偏差和方差

01 为什么会有偏差和方差?

02 偏差

03 方差

04 图形解释偏差和方差

05 机器学习中的举例

01 为什么会有偏差和方差?

1 为什么会有偏差和方差？

对机器学习算法，除了通过实验估计其泛化性能之外，人们往往还希望了解它**为什么具有这样的性能**。我们可以从**偏差**和**方差**的角度来**解释机器学习算法泛化性能**。

泛化能力 语音 编辑 讨论 上传视频

泛化能力（generalization ability）是指机器学习算法对新鲜样本的适应能力。[1] 学习的目的是学到隐含在数据背后的规律，对具有同一规律的学习集以外的数据，经过训练的网络也能给出合适的输出，该能力称为泛化能力。

中文名	泛化能力	特　指	学习算法对新鲜样本的适应能力
外文名	generalization ability	隶　属	计算机科学

1 为什么会有偏差和方差？

真实模型：如果我们能够获得所有可能的数据集合，并在这个数据集合上将损失最小化，那么学习得到的模型就可以称之为“真实模型”。当然，“真实模型”理论上存在，但是工程上无法获得。

机器学习的目的就是：是让机器学习一个模型，使其**更加接近**这个真实模型。

偏差和方差，就是分别从两个方面来描述我们学习到的模型与真实模型之间的**差距**。

02 偏差

2 偏差

偏差（Bias），反映的是期望输出与真实标记的差别。

$$bias^2(\boldsymbol{x}) = \left(\bar{f}(\boldsymbol{x}) - y\right)^2$$

符号	涵义
$\mathbf{x}$	测试样本
D	数据集
y_D	$\mathbf{x}$ 在数据集中的标记
y	$\mathbf{x}$ 的真实标记
f	训练集 D 学得的模型
$f(\mathbf{x}; D)$	由训练集 D 学得的模型 f 对 $\mathbf{x}$ 的预测输出
$\bar{f}(\mathbf{x})$	模型 f 对 $\mathbf{x}$ 的 **期望预测** 输出

偏差度量了模型的期望预测与真实结果的偏离程度，换句话说，就是刻画了模型本身的拟合能力。

03 方差

3 方差

初中数学课本中的方差，是指统计学中的方差。

统计学中：方差（样本方差）是**每个样本值**与**全体样本值的平均数**之差的平方值的**平均数**。

$$s^2 = \frac{1}{n}[(x_1 - x)^2 + (x_2 - x)^2 + \cdots (x_n - x)^2]$$

两人的5次测验成绩如下：

A: 50, 100, 100, 60, 50 -->Average(A) = 72

B: 73, 70, 75, 72, 70 -->Average(B) = 72

A组的方差大？还是B组？

方差(variance)是一个常见的分布描述量。

方差就是分布的离散程度。方差越大，说明随机变量取值越离散。

3 方差

概率论中：方差用来度量随机变量和其数学期望（即均值）之间的偏离程度。

应用在**机器学习**中，可以描述**不同的训练集**训练出的模型输出值之间的**差异**。

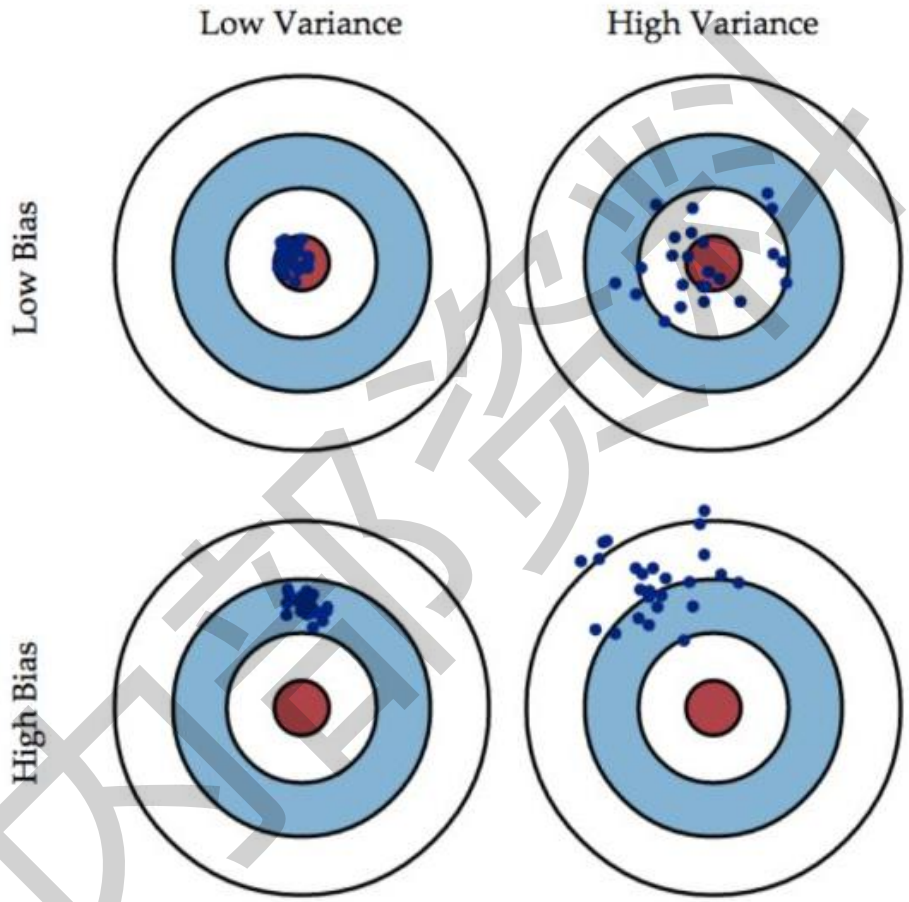
$$var(\boldsymbol{x}) = \mathbb{E}_D\left[\left(f(\boldsymbol{x};D) - \bar{f}(\boldsymbol{x})\right)^2\right]$$

符号	涵义
$\mathbf{x}$	测试样本
D	数据集
y_D	$\mathbf{x}$ 在数据集中的标记
y	$\mathbf{x}$ 的真实标记
f	训练集 D 学得的模型
$f(\mathbf{x};D)$	由训练集 D 学得的模型 f 对 $\mathbf{x}$ 的预测输出
$\bar{f}(\mathbf{x})$	模型 f 对 $\mathbf{x}$ 的 **期望预测** 输出

方差度量了同样大小的训练集的变动所导致的学习性能的变化，换句话说，就是刻画了数据扰动所造成的影响。

04 图形解释偏差和方差

4 图形解释偏差和方差

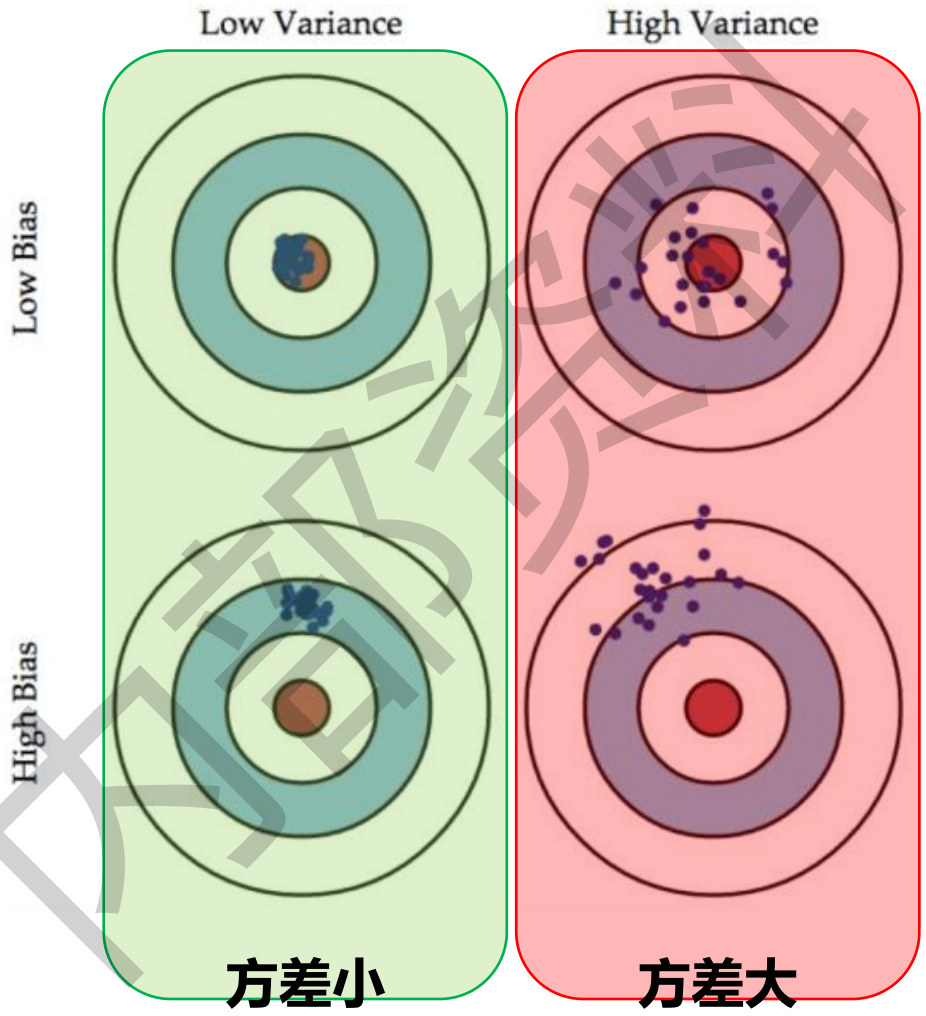


红色的靶心：是模型的正确预测值

蓝色点：训练集所训练出的模型对样本的预测值

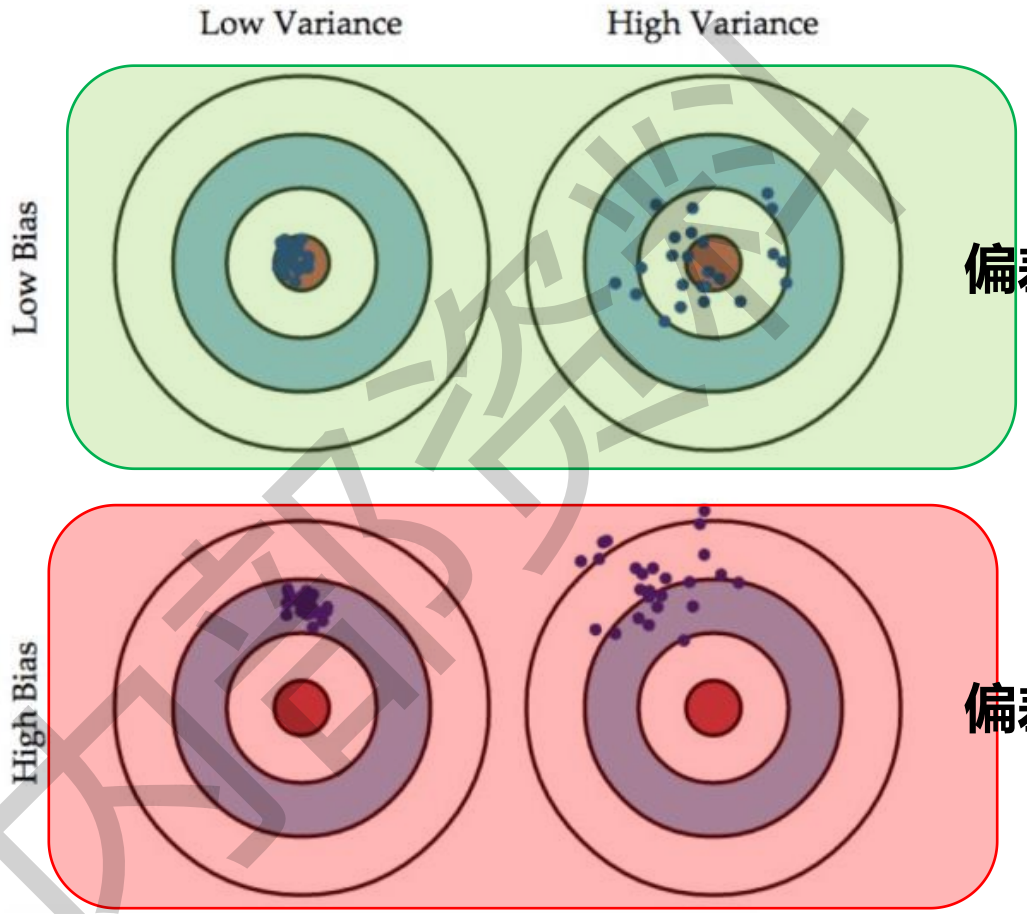
可以看到，当我们从靶心逐渐往外移动时，预测效果逐渐变差。

4 图形解释偏差和方差



	离散度	方差大小
左边	集中	方差较小
右边	分散	方差较大

4 图形解释偏差和方差



偏差小

偏差大

	距离靶心	方差大小
上边	靠近靶心	偏差较小
下边	远离靶心	偏差较大

05 机器学习中的举例

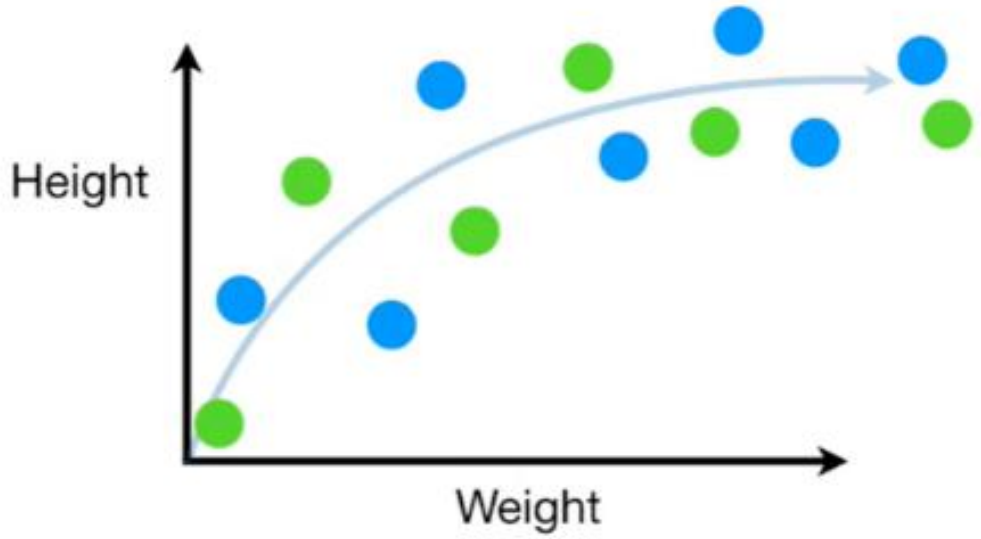
5 机器学习中的举例

基于小鼠的体重预测小鼠的身高:

首先，将所有样本随机分为:

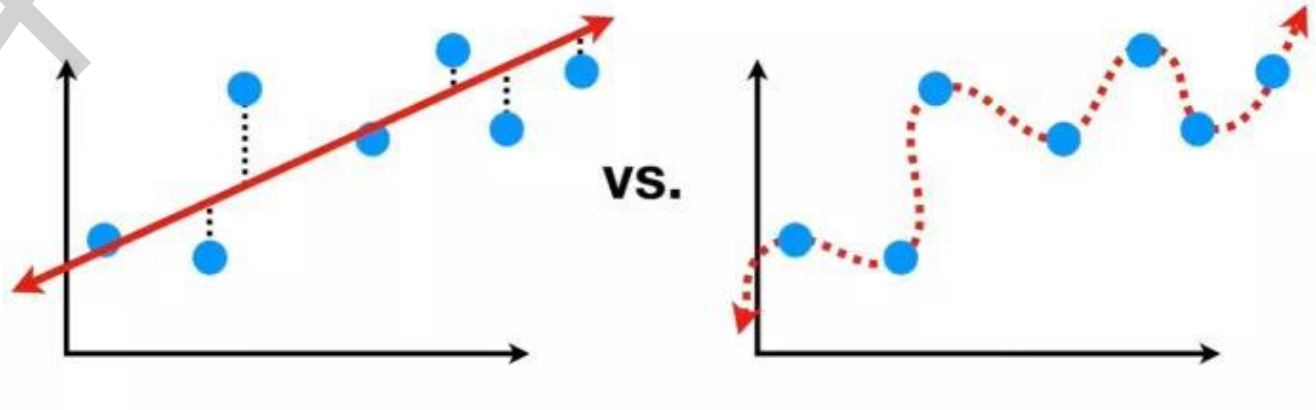
训练样本（蓝色圆点）

测试样本（绿色圆点）



5 机器学习中的举例

基于训练样本训练模型

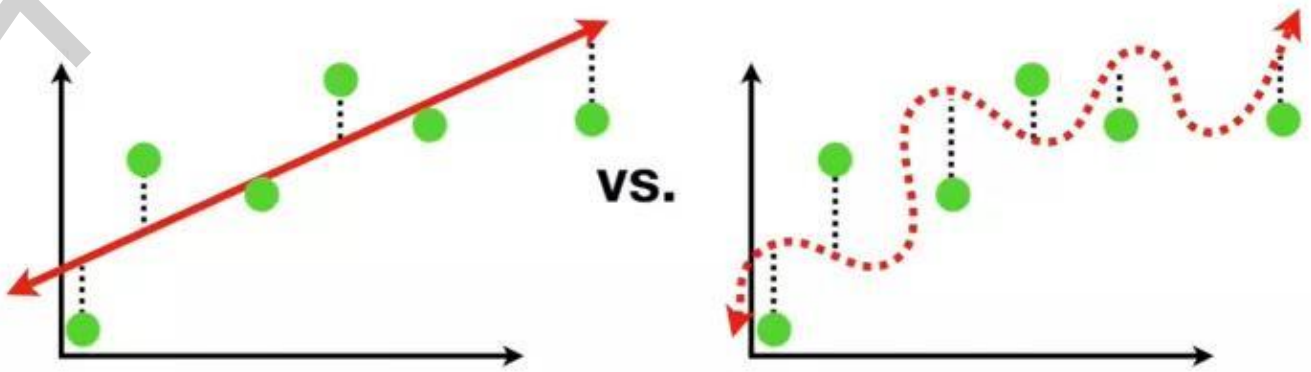


直线回归模型：该算法不能准确描述数据间的真实关系，预测数据与真实数据之间**偏差大**

曲线回归模型：能够在训练数据集中准确的描述数据间真实的关系，该模型的**偏差小**

5 机器学习中的举例

基于测试样本测试模型

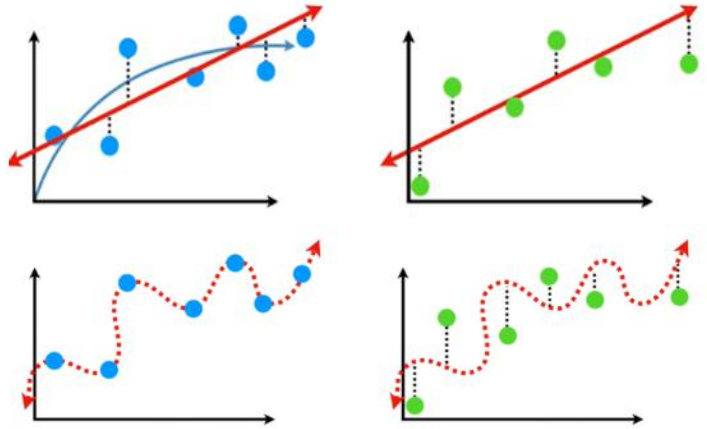


关注某个机器学习算法在训练样本以外的数据（即测试样本）的表现

同一模型在不同数据集间拟合效果（fits）的差异称为方差

拟合直线的方差：直线拟合方差**较小**

拟合曲线的方差：曲线拟合方差**较大**



5 机器学习中的举例

在机器学习中，理想的算法：

具有较小的偏差，能够准确的描述真实数据间的关系；

具有较小的方差，能够在不同数据集间表现出一致的预测性能。

简单来说，我们可以在简单模型和复杂模型（过拟合模型）之间选择一个折中模型，以权衡偏差与方差，实现这一目标的方法有适当增减特征数量和复杂化或简单化模型，以及正则化（regularization）、加入随机因子（如boosting和bagging）。我们将在后续的学习中一一学习。

百度智能云

THANK YOU

CLOUD.BAIDU.COM ABCXUEYUAN.BAIDU.COM