一方的降低，这就是接下来要讲的偏差-方差权衡。

2 偏差-方差权衡与模型选择

我们在上一节从理论推导的角度解释了偏差与方差的关系，这一节来谈谈偏差与方差对选择模型有什么指导意义。首先，要意识到训练集和测试集的随机性。如图2所示，图中的深蓝色和深红色线条分别表示训练误差的期望和测试误差的期望。每次的训练集和测试集可能都不一样，但是模型复杂度与误差的期望总是呈现这样的规律。

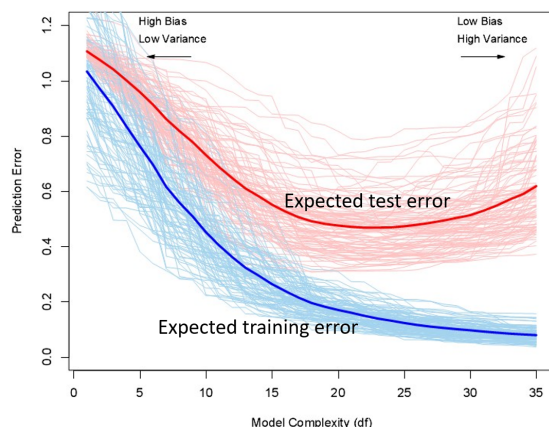


Figure 2: 模型复杂度与预测误差的关系图

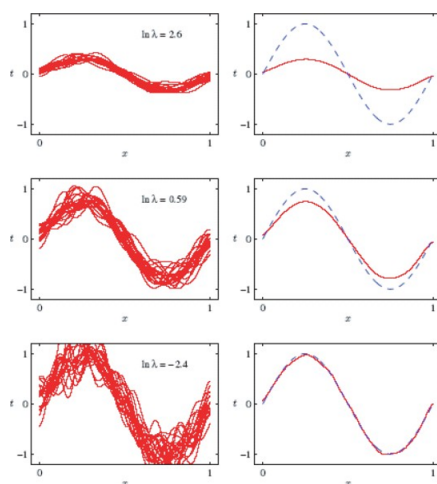


Figure 3: 不同 λ 对应的回归效果图

例如在回归问题中，模型的复杂度具体表现为模型特征的个数。若正则化参数为 λ ，那么 λ 的值越小，意味着特征越多，即模型越复杂。从图3可以看出，简单的（高度正则化的）模型拥有更小的方差，但是偏差也更大，复杂的模型有更小的偏差和更大的方差，这是一个需要权衡的问题。训练集的随机性导致了模型的方差。简单模型的学习性能对不同的训练集不是很敏感，而复杂模型对训练集的改变很敏感，即方差大。如图4所示，从左往右模型越来越复杂， $\hat{\theta}_1$ 表示用原训练集训练出来的回归曲线的参数， $\hat{\theta}_2$ 表示去掉一个训练样本后回归曲线的参数，蓝叉盖着的地方即是去掉的训练样本。可以发现简单的模型变化不大，而复杂模型变了许多。

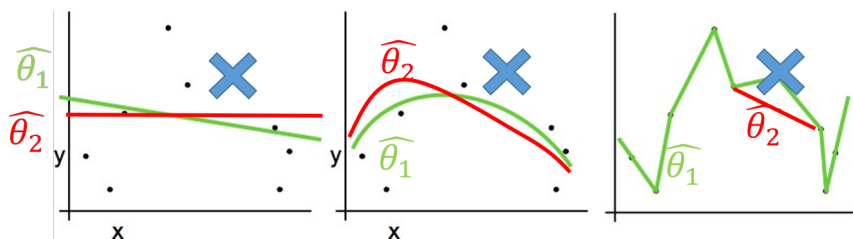


Figure 4: 去掉一个训练样本对不同模型的影响

又比如图5中的 KNN 模型，左图为 $K = 15$ 的情况，右图为 $K = 1$ 的情况。在 KNN 模型中，越大的 K 值对应越模型复杂度越低。若在训练集中增加一个样本，如图中的大蓝圆圈，我们可以发现，简单模型的决策边界没有改变，而复杂模型的决策边界发生了改变。这两个例子都体现出了模型的复杂度越高，模型的方差就越大。

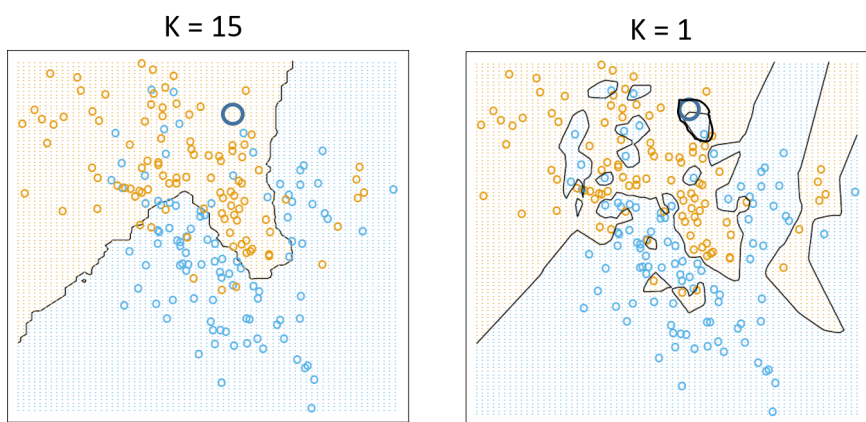


Figure 5: 增加一个训练样本对不同模型的影响

另一方面，当模型复杂度提高时，训练误差总能降低，但是测试误差却不会一直降低，因为随着复杂度提高方差也会变大。当测试误差达到最小值时，可以说近似达到了很好的泛化效果。因此需要解决如下的偏差-方差权衡问题：

- 模型参数太少的话，会因为偏差太大而不准确（不够灵活）
- 模型参数太多的话，会因为方差太大而不准确（对样本随机性太敏感）

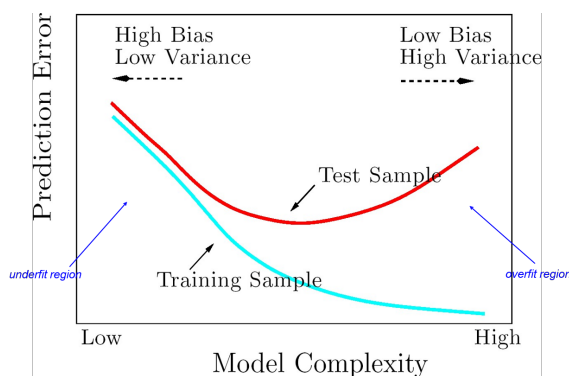


Figure 6: 偏差-方差权衡示意图

如图6所示，当偏差较大时，方差较小，模型处于欠拟合状态，可以通过复杂化模型，增加特征，增长训练时间等方法来应对；当方差较大时，偏差较小，模型处于过拟合状态，可以通过正则化，获得更多数据，以及特征选择来优化模型。根据上一节讲的误差与偏差、方差的关系，可以用 $\text{bias}^2 + \text{variance}$ 来估计测试误差，但是偏差和误差往往无法计算，因为样本 x 和标签 y 的真正分布无从得知。

3 预测误差期望 (EPE)

定义：预测误差期望

对于样本 X 和标签 Y ，函数 f 的预测期望误差为

$$\begin{aligned} EPE(f) &= E(L(Y, f(X))) \\ &= \int L(y, f(x)) Pr(dx, dy) \end{aligned} \quad (4)$$

例如 L_2 损失函数， $EPE = \int (y - f(x))^2 Pr(dx, dy)$ ，理论分析上 L_2 预测误差期望最小时，函数为条件均值，即

$$\hat{f}(x) = E(Y|X = x) \quad (5)$$

KNN 算法就是对该函数的直接近似，它假设 $f(x)$ 可以通过最邻近的几个值来估计。下面说明为什么 L_2 损失函数的 EPE 的最佳估计函数为条件均值。由于 $Pr(X, Y) = Pr(Y|X)Pr(X)$ ，那么 EPE 可以写成

$$EPE(f) = E_X E_{Y|X}[(Y - f(x))^2|X] \quad (6)$$

求的最佳估计函数使 EPE 最小，即

$$f(x) = \arg \min_c E_{Y|X}[(Y - c)^2|X = x] \quad (7)$$

这个最优化问题的解即是条件均值。我们再看另一种证明方法。假设 t 是真实的标签， $y(x)$ 是我们的估计值，那么误差期望为

$$\begin{aligned} E(L) &= \iint L(t, y(x)) p(x, t) dx dt \\ &= \iint (t - y(x))^2 p(x, t) dx dt \end{aligned} \quad (8)$$

我们需要求使 $E(L)$ 最小的 $y(x)$ ，那么对 $E(L)$ 求偏导

$$\frac{\partial E(L)}{\partial y(x)} = 2 \int (t - y(x)) p(x, t) dt = 0 \quad (9)$$

可以得到

$$\int y(x) p(x, t) dt = \int t p(x, t) dt \quad (10)$$

$$\begin{aligned} y^*(x) &= \int \frac{t p(x, t)}{p(x)} dt = \int t p(t|x) dt \\ &= E_{t|x}[t] = E[t|x] \end{aligned} \quad (11)$$

图7给出了不同损失函数的最佳估计函数。其中，当损失函数为 0-1 损失函数时， $\hat{f}(x)$ 就是贝叶斯分类器的损失函数。

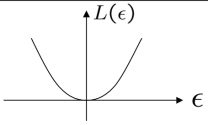
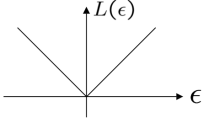
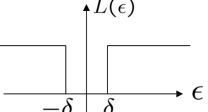
Loss Function	Estimator $\hat{f}(x)$
L_2 	$\hat{f}(x) = E[Y X = x]$
L_1 	$\hat{f}(x) = \text{median}(Y X = x)$
0-1 	$\hat{f}(x) = \arg \max_Y P(Y X = x)$ (Bayes classifier / MAP)

Figure 7: 不同损失函数的估计函数

关于交叉验证的几个注意点

- 如何决定 K 折交叉验证的 K ($\alpha = 1/K$)
通常 $K = 10$, $\alpha = 0.1$ 。当数据集相对较少时，需要增大 K 和减小 α 。 K 需要足够大来使方差足够小，但是这样会比较耗费时间。

- 偏差-方差权衡

较小的 α 会导致较小的偏差。原则上，LOOCV 提供了对分类器泛化能力的几乎无偏差的估计，尤其是当可用训练样本的数量受到严重限制时，但它也可能会有很大的方差。

较大的 α 可以减小方差，但是会导致无法充足利用数据和比较高的偏差。

- 重要的一点是测试集不能用于交叉验证，因为这样做会导致测试阶段过于乐观的准确率（实际是不真实的）。