Project 2 偏差方差分析实验报告

一、 实验目的

实验目的有以下三个：

1) 根据目标函数生成多个有两个实例的数据集，拟合两种模型。
2) 研究不同模型偏差和方差的关系，选择更好的模型。

二、 实验原理

1. 模型均方误差推导

在有监督学习中，对于任何学习算法而言，其预测误差可分解为三部分：样本噪音、模型预测值的方差、预测值与真实值偏差的平方。

$$\mathrm{E}\left[\left(\mathrm{y}-\hat{\mathrm{f}}(x)\right)^2\right]=\epsilon^2+var\left[\hat{f}(x)\right]+\left(bias\left[\hat{f}(x)\right]\right)^2$$

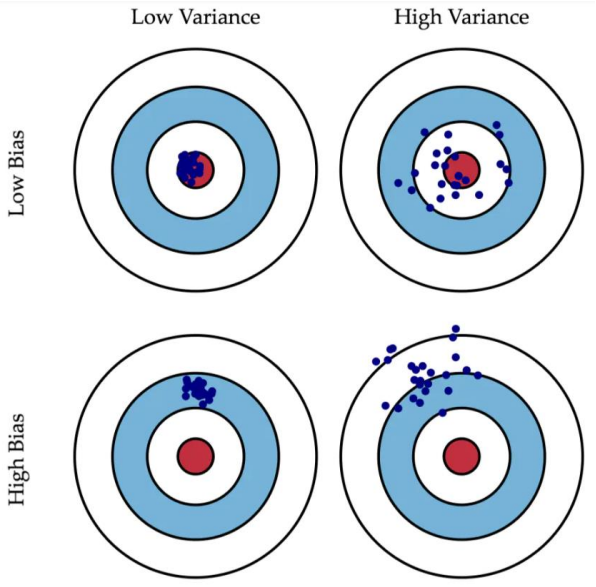
推导如下：

$$\begin{aligned}\mathrm{E}\left[\left(\mathrm{y}-\hat{\mathrm{f}}\right)^2\right]&=\mathrm{E}\left[\mathrm{y}^2+\hat{\mathrm{f}}^2-2\mathrm{y}\hat{\mathrm{f}}\right]=\mathrm{E}[\mathrm{y}^2]+\mathrm{E}\left[\hat{\mathrm{f}}^2\right]-2\mathrm{E}\left[(\mathrm{f}+\epsilon)\hat{\mathrm{f}}\right]\\&=\mathrm{Var}[\mathrm{y}]+(\mathrm{E}[\mathrm{y}])^2+\mathrm{Var}\left[\hat{\mathrm{f}}\right]+\left(\mathrm{E}\left(\hat{\mathrm{f}}\right)\right)^2-2\mathrm{fE}\left[\hat{\mathrm{f}}\right]-2\mathrm{E}[\epsilon]\mathrm{E}\left[\hat{\mathrm{f}}\right]\\&=\mathrm{Var}[\mathrm{y}]+\mathrm{Var}\left[\hat{\mathrm{f}}\right]+\left(f-E\left[\hat{f}\right]\right)^2=\sigma^2+Var\left[\hat{f}\right]+\left(Bias\left[\hat{f}\right]\right)^2\end{aligned}$$

其中，噪声属于不可约减误差，无论使用哪种算法，都无法减少噪声。 通常噪声是从问题的选定框架中引入的错误，也可能是由诸如未知变量之类的因素引起的，这些因素会影响输入变量到输出变量的映射。噪声表达了在当前任务上任何学习算法所能达到的期望泛化误差的下界，即刻画了学习问题本身的难度。而剩下两种误差则与我们选择的学习算法相关，并且可以通过一些方法减小。

2. 偏差和方差

当模型表现不佳时，通常是出现高偏差问题或高方差问题。偏差描述模型输出结果的期望与样本真实结果的差距，方差描述模型对于给定值的输出稳定性。下图展示了方差和偏差关系。



若具体问题没有对方差和偏差有特别的偏好时，往往认为相对较好的模型的顺序：方差小，偏差小>方差小，偏差大>方差大，偏差小>方差大，偏差大。

方差小，偏差大之所以在实际中排位比较靠前，是因为它比较稳定。很多时候，实际中无法获得非常全面的数据集。因此，如果一个模型在可获得的样本上有较小的方差，说明它对不同数据集的敏感度不高，可以期望它对新数据集的预测效果比较稳定。但更重要的还是结合具体问题具体分析。

三、实验过程

1. 试验器材

JDK1.8.0。

2. 实验过程

S1. 生成两个实例，要求特征 X 服从 U[-1,1]，Y 根据f(x) = sin(πx)计算得到。

S2. 估计模型 h(x) = b，其中 b=(y1+y2)/2，并计算偏差和方差。估计模型 h(x) = ax+b，其中a = $\frac{y_1-y_2}{x_1-x_2}$, b = $(y_1 - ax_1)$，并计算偏差和方差。

3. 实验代码

```
public class BVDemo{

	private static double[] x = new double[2];
	private static double[] y = new double[2];

	private static void generateData() {
		x[0] = Math.random()*2-1; x[1] = Math.random()*2-1;
		y[0] = Math.sin(x[0]*Math.PI); y[1] = Math.sin(x[1]*Math.PI);
	}

	private static void firstModel(){
//		H0: h(x)=b0
		double b0 = (y[0]+y[1])/2;
		double biasSq0 = Math.pow((y[0]-b0),2)+Math.pow((y[1]-b0),2);
		double var0 = 0;
		System.out.printf("H0, bias: %.3f, variance: %.3f; ",biasSq0,
var0);
	}

	private static void secondModel(){
//		H1: h(x)=a1*x + b1
		double a1 = (y[0]-y[1])/(x[0]-x[1]);
		double b1 = y[0] - a1*x[0];
		double biasSq1 = Math.pow((y[0]-a1*x[0]-b1),2) +
Math.pow((y[1]-a1*x[1]-b1),2);
		double var1 =  Math.pow((a1*x[0]+b1)-(y[0]+y[1]/2),2) +
Math.pow((a1*x[1]+b1)-(y[0]+y[1]/2),2);
		System.out.printf("H1, bias: %.3f, variance: %.3f\n",biasSq1,
var1);
	}
```

```
	public static void main(String[] args) {
		for (int i=1; i<=10; i++) {
			System.out.println("DataSet "+i);
			generateData();
			firstModel();
			secondModel();
		}
	}
}
```

四、实验结果

```
DataSet 1
H0, bias: 1.130, variance: 0.000; H1, bias: 0.000, variance: 1.299
DataSet 2
H0, bias: 0.151, variance: 0.000; H1, bias: 0.000, variance: 0.398
DataSet 3
H0, bias: 0.004, variance: 0.000; H1, bias: 0.000, variance: 0.474
DataSet 4
H0, bias: 0.004, variance: 0.000; H1, bias: 0.000, variance: 0.246
DataSet 5
H0, bias: 1.796, variance: 0.000; H1, bias: 0.000, variance: 2.247
DataSet 6
H0, bias: 1.253, variance: 0.000; H1, bias: 0.000, variance: 1.474
DataSet 7
H0, bias: 0.221, variance: 0.000; H1, bias: 0.000, variance: 0.351
DataSet 8
H0, bias: 0.430, variance: 0.000; H1, bias: 0.000, variance: 0.668
DataSet 9
H0, bias: 0.105, variance: 0.000; H1, bias: 0.000, variance: 0.106
DataSet 10
H0, bias: 0.731, variance: 0.000; H1, bias: 0.000, variance: 0.853
```

五、结果分析

上述实验结果表明，对于模型假设 0：h(x) = b，其偏差平方和不断变动，但方差始终为 0；而对于模型假设 1：h(x) = ax+b，其偏差始终为 0，方差不断变动。

这是因为模型假设 0 是常函数，将任何实例不加区分地全都预测为$(y_1+y_2)/2$，这与真实目标函数 sin(πx)常常会有出入，所以偏差较大。而且此时，实例的预测值和预测值期望始终等于$(y_1+y_2)/2$，方差则为 0。可以认为模型 0“预测的不准，但预测的很一致”。

而模型假设 1 在假设 0 的基础上加入了一个“特征”x，此时能有效地减低模型的偏差。由于只有两个实例，此时假设 1 为穿过(x_1, y_1) 和(x_2, y_2)的直线，对这两个实例，模型的偏差将始终为 0。若用该模型对另外的新实例预测，假设 1 的模型偏差往往也会小于假设 0 的模型偏差。而由于模型每次会预测出不同的结果，所以其有更大的方差。可以认为模型 1“预测的更准，但也预测的更不一致”。

偏差过大的模型往往是欠拟合的，可以通过增加更多特征、改变特征形式（如对连续变量可以取对数或高次幂）、减小正则化项的系数值等方式缓解；偏差小而方差大的模型

往往是过拟合的，此时可以通过增加样本量、简化模型（如进行特征选择）、对损失函数加入正则化项解决。

总的来说，选择偏差更小还是方差更小的模型要结合实际应用来看。如果该问题更关注预测的准确性，则选择偏差更小的模型，如过更关注模型的稳健性，则在偏差可以接受的情况下，选择方差更小的模型。