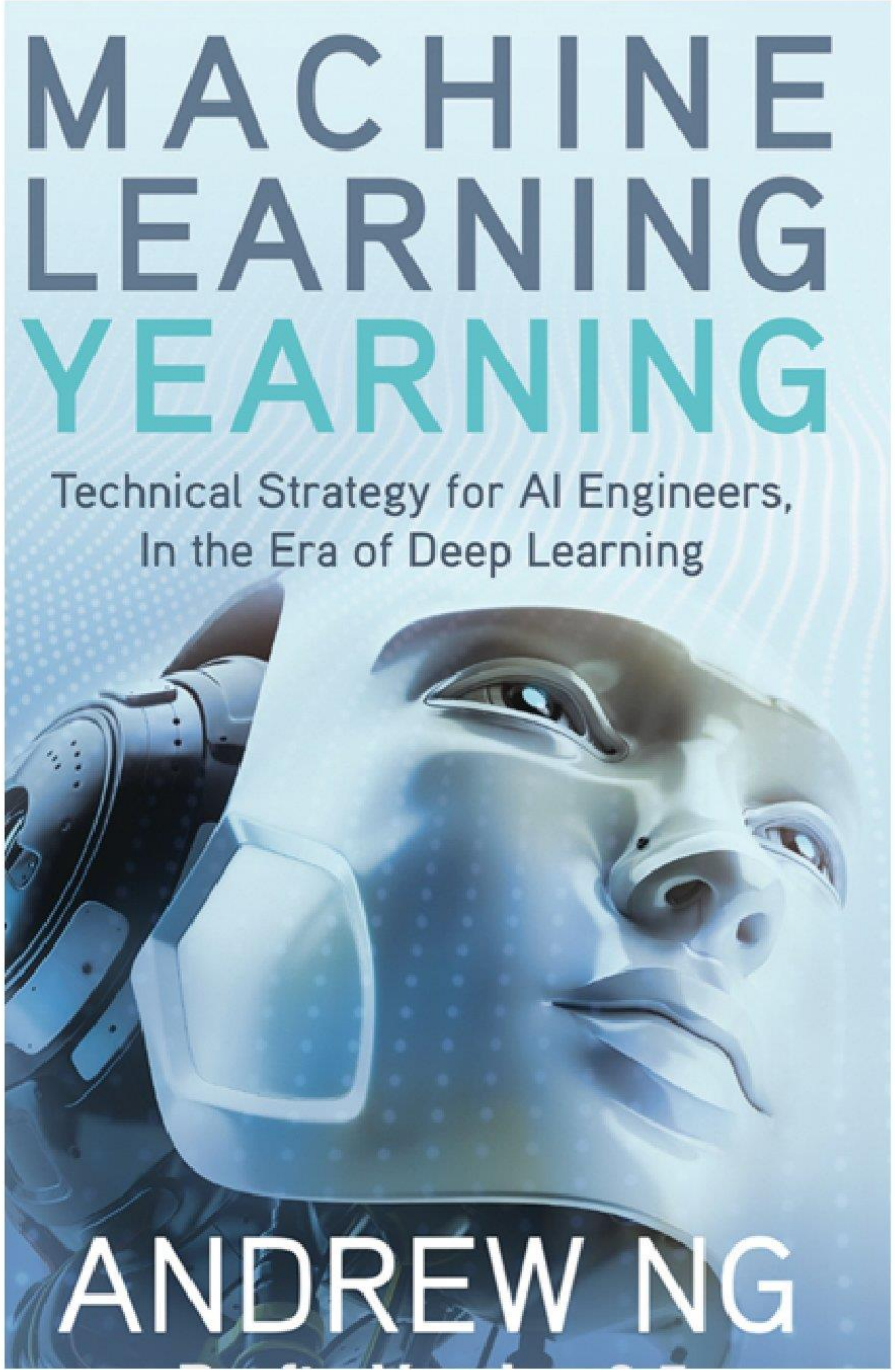
机器学习基础


MACHINE
LEARNING
YEARNING
Technical Strategy for AI Engineers,
In the Era of Deep Learning
ANDREW NG

机器学习：

对于某个任务T和性能度量P，一个计算机程序被认为可以从经验E中学习是指：通过经验E改进后，它在任务T上由性能度量P衡量的性能有提升。

机器学习三要素：

1. 任务T；
2. 性能度量P；
3. 经验E

分类：在这类任务中，计算机程序需要指定某些输入属于 k 类中的哪一类。为了完成这个任务，学习算法通常会返回一个函数 $f: \mathbb{R}^n \rightarrow \{1, \dots, k\}$。当 $y = f(\boldsymbol{x})$ 时，模型将向量 $\boldsymbol{x}$ 所代表的输入分类到数字码 y 所代表的类别。还有一些其他的分类问题，例如，f 输出的是不同类别的概率分布。分类任务中有一个任务是对象识别，其中输入是图片（通常由一组像素亮度值表示），输出是表示图片物体的数字码。例如，Willow Garage PR2 机器人能像服务员一样识别不同饮料，并送给点餐的顾客 (Goodfellow *et al.*, 2010)。目前，最好的对象识别工作正是基于深度学习的 (Krizhevsky *et al.*, 2012a; Ioffe and Szegedy, 2015)。对象识别同时也是计算机识别人脸的基本技术，可用于标记相片合辑中的人脸 (Taigman *et al.*, 2014)，有助于计算机更自然地与用户交互。

- **回归**：在这类任务中，计算机程序需要对给定输入预测数值。为了解决这个任务，学习算法需要输出函数 $f:\mathbb{R}^n \to \mathbb{R}$。除了返回结果的形式不一样外，这类问题和分类问题是很像的。这类任务的一个示例是预测投保人的索赔金额（用于设置保险费），或者预测证券未来的价格。这类预测也用在算法交易中。

转录：这类任务中，机器学习系统观测一些相对非结构化表示的数据，并转录信息为离散的文本形式。例如，光学字符识别要求计算机程序根据文本图片返回文字序列（ASCII 码或者 Unicode 码）。谷歌街景以这种方式使用深度学习处理街道编号 (Goodfellow *et al.*, 2014d)。另一个例子是语音识别，计算机程序输入一段音频波形，输出一序列音频记录中所说的字符或单词 ID 的编码。深度学习是现代语音识别系统的重要组成部分，被各大公司广泛使用，包括微软，IBM 和谷歌 (Hinton *et al.*, 2012a)。

机器翻译：在机器翻译任务中，输入是一种语言的符号序列，计算机程序必须将其转化成另一种语言的符号序列。这通常适用于自然语言，如将英语译成法语。最近，深度学习已经开始在这个任务上产生重要影响 (Sutskever *et al.*, 2014; Bahdanau *et al.*, 2015)。

结构化输出：输出是向量或者包含多个值的数据结构。如语法分析，图像像素级分割，用句子描述图像。

异常检测：检测一个不符合规范的数据或者过程，如信用卡欺诈，网络攻击异常检测，医疗异常检测，

合成和采用：生成一些和训练数据相似的数据样本。如图像生成、语音合成。

缺失值填充：根据数据统计规律，为一个缺失数据样本填充值。如多源异构数据的融合处理。

去噪：在这类任务中，机器学习算法的输入是，干净样本 $\boldsymbol{x} \in \mathbb{R}^n$ 经过未知损坏过程后得到的损坏样本 $\tilde{\boldsymbol{x}} \in \mathbb{R}^n$。算法根据损坏后的样本 $\tilde{\boldsymbol{x}}$ 预测干净的样本 $\boldsymbol{x}$，或者更一般地预测条件概率分布 $p(\boldsymbol{x} \mid \tilde{\boldsymbol{x}})$。

密度估计或概率质量函数估计：在密度估计问题中，机器学习算法学习函数 $p_{\text{model}}: \mathbb{R}^n \to \mathbb{R}$，其中 $p_{\text{model}}(x)$ 可以解释成样本采样空间的概率密度函数（如果x 是连续的）或者概率质量函数（如果x 是离散的）。要做好样的任务,算法需要学习观测到的数据的结构。算法必须知道什么情况下样本聚集出现，什么情况下不太可能出现。以上描述的大多数任务都要求学习算法至少能隐式地捕获概率分布的结构。密度估计可以让我们显式地捕获该分布。原则上，我们可以在该分布上计算以便解决其他任务。例如，如果我们通过密度估计得到了概率分布p(x)，我们可以用该分布解决缺失值填补任务。 。如果xi 的值是缺失的，但是其他的变量值 x_{-i} 已知，那么我们可以得到条件概率分布 $p(x_i \mid x_{-i})$

性能度量P：

1. 准确率；2. 错误率；3. 样本概率的对数平均值

训练集：用来训练机器学习系统的数据集；

测试集：训练后的系统在一个数据集的性能度量。

经验E：从数据集中获取经验。

数据集由许多样本组成，样本也称为数据点

无监督学习：数据集无标签，如数据的生成分布，密度函数，去燥

有监督学习：数据集中样本含有标签。

例子：线性回归

$$\hat{y} = \boldsymbol{w}^\top \boldsymbol{x},$$

任务T ：通过输出 $\hat{y} = \boldsymbol{w}^\top \boldsymbol{x}$ 从 $\boldsymbol{x}$ 预测 y，输入x，预测输出y

性能度量：测试集上的均方误差

$$\text{MSE}_{\text{test}} = \frac{1}{m}\sum_i (\hat{\boldsymbol{y}}^{(\text{test})} - \boldsymbol{y}^{(\text{test})})_i^2.$$

例子：线性回归一个机器学习算法：找到最优W，使得误差最小。

采用梯度下降算法：

最小化 $\mathrm{MSE}_{\mathrm{train}}$，我们可以简单地求解其导数为 **0** 的情况：

$$\nabla_{w}\mathrm{MSE}_{\mathrm{train}} = 0$$

$$\Rightarrow \nabla_{w}\frac{1}{m}\left\|\hat{\boldsymbol{y}}^{(\mathrm{train})} - \boldsymbol{y}^{(\mathrm{train})}\right\|_2^2 = 0$$

$$\Rightarrow \frac{1}{m}\nabla_{w}\left\|\boldsymbol{X}^{(\mathrm{train})}\boldsymbol{w} - \boldsymbol{y}^{(\mathrm{train})}\right\|_2^2 = 0$$

$$\Rightarrow \nabla_{w}\left(\boldsymbol{X}^{(\mathrm{train})}\boldsymbol{w} - \boldsymbol{y}^{(\mathrm{train})}\right)^{\top}\left(\boldsymbol{X}^{(\mathrm{train})}\boldsymbol{w} - \boldsymbol{y}^{(\mathrm{train})}\right) = 0$$

$$\Rightarrow \nabla_{w}\left(\boldsymbol{w}^{\top}\boldsymbol{X}^{(\mathrm{train})\top}\boldsymbol{X}^{(\mathrm{train})}\boldsymbol{w} - 2\boldsymbol{w}^{\top}\boldsymbol{X}^{(\mathrm{train})\top}\boldsymbol{y}^{(\mathrm{train})} + \boldsymbol{y}^{(\mathrm{train})\top}\boldsymbol{y}^{(\mathrm{train})}\right) = 0$$

$$\Rightarrow 2\boldsymbol{X}^{(\mathrm{train})\top}\boldsymbol{X}^{(\mathrm{train})}\boldsymbol{w} - 2\boldsymbol{X}^{(\mathrm{train})\top}\boldsymbol{y}^{(\mathrm{train})} = 0$$

$$\Rightarrow \boldsymbol{w} = \left(\boldsymbol{X}^{(\mathrm{train})\top}\boldsymbol{X}^{(\mathrm{train})}\right)^{-1}\boldsymbol{X}^{(\mathrm{train})\top}\boldsymbol{y}^{(\mathrm{train})}$$

线性回归

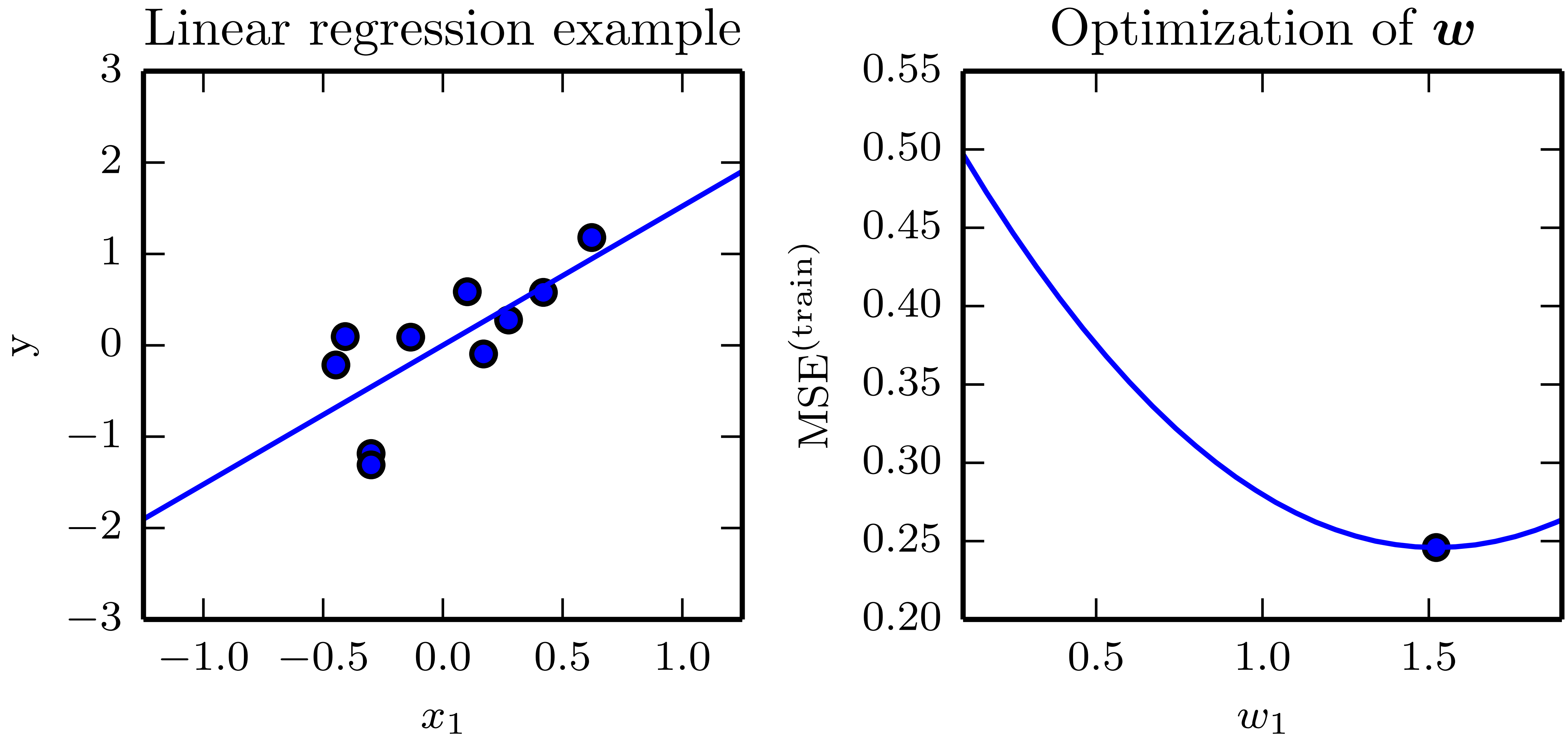


Figure 5.1

(Goodfellow 2016)

学习部件：

1. 任务：回归。

2. 模型： $\hat{y} = \boldsymbol{w}^\top \boldsymbol{x},$

3. 性能度量： $\mathrm{MSE}_{\mathrm{test}} = \frac{1}{m} \sum_i (\hat{\boldsymbol{y}}^{(\mathrm{test})} - \boldsymbol{y}^{(\mathrm{test})})_i^2.$

4. 学习算法：是目标函数最优

5. 经验：数据。

目标：对未观察的数据表现良好，而不是在训练集上表现良好，这里能力称为泛化能力。

通过误差来描述：

训练误差：在训练集上的误差；

泛化误差：在输入空间上的误差（测试误差）；

在我们的线性回归示例中，我们通过最小化训练误差来训练模型，

$$\frac{1}{m^{(\text{train})}} \left\| \boldsymbol{X}^{(\text{train})} \boldsymbol{w} - \boldsymbol{y}^{(\text{train})} \right\|_2^2,$$

但是我们真正关注的是测试误差 $\frac{1}{m^{(\text{test})}} \left\| \boldsymbol{X}^{(\text{test})} \boldsymbol{w} - \boldsymbol{y}^{(\text{test})} \right\|_2^2$。

如果数据独立来自同一分布，则有可能通过减少训练误差来达到减少测试误差。

统计学习假设：所有的数据都独立同分布的，这个分布称为数据生成分布。

目标：1. 降低训练误差；2. 缩小训练误差和泛化误差差距。

欠拟合：模型在训练集上的误差不够低；

过拟合：训练误差与泛化误差差距太大。

容量：模型能拟合各种函数的能力。容量低的模型很难拟合训练集导致欠拟合，容量高的模型导致过拟合。通过调节容量达到两者平衡。

线性回归：

模型容量：

1. 一元多项式：

$$\hat{y} = b + wx.$$

2. 二元多项式：

$$\hat{y} = b + w_1 x + w_2 x^2.$$

3. 九次多项式：

$$\hat{y} = b + \sum_{i=1}^{9} w_i x^i.$$

欠拟合 与过拟合

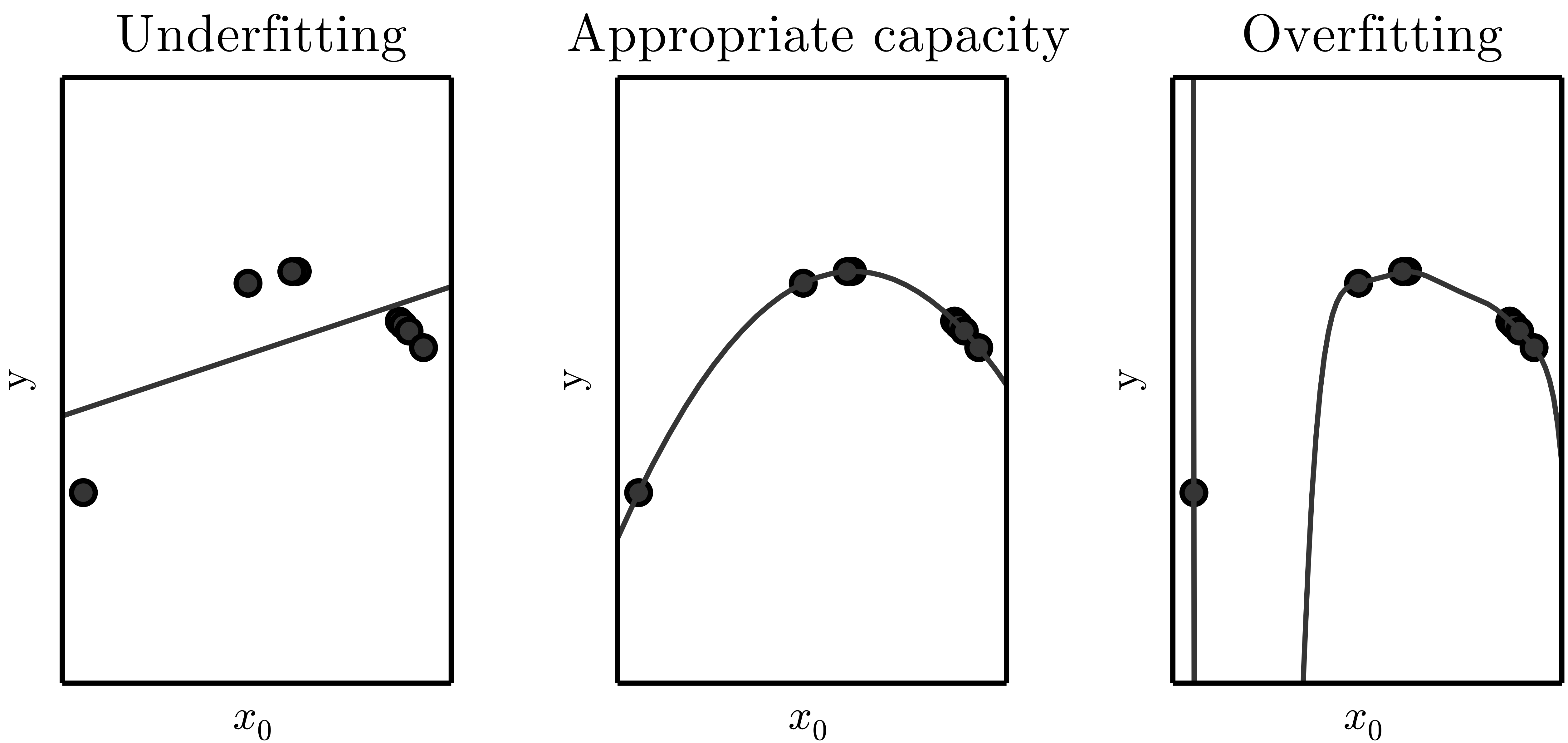


Figure 5.2

(Goodfellow 2016)

模型的表示容量：模型能表达的函数簇

模型的有效容量：由于算法，使得实际搜索的函数集是表示容量子集。

容量的度量：VC维理论

结论：训练误差与泛化误差随模型容量增加而增加，随着样本增加而减少。

泛化与容量

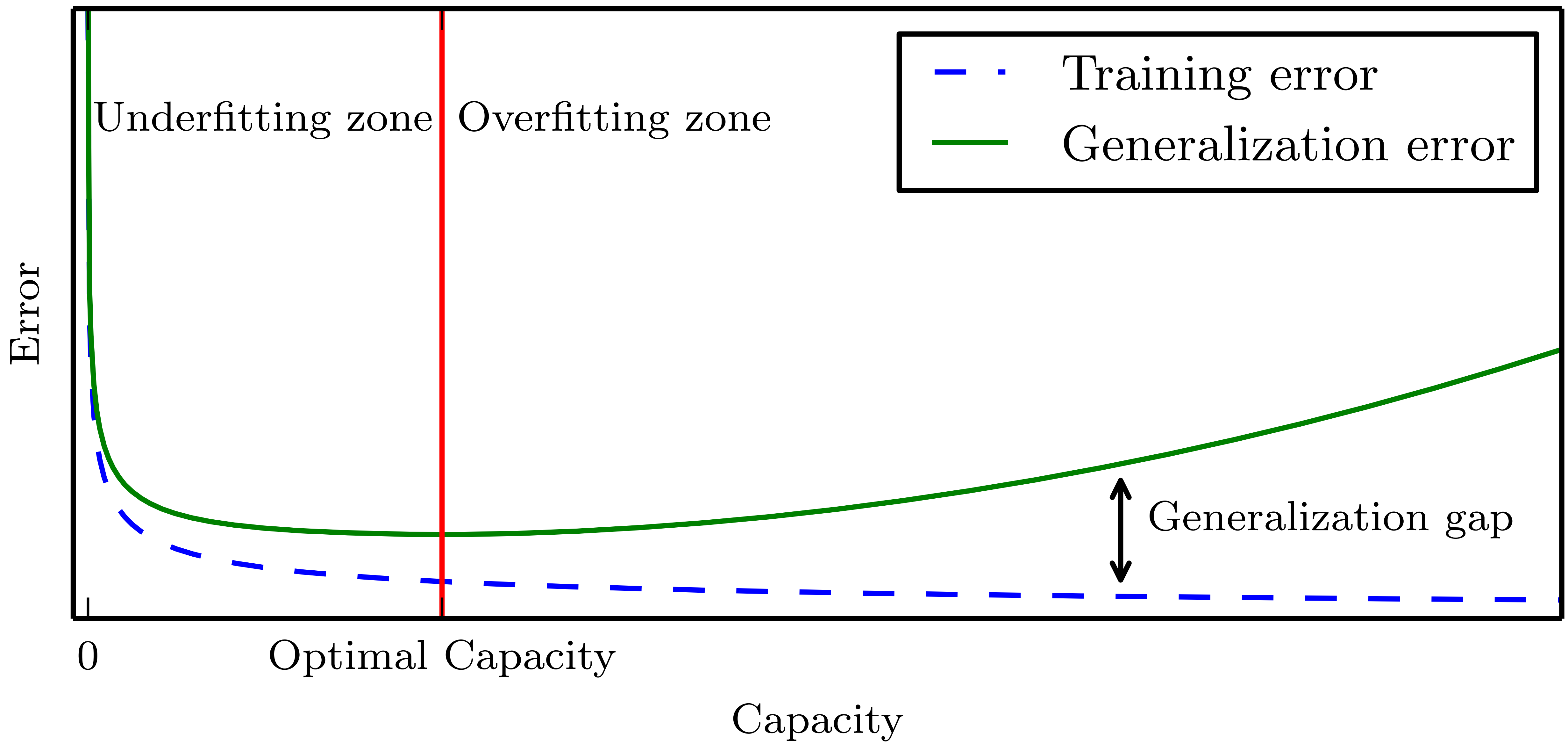


Figure 5.3

(Goodfellow 2016)

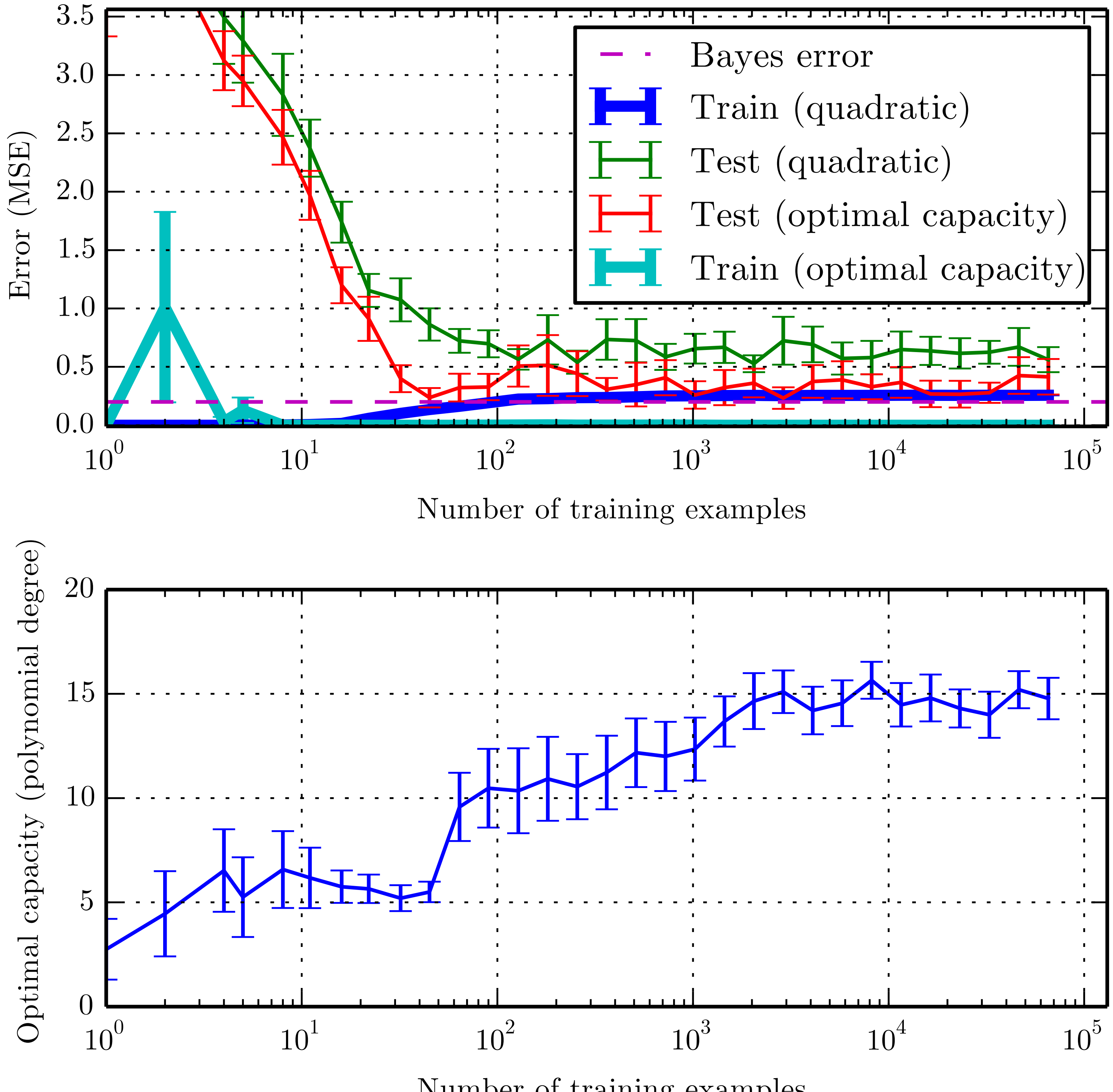
没有免费的午餐定理

任意的分类算法在所有的分布上的平均错误率是相等的。

实际：根据任务，对真实数据分布进行假设，设计最优的模型和算法。

测试集大小与训练误差

Figure 5.4



(Goodfellow 2016)

正则化：

当模型容量过大时，训练集误差小，但泛化误差大。为此，通过引入偏好或者限制，使得模型有效容量减少，从而减少训练误差和泛化误差的距离。

如：权值衰减

$$J(\boldsymbol{w}) = \mathrm{MSE}_{\mathrm{train}} + \lambda \boldsymbol{w}^{\top} \boldsymbol{w},$$

权值衰减

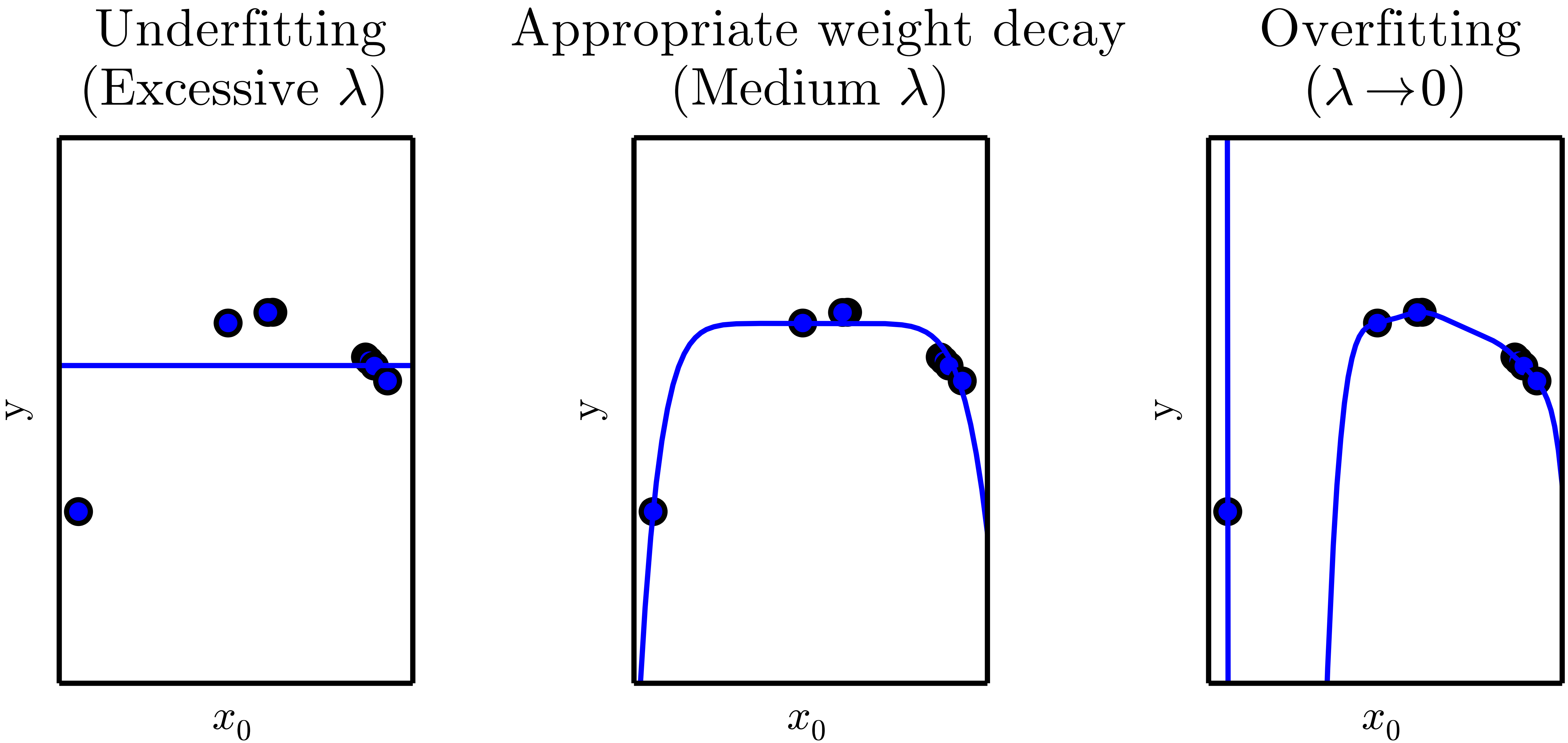


Figure 5.5

(Goodfellow 2016)

超参数与测试集：

不是被学习的参数，而是来控制学习算法行为的参数。

如权值衰减中的超参数。

如何选择超参数？

通过设置验证集来选择：在验证集上最优的超参数。

数据被分为：训练集，测试集，验证集。如果数据少怎么办？

交叉验证。

采用统计观点来描述泛化、欠拟合和过拟合

1. 点估计

对一个感兴趣的量进行单个“最优”预测。

为了区分参数估计和真实值，我们习惯将参数 $\boldsymbol{\theta}$ 的点估计表示为 $\hat{\boldsymbol{\theta}}$

令 $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$ 是 m 个独立同分布（i.i.d.）的数据点。**点估计**（point estimator）或**统计量**（statistics）是这些数据的任意函数：

$$\hat{\boldsymbol{\theta}}_m = g(\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}). \tag{5.19}$$

这个定义不要求 g 返回一个接近真实 $\boldsymbol{\theta}$ 的值，或者 g 的值域恰好是 $\boldsymbol{\theta}$ 的允许取值范围。点估计的定义非常宽泛，给了估计量的设计者极大的灵活性。虽然几乎所有的函数都可以称为估计量，但是一个良好的估计量的输出会接近生成训练数据的真实参数 $\boldsymbol{\theta}$。

频率派的观点：

1. 真实的参数θ是固定但未知，而点估计是数据的函数。

2. 由于数据是随机采用的，所以估计值是一个随机变量。

函数估计 有时我们会关注函数估计（或函数近似）。这时我们试图从输入向量 $\boldsymbol{x}$ 预测变量 $\boldsymbol{y}$。我们假设有一个函数 $f(\boldsymbol{x})$ 表示 $\boldsymbol{y}$ 和 $\boldsymbol{x}$ 之间的近似关系。例如，我们可能假设 $\boldsymbol{y} = f(\boldsymbol{x}) + \boldsymbol{\epsilon}$，其中 $\boldsymbol{\epsilon}$ 是 $\boldsymbol{y}$ 中未能从 $\boldsymbol{x}$ 预测的一部分。在函数估计中，我们感兴趣的是用模型估计去近似 f，或者估计 $\hat{f}$。函数估计和估计参数 $\boldsymbol{\theta}$ 是一样的；函数估计 $\hat{f}$ 是函数空间中的一个点估计。线性回归示例（第 5.1.4 节中讨论的）和多项式回归示例（第 5.2 节中讨论的）都既可以被解释为估计参数 $\boldsymbol{w}$，又可以被解释为估计从 $\boldsymbol{x}$ 到 y 的函数映射 $\hat{f}$。

偏差:

估计的偏差被定义为:

$$\mathrm{bias}(\hat{\boldsymbol{\theta}}_m) = \mathbb{E}(\hat{\boldsymbol{\theta}}_m) - \boldsymbol{\theta}, \tag{5.20}$$

其中期望作用在所有数据(看作是从随机变量采样得到的)上,$\boldsymbol{\theta}$ 是用于定义数据生成分布的 $\boldsymbol{\theta}$ 的真实值。如果 $\mathrm{bias}(\hat{\boldsymbol{\theta}}_m) = 0$,那么估计量 $\hat{\boldsymbol{\theta}}_m$ 被称为是**无偏**(unbiased),这意味着 $\mathbb{E}(\hat{\boldsymbol{\theta}}_m) = \boldsymbol{\theta}$。如果 $\lim_{m\to\infty} \mathrm{bias}(\hat{\boldsymbol{\theta}}_m) = 0$,那么估计量 $\hat{\boldsymbol{\theta}}_m$ 被称为是**渐近无偏**(asymptotically unbiased),这意味着 $\lim_{m\to\infty} \mathbb{E}(\hat{\boldsymbol{\theta}}_m) = \boldsymbol{\theta}$。

示例:伯努利分布 考虑一组服从均值为 θ 的伯努利分布的独立同分布的样本 $\{x^{(1)}, \ldots, x^{(m)}\}$:

$$P(x^{(i)}; \theta) = \theta^{x^{(i)}} (1 - \theta)^{(1 - x^{(i)})}. \tag{5.21}$$

这个分布中参数 θ 的常用估计量是训练样本的均值:

$$\hat{\theta}_m = \frac{1}{m} \sum_{i=1}^{m} x^{(i)}. \tag{5.22}$$

偏差：贝努力分布参数无偏估计

$$
\begin{aligned}
\operatorname{bias}(\hat{\theta}_m) &= \mathbb{E}[\hat{\theta}_m] - \theta \\
&= \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} x^{(i)}\right] - \theta \\
&= \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}\left[x^{(i)}\right] - \theta \\
&= \frac{1}{m}\sum_{i=1}^{m} \sum_{x^{(i)}=0}^{1} \left(x^{(i)}\theta^{x^{(i)}}(1-\theta)^{(1-x^{(i)})}\right) - \theta \\
&= \frac{1}{m}\sum_{i=1}^{m}(\theta) - \theta \\
&= \theta - \theta = 0
\end{aligned}
$$

例子：高斯分布的均值无偏估计

示例：均值的高斯分布估计 现在，考虑一组独立同分布的样本 $\{x^{(1)}, \dots, x^{(m)}\}$ 服从高斯分布 $p(x^{(i)}) = \mathcal{N}(x^{(i)}; \mu, \sigma^2)$，其中 $i \in \{1, \dots, m\}$。回顾高斯概率密度函数如下：

$$p(x^{(i)}; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x^{(i)} - \mu)^2}{\sigma^2}\right). \tag{5.29}$$

高斯均值参数的常用估计量被称为 **样本均值**（sample mean）：

$$\hat{\mu}_m = \frac{1}{m}\sum_{i=1}^{m} x^{(i)} \tag{5.30}$$

例子：高斯分布的均值无偏估计

$$
\begin{aligned}
\mathrm{bias}(\hat{\mu}_m) &= \mathbb{E}[\hat{\mu}_m] - \mu \\
&= \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} x^{(i)}\right] - \mu \\
&= \left(\frac{1}{m}\sum_{i=1}^{m} \mathbb{E}\left[x^{(i)}\right]\right) - \mu \\
&= \left(\frac{1}{m}\sum_{i=1}^{m} \mu\right) - \mu \\
&= \mu - \mu = 0
\end{aligned}
$$

例子：高斯分布的方差有偏估计

示例：高斯分布方差估计 本例中，我们比较高斯分布方差参数 σ^2 的两个不同估计。我们探讨是否有一个是有偏的。

我们考虑的第一个方差估计被称为 **样本方差**（sample variance）：

$$\hat{\sigma}_m^2 = \frac{1}{m}\sum_{i=1}^{m}\left(x^{(i)} - \hat{\mu}_m\right)^2, \tag{5.36}$$

例子：高斯分布的方差有偏估计

$$\mathbb{E}[\hat{\sigma}_m^2] = \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}\left(x^{(i)} - \hat{\mu}_m\right)^2\right]$$
$$= \frac{m-1}{m}\sigma^2$$

偏差为： $-\sigma^2/m$

高斯分布的无偏方差估计：

$$\tilde{\sigma}_m^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(x^{(i)} - \hat{\mu}_m\right)^2$$

无偏估计非常好，但实际应用中我们需要满足某种性质的有偏估计

估计的方差和标准差：

估计量的方差就是该估计量的方差：

$$\mathrm{Var}(\hat{\theta})$$

其中训练集是随机变量。

方差的平方根是标准差：

$$\mathrm{SE}(\hat{\theta})$$

均值的标准差被记作

$$\mathrm{SE}(\hat{\mu}_m) = \sqrt{\mathrm{Var}\left[\frac{1}{m}\sum_{i=1}^{m} x^{(i)}\right]} = \frac{\sigma}{\sqrt{m}},$$

示例：伯努利分布 我们再次考虑从伯努利分布（回顾 $P(x^{(i)};\theta)=\theta^{x^{(i)}}(1-\theta)^{1-x^{(i)}}$）中独立同分布采样出来的一组样本 $\{x^{(1)},\ldots,x^{(m)}\}$。这次我们关注估计 $\hat{\theta}_m=\frac{1}{m}\sum_{i=1}^{m}x^{(i)}$ 的方差：

$$\begin{aligned}
\operatorname{Var}\left(\hat{\theta}_m\right) &= \operatorname{Var}\left(\frac{1}{m}\sum_{i=1}^{m}x^{(i)}\right)\\
&= \frac{1}{m^2}\sum_{i=1}^{m}\operatorname{Var}\left(x^{(i)}\right)\\
&= \frac{1}{m^2}\sum_{i=1}^{m}\theta(1-\theta)\\
&= \frac{1}{m^2}m\theta(1-\theta)\\
&= \frac{1}{m}\theta(1-\theta)
\end{aligned}$$

最小化均方误差

问题：存在一个偏差大的估计和另外一个方差大的估计，我们如何选择？

解决方法：

1. 交叉验证；
2. 最小化均方误差

均方误差定义为：

$$
\begin{aligned}
\mathrm{MSE} &= \mathbb{E}[(\hat{\theta}_m - \theta)^2] \\
&= \mathrm{Bias}(\hat{\theta}_m)^2 + \mathrm{Var}(\hat{\theta}_m)
\end{aligned}
$$

偏差、方差与拟合关系如下图：

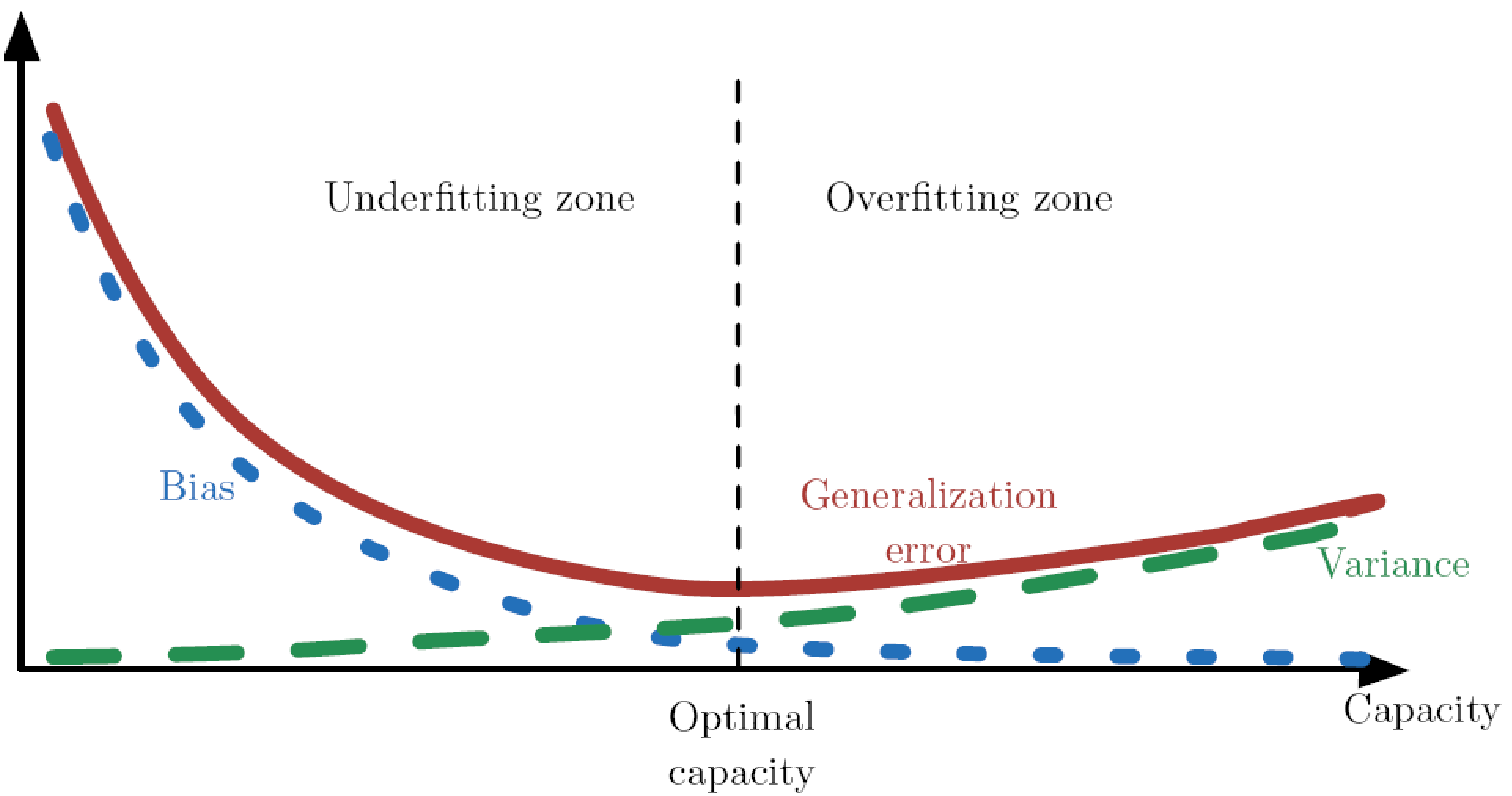
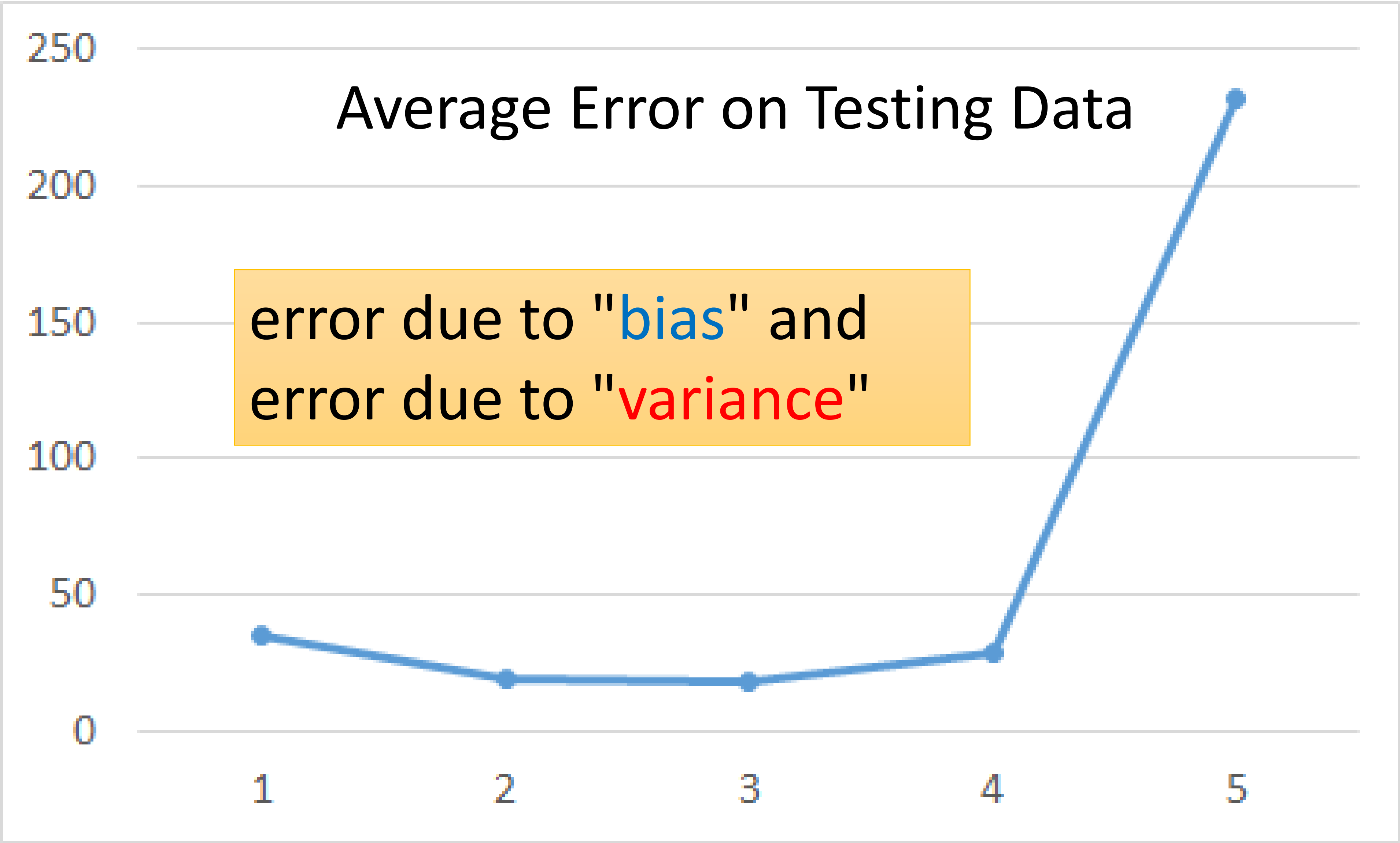


图 5.6: 当容量增大（x 轴）时，偏差（用点表示）随之减小，而方差（虚线）随之增大，使得泛化误差（加粗曲线）产生了另一种 U 形。如果我们沿着轴改变容量，会发现最佳容量，当容量小于最佳容量会呈现欠拟合，大于时导致过拟合。这种关系与第 5.2 节以及图 5.3 中讨论的容量、欠拟合和过拟合之间的关系类似。

线性回归例子



A more complex model does not always lead to better performance on ***testing data***.

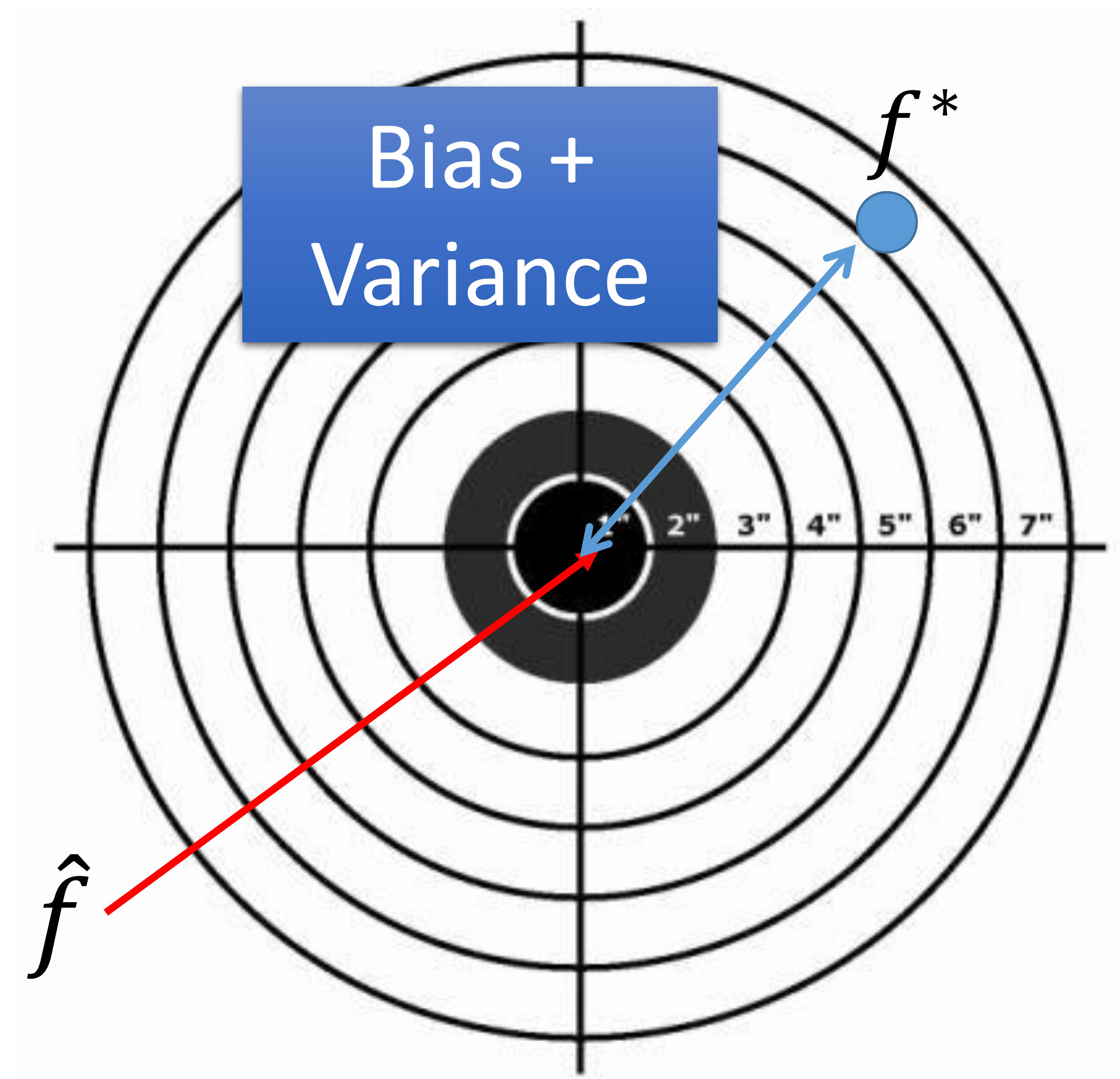
Estimator

$\hat{y} = \hat{f}($  $)$

Only Niantic knows $\hat{f}$

From training data,
we find f^*

f^* is an estimator of $\hat{f}$



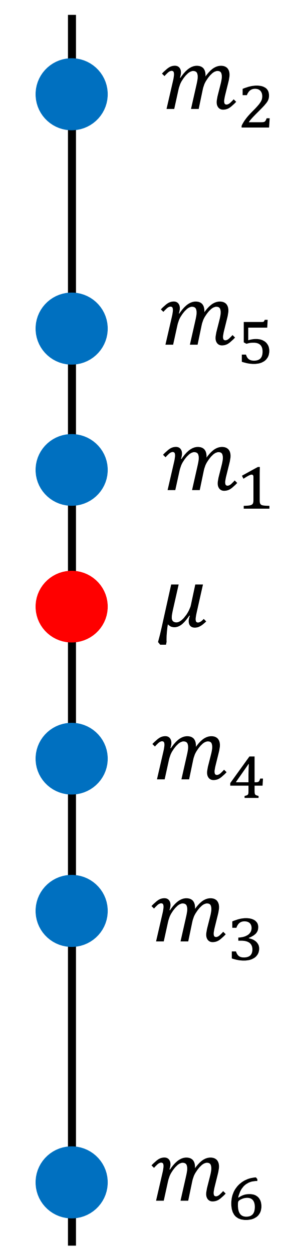
Bias and Variance of Estimator

- Estimate the mean of a variable x
  - assume the mean of x is μ
  - assume the variance of x is σ^2
- Estimator of mean μ
  - Sample N points: $\{x^1, x^2, \dots, x^N\}$

$$m = \frac{1}{N}\sum_n x^n \neq \mu$$

$$E[m] = E\left[\frac{1}{N}\sum_n x^n\right] = \frac{1}{N}\sum_n E[x^n] = \mu$$

unbiased



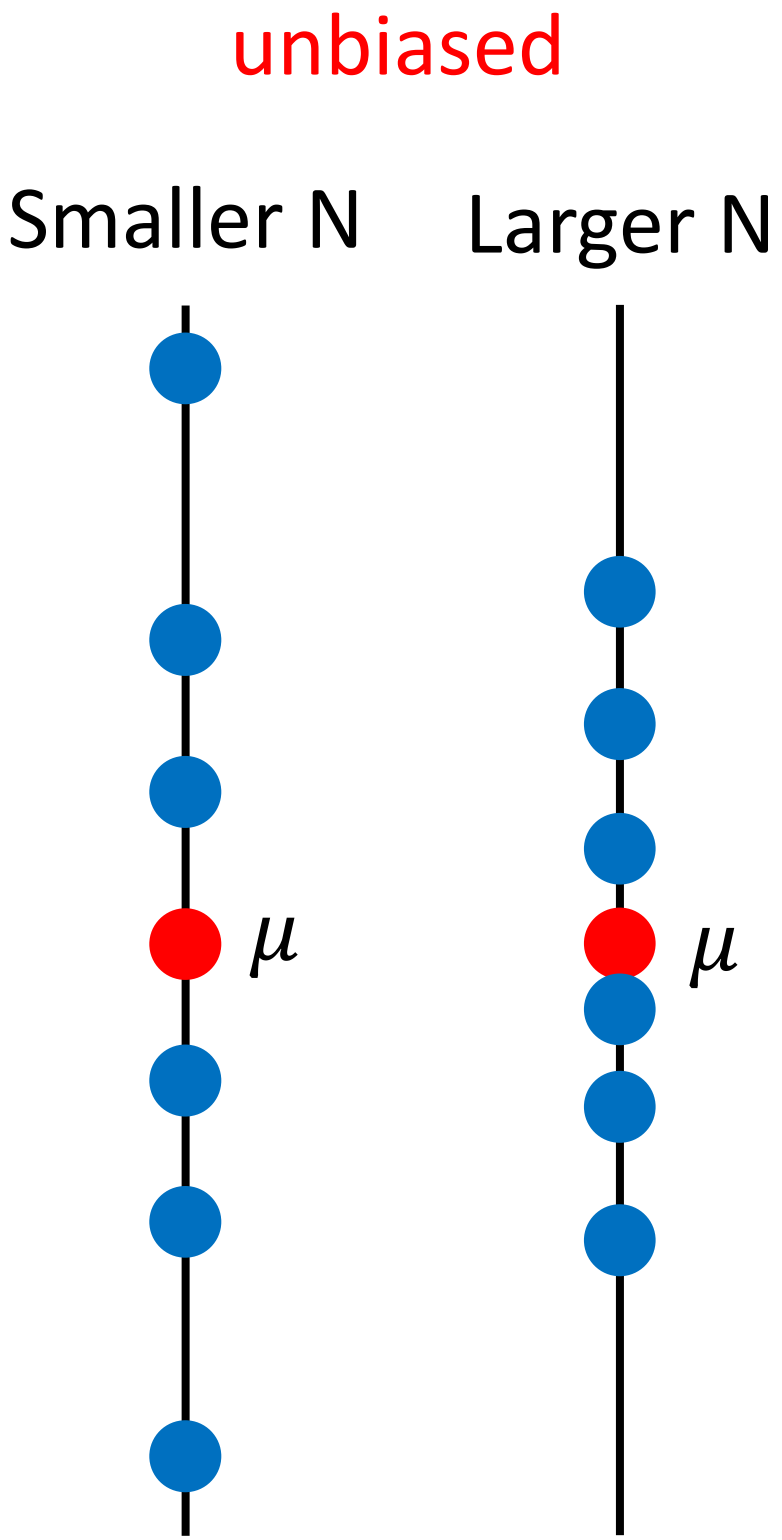
Bias and Variance of Estimator

- Estimate the mean of a variable x
  - assume the mean of x is μ
  - assume the variance of x is σ^2
- Estimator of mean μ
  - Sample N points: $\{x^1, x^2, \ldots, x^N\}$

$$m = \frac{1}{N}\sum_n x^n \neq \mu$$

$$Var[m] = \frac{\sigma^2}{N}$$

Variance depends on the number of samples



Bias and Variance of Estimator

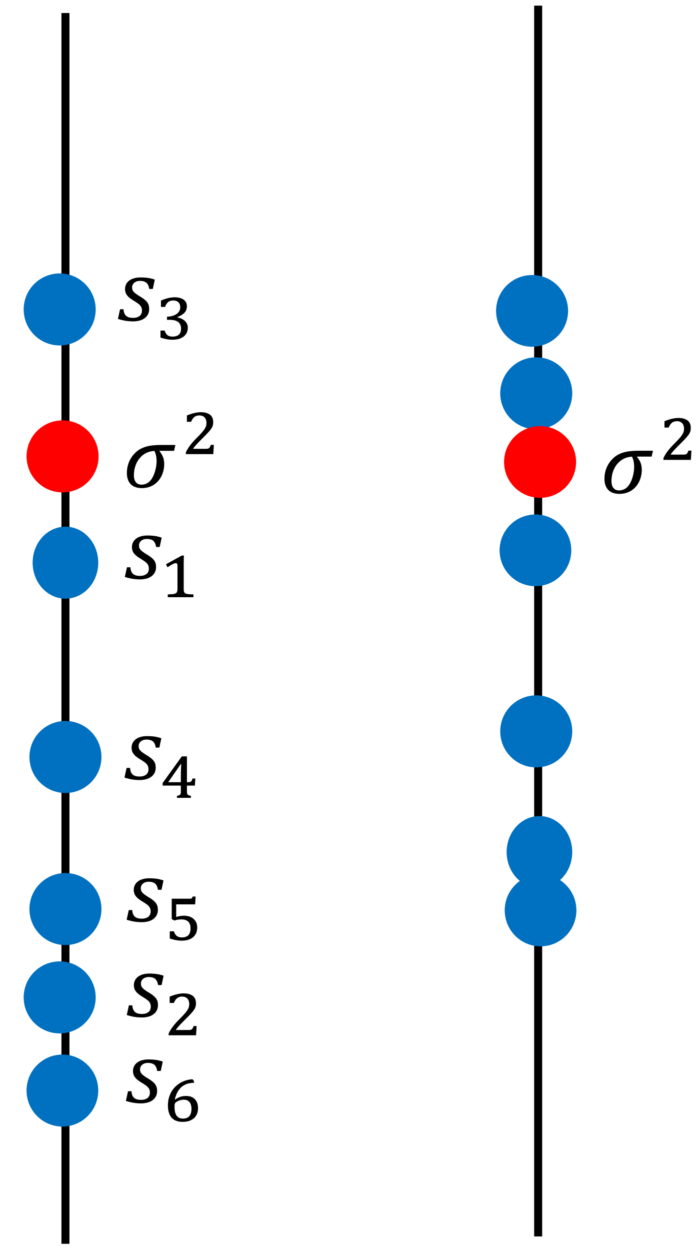
- Estimate the mean of a variable x
  - assume the mean of x is μ
  - assume the variance of x is σ^2
- Estimator of variance σ^2
  - Sample N points: $\{x^1, x^2, \ldots, x^N\}$

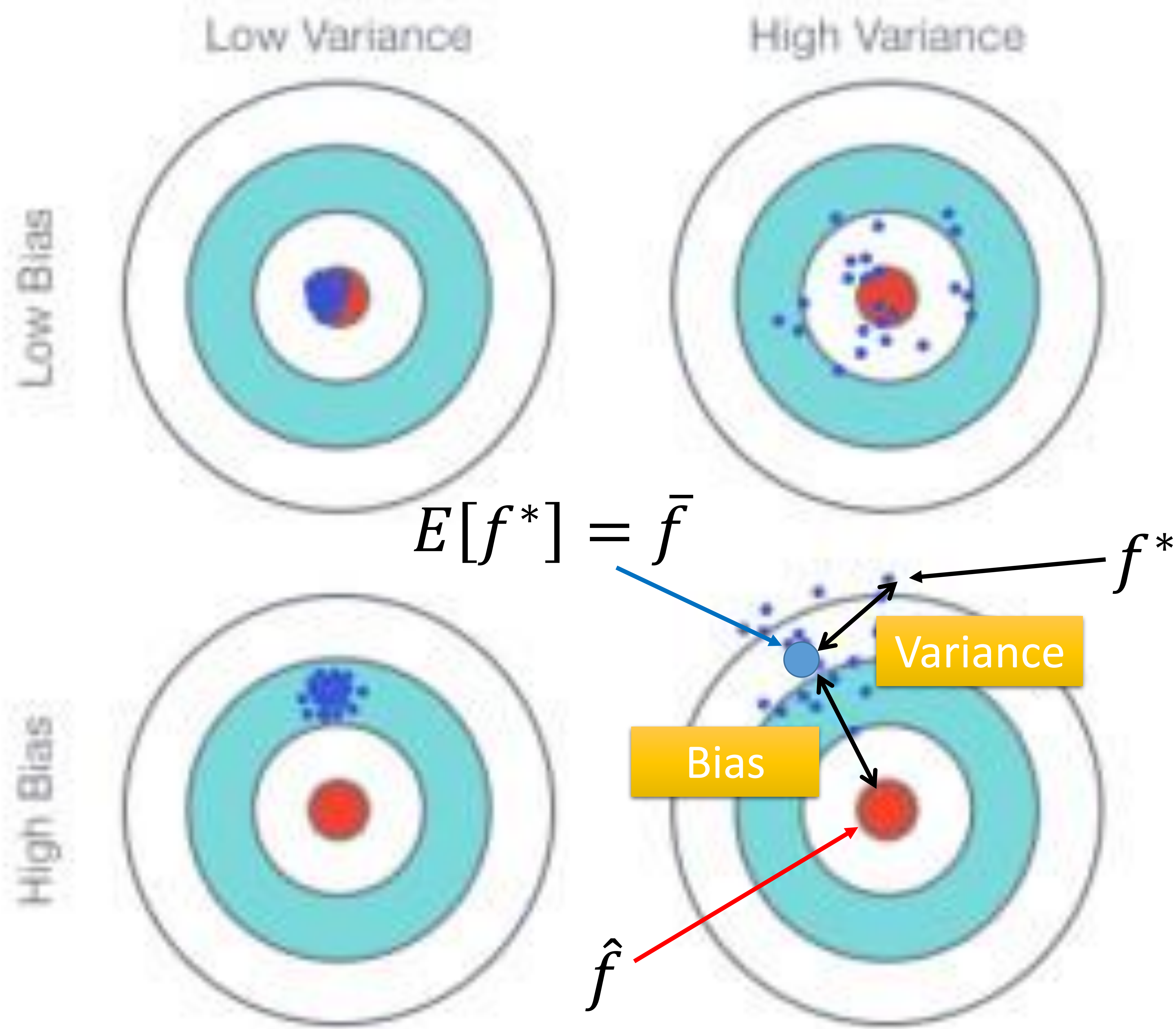
$$m = \frac{1}{N}\sum_n x^n \quad s = \frac{1}{N}\sum_n (x^n - m)^2$$

Biased estimator

$$E[s] = \frac{N-1}{N}\sigma^2 \quad \neq \sigma^2$$

Increase N





Low Variance
High Variance
Low Bias
High Bias
$E[f^*] = \bar{f}$
f^*
Variance
Bias
$\hat{f}$

一致性：

目前我们已经探讨了固定大小训练集下不同估计量的性质。通常，我们也会关注训练数据增多后估计量的效果。特别地，我们希望当数据集中数据点的数量 m 增加时，点估计会收敛到对应参数的真实值。更形式地，我们想要

$$\mathrm{plim}_{m\to\infty}\hat{\theta}_m = \theta. \tag{5.55}$$

符号 plim 表示依概率收敛，即对于任意的 $\epsilon > 0$，当 $m \to \infty$ 时，有 $P(|\hat{\theta}_m - \theta| > \epsilon) \to 0$。式 (5.55) 表示的条件被称为**一致性**（consistency）。有时它是指弱一致性，强一致性是指**几乎必然**（almost sure）从 $\hat{\theta}$ 收敛到 θ。**几乎必然收敛**（almost sure convergence）是指当 $p(\lim_{m\to\infty} \mathbf{x}^{(m)} = \boldsymbol{x}) = 1$ 时，随机变量序列 $\mathbf{x}^{(1)}$，$\mathbf{x}^{(2)}$，$\ldots$ 收敛到 $\boldsymbol{x}$。

背景

之前，我们已经看过常用估计的定义，并分析了它们的性质。但是这些估计是从哪里来的呢？我们希望有些准则可以让我们从不同模型中得到特定函数作为好的估计，而不是猜测某些函数可能是好的估计，然后分析其偏差和方差。

最常用的准则是最大似然估计。

考虑一组含有 m 个样本的数据集 $\mathbb{X} = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$，独立地由未知的真实数据生成分布 $p_{\text{data}}(\mathbf{x})$ 生成。

令 $p_{\text{model}}(\mathbf{x}; \boldsymbol{\theta})$ 是一族由 $\boldsymbol{\theta}$ 确定在相同空间上的概率分布。换言之，$p_{\text{model}}(\boldsymbol{x}; \boldsymbol{\theta})$ 将任意输入 $\boldsymbol{x}$ 映射到实数来估计真实概率 $p_{\text{data}}(\boldsymbol{x})$。

原理

对 $\boldsymbol{\theta}$ 的最大似然估计被定义为：

$$\boldsymbol{\theta}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} p_{\mathrm{model}}(\mathbb{X}; \boldsymbol{\theta}),$$

$$= \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{m} p_{\mathrm{model}}(\boldsymbol{x}^{(i)}; \boldsymbol{\theta}).$$

对概率计算乘积计算导致下溢，所以在概率的对数空间计算：

$$\boldsymbol{\theta}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{m} \log p_{\mathrm{model}}(\boldsymbol{x}^{(i)}; \boldsymbol{\theta}).$$

除以m，得到：

$$\boldsymbol{\theta}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{x} \sim \hat{p}_{\mathrm{data}}} \log p_{\mathrm{model}}(\boldsymbol{x}; \boldsymbol{\theta})$$

核心观点：

最大似然估计被看作是数据经验分布与模型分布之间差异最小。

用KL散度来描述分布之间最小：

$$D_{\mathrm{KL}}(\hat{p}_{\mathrm{data}} \| p_{\mathrm{model}}) = \mathbb{E}_{\mathbf{x}\sim\hat{p}_{\mathrm{data}}}[\log \hat{p}_{\mathrm{data}}(\boldsymbol{x}) - \log p_{\mathrm{model}}(\boldsymbol{x})].$$

左边项与模型参数无关，可以看作是常数，因此最小化：

$$-\mathbb{E}_{\mathbf{x}\sim\hat{p}_{\mathrm{data}}}[\log p_{\mathrm{model}}(\boldsymbol{x})],$$

也是模型分布在经验分布上交叉熵

最大似然应用于条件概率

最大似然估计很容易扩展到估计条件概率 $P(\mathrm{y} \mid \mathrm{x}; \boldsymbol{\theta})$，从而给定 x 预测 y。实际上这是最常见的情况，因为这构成了大多数监督学习的基础。如果 $\boldsymbol{X}$ 表示所有的输入，$\boldsymbol{Y}$ 表示我们观测到的目标，那么条件最大似然估计是

$$\boldsymbol{\theta}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} P(\boldsymbol{Y} \mid \boldsymbol{X}; \boldsymbol{\theta}). \tag{5.62}$$

如果假设样本是独立同分布的，那么这可以分解成

$$\boldsymbol{\theta}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{m} \log P(\boldsymbol{y}^{(i)} \mid \boldsymbol{x}^{(i)}; \boldsymbol{\theta}). \tag{5.63}$$

例子

示例：线性回归作为最大似然 第 5.1.4 节介绍的线性回归，可以被看作是最大似然过程。之前，我们将线性回归作为学习从输入 $\boldsymbol{x}$ 映射到输出 $\hat{y}$ 的算法。从 $\boldsymbol{x}$ 到 $\hat{y}$ 的映射选自最小化均方误差（我们或多或少介绍的一个标准）。现在，我们以最大似然估计的角度重新审视线性回归。我们现在希望模型能够得到条件概率 $p(y \mid \boldsymbol{x})$，而不只是得到一个单独的预测 $\hat{y}$。想象有一个无限大的训练集，我们可能会观测到几个训练样本有相同的输入 $\boldsymbol{x}$ 但是不同的 y。现在学习算法的目标是拟合分布 $p(y \mid \boldsymbol{x})$ 到和 $\boldsymbol{x}$ 相匹配的不同的 y。为了得到我们之前推导出的相同的线性回归算法，我们定义 $p(y \mid \boldsymbol{x}) = \mathcal{N}(y; \hat{y}(\boldsymbol{x}; \boldsymbol{w}), \sigma^2)$。函数 $\hat{y}(\boldsymbol{x}; \boldsymbol{w})$ 预测高斯的均值。在这个例子中，我们假设方差是用户固定的某个常量 σ^2。这种函数形式 $p(y \mid \boldsymbol{x})$ 会使得最大似然估计得出和之前相同的学习算法。由于假设样本是独立同分布的，条件对数似然（式 (5.63)）如下

例子

$$\sum_{i=1}^{m} \log p(y^{(i)} \mid \boldsymbol{x}^{(i)}; \boldsymbol{\theta})$$

$$= -m \log \sigma - \frac{m}{2} \log(2\pi) - \sum_{i=1}^{m} \frac{\left\| \hat{y}^{(i)} - y^{(i)} \right\|^2}{2\sigma^2},$$

等价：

其中 $\hat{y}^{(i)}$ 是线性回归在第 i 个输入 $\boldsymbol{x}^{(i)}$ 上的输出，m 是训练样本的数目。对比于均方误差的对数似然，

$$\mathrm{MSE}_{\mathrm{train}} = \frac{1}{m} \sum_{i=1}^{m} \left\| \hat{y}^{(i)} - y^{(i)} \right\|^2, \tag{5.66}$$

最大化似然估计性质：

当最大似然满足下列性质时，其估计满足一致性。

- 真实分布 p_{data} 必须在模型族 $p_{\text{model}}(\cdot;\boldsymbol{\theta})$ 中。

- 真实分布 p_{data} 必须刚好对应一个 $\boldsymbol{\theta}$ 值。否则，最大似然估计恢复出真实分布 p_{data} 后，也不能决定数据生成过程使用哪个 $\boldsymbol{\theta}$。

频率派的观点：

1. 真实的参数是未知的固定值；
2. 估计是样本的函数，随样本发生变化，是随机变量。
3. 估计最优准则：最大似然估计

贝叶斯观点：

1. 真实的参数是未知不确定的，是随机变量；
2. 观察的样本是确定；
3. 估计的最优准则：最大后经估计

贝叶斯观点：

1. 真实的参数是未知不确定的，是随机变量，其分布为先验分布；
2. 观察的样本是确定；
3. 估计的最优准则：最大后经估计。

现在假设我们有一组数据样本 $\{x^{(1)}, \ldots, x^{(m)}\}$。通过贝叶斯规则结合数据似然 $p(x^{(1)}, \ldots, x^{(m)} \mid \boldsymbol{\theta})$ 和先验，我们可以恢复数据对我们关于 $\boldsymbol{\theta}$ 信念的影响：

$$p(\boldsymbol{\theta} \mid x^{(1)}, \ldots, x^{(m)}) = \frac{p(x^{(1)}, \ldots, x^{(m)} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(x^{(1)}, \ldots, x^{(m)})} \tag{5.67}$$

贝叶斯统计与最大似然不同：

1. 最大似然使用参数的点估计，贝叶斯使用参数的概率分布

下一个数据样本 $x^{(m+1)}$ 的预测分布如下：

$$p(x^{(m+1)} \mid x^{(1)}, \ldots, x^{(m)}) = \int p(x^{(m+1)} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid x^{(1)}, \ldots, x^{(m)}) \, d\boldsymbol{\theta}.$$

2. 最大似然方法用均方误差（偏差和方差）评估估计的不确定问题。而贝叶斯方法使用积分处理不确定。

3. 贝叶斯估计中加入了先验分布，先验分布实现了对参数的选择和偏好，因此有正则化作用。

4. 适用场景：数据有限，数据量大时计算高

例子：贝叶斯线性回归

$\boldsymbol{w} \in \mathbb{R}^n$ 参数化：

$$\hat{y} = \boldsymbol{w}^\top \boldsymbol{x}. \tag{5.69}$$

给定一组 m 个训练样本 $(\boldsymbol{X}^{(\text{train})}, \boldsymbol{y}^{(\text{train})})$，我们可以表示整个训练集对 y 的预测：

$$\hat{\boldsymbol{y}}^{(\text{train})} = \boldsymbol{X}^{(\text{train})}\boldsymbol{w}. \tag{5.70}$$

表示为 $\boldsymbol{y}^{(\text{train})}$ 上的高斯条件分布，我们得到

$$p(\boldsymbol{y}^{(\text{train})} \mid \boldsymbol{X}^{(\text{train})}, \boldsymbol{w}) = \mathcal{N}(\boldsymbol{y}^{(\text{train})}; \boldsymbol{X}^{(\text{train})}\boldsymbol{w}, \boldsymbol{I}) \tag{5.71}$$

$$\propto \exp\left(-\frac{1}{2}(\boldsymbol{y}^{(\text{train})} - \boldsymbol{X}^{(\text{train})}\boldsymbol{w})^\top(\boldsymbol{y}^{(\text{train})} - \boldsymbol{X}^{(\text{train})}\boldsymbol{w})\right), \tag{5.72}$$

其中，我们根据标准的MSE公式假设 y 上的高斯方差为 1。在下文中，为减少符号负担，我们将 $(\boldsymbol{X}^{(\text{train})}, \boldsymbol{y}^{(\text{train})})$ 简单表示为 $(\boldsymbol{X}, \boldsymbol{y})$。

例子：贝叶斯线性回归

确定先验分布

$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}; \boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0) \propto \exp\left(-\frac{1}{2}(\boldsymbol{w} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Lambda}_0^{-1} (\boldsymbol{w} - \boldsymbol{\mu}_0)\right),$$

其中，$\boldsymbol{\mu}_0$ 和 $\boldsymbol{\Lambda}_0$ 分别是先验分布的均值向量和协方差矩阵。[1]

计算后验分布：

$$p(\boldsymbol{w} \mid \boldsymbol{X}, \boldsymbol{y}) \propto p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{w}) p(\boldsymbol{w}) \tag{5.74}$$

$$\propto \exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})\right) \exp\left(-\frac{1}{2}(\boldsymbol{w} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Lambda}_0^{-1} (\boldsymbol{w} - \boldsymbol{\mu}_0)\right) \tag{5.75}$$

$$\propto \exp\left(\frac{1}{2}\left(-2\boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{w} + \boldsymbol{w}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w} + \boldsymbol{w}^\top \boldsymbol{\Lambda}_0^{-1} \boldsymbol{w} - 2\boldsymbol{\mu}_0^\top \boldsymbol{\Lambda}_0^{-1} \boldsymbol{w}\right)\right). \tag{5.76}$$

例子：贝叶斯线性回归

现在我们定义 $\boldsymbol{\Lambda}_m = (\boldsymbol{X}^\top \boldsymbol{X} + \boldsymbol{\Lambda}_0^{-1})^{-1}$ 和 $\boldsymbol{\mu}_m = \boldsymbol{\Lambda}_m(\boldsymbol{X}^\top \boldsymbol{y} + \boldsymbol{\Lambda}_0^{-1}\boldsymbol{\mu}_0)$

后验分布可写为：

$$
\begin{aligned}
p(\boldsymbol{w} \mid \boldsymbol{X}, \boldsymbol{y}) &\propto \exp\left(-\frac{1}{2}(\boldsymbol{w} - \boldsymbol{\mu}_m)^\top \boldsymbol{\Lambda}_m^{-1}(\boldsymbol{w} - \boldsymbol{\mu}_m) + \frac{1}{2}\boldsymbol{\mu}_m^\top \boldsymbol{\Lambda}_m^{-1}\boldsymbol{\mu}_m\right) \\
&\propto \exp\left(-\frac{1}{2}(\boldsymbol{w} - \boldsymbol{\mu}_m)^\top \boldsymbol{\Lambda}_m^{-1}(\boldsymbol{w} - \boldsymbol{\mu}_m)\right).
\end{aligned}
$$

最大后验估计

$$\boldsymbol{\theta}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta} \mid \boldsymbol{x}) = \arg\max_{\boldsymbol{\theta}} \log p(\boldsymbol{x} \mid \boldsymbol{\theta}) + \log p(\boldsymbol{\theta}).$$

正如全贝叶斯推断，MAP 贝叶斯推断的优势是能够利用来自先验的信息，这些信息无法从训练数据中获得。该附加信息有助于减少最大后验点估计的方差（相比于 ML 估计）。然而，这个优点的代价是增加了偏差。

监督学习：

1. 逻辑回归：

$Wx+b=\begin{cases}1, & x \succ 0 \\ 0, & x \leq 0\end{cases}$

任务：二分类问题，模型如下：

$$p(y=1 \mid \boldsymbol{x}; \boldsymbol{\theta}) = \sigma(\boldsymbol{\theta}^{\top}\boldsymbol{x}).$$

代价函数：最大似然

优化算法：梯度下降算法

2. 支出向量机

任务：二分类问题，模型如下：

$$Wx+b=\begin{cases}1, & x \succ 0 \\ 0, & x \leq 0\end{cases}$$

代价函数：

$$\min_{w} \quad \frac{1}{2}\|w\|^2$$
$$s.t. \quad y_i(w^T x_i) \geq 1, \quad (i = 1, \cdots, N)$$

优化算法：凸优化算法

监督学习：

核技巧：

$$\boldsymbol{w}^\top \boldsymbol{x} + b = b + \sum_{i=1}^{m} \alpha_i \boldsymbol{x}^\top \boldsymbol{x}^{(i)},$$

替换为核函数：

$$f(\boldsymbol{x}) = b + \sum_i \alpha_i k(\boldsymbol{x}, \boldsymbol{x}^{(i)}).$$

其中

$$k(\boldsymbol{x}, \boldsymbol{x}^{(i)}) = \phi(\boldsymbol{x}) \cdot \phi(\boldsymbol{x}^{(i)})$$

高斯核（径向基函数）：

$$k(\boldsymbol{u}, \boldsymbol{v}) = \mathcal{N}(\boldsymbol{u} - \boldsymbol{v}; \boldsymbol{0}, \sigma^2 I)$$

监督学习：

3. K近邻算法

任务：分类或者回归，非参数方法

没有训练和学习过程，其容量相当高

其错误率可降低为贝叶斯错误率的两倍

缺点：算法慢，需要大量的样本，没有特征学习过程

4. 决策树或ID3算法

任务：分类。模型：分类树

非参数方法。

优化算法：信息增益算法

决策树

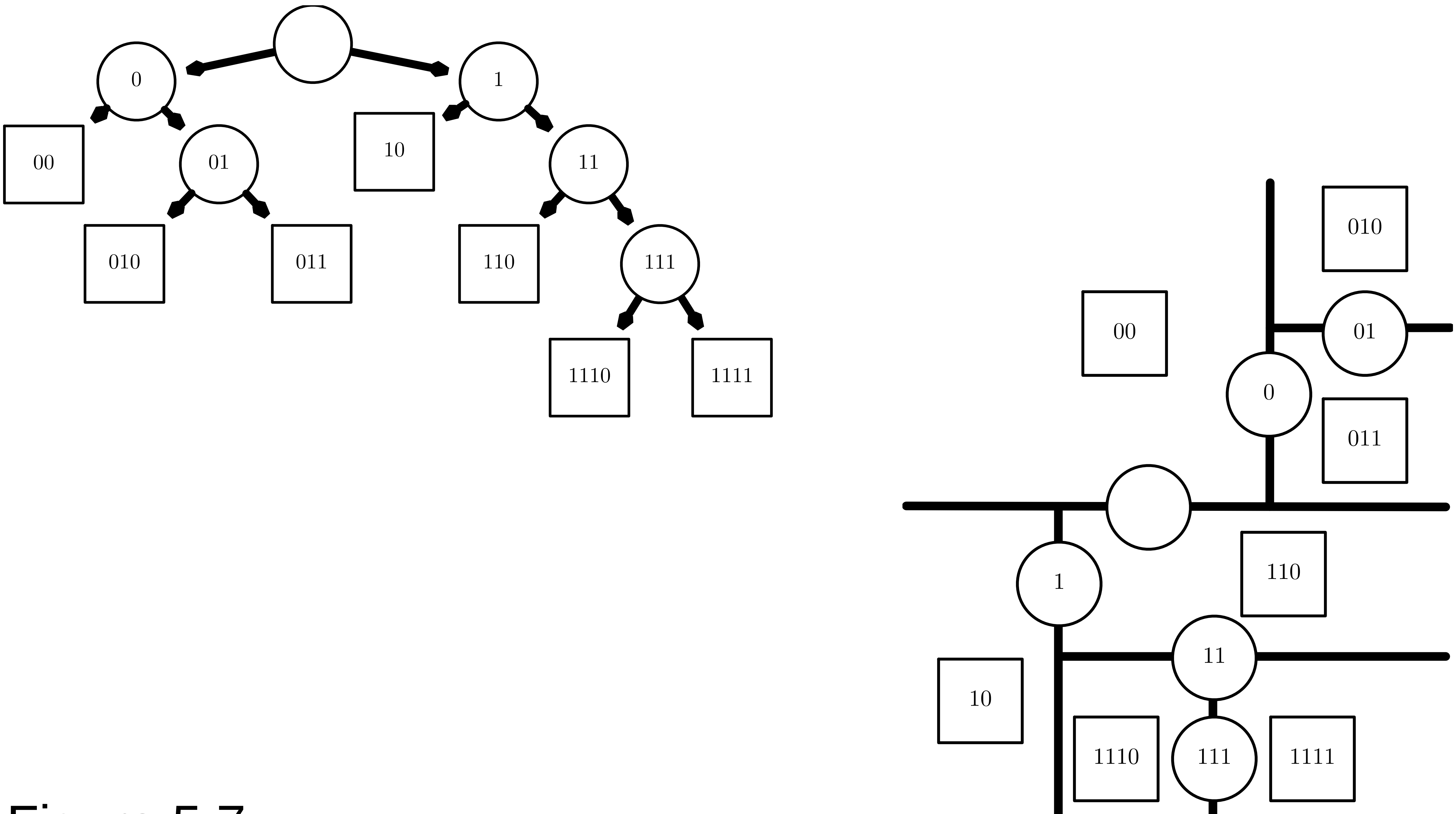


Figure 5.7

(Goodfellow 2016)

无监督学习：没有人工标准的标签

1. 去噪，数据分布，数据聚类
2. 学习数据的最佳表示
3. 什么是最佳表示？ 稀疏表示、压缩表示、独立性表示

（1）主成分分析（PCA变换）

设X是均值为0的数据矩阵

样本协方差矩阵为：

$$\mathrm{Var}[\boldsymbol{x}] = \frac{1}{m-1}\boldsymbol{X}^\top \boldsymbol{X}.$$

通过一个线性变换z=Wx，使得Var[x]为对角矩阵

PCA分析

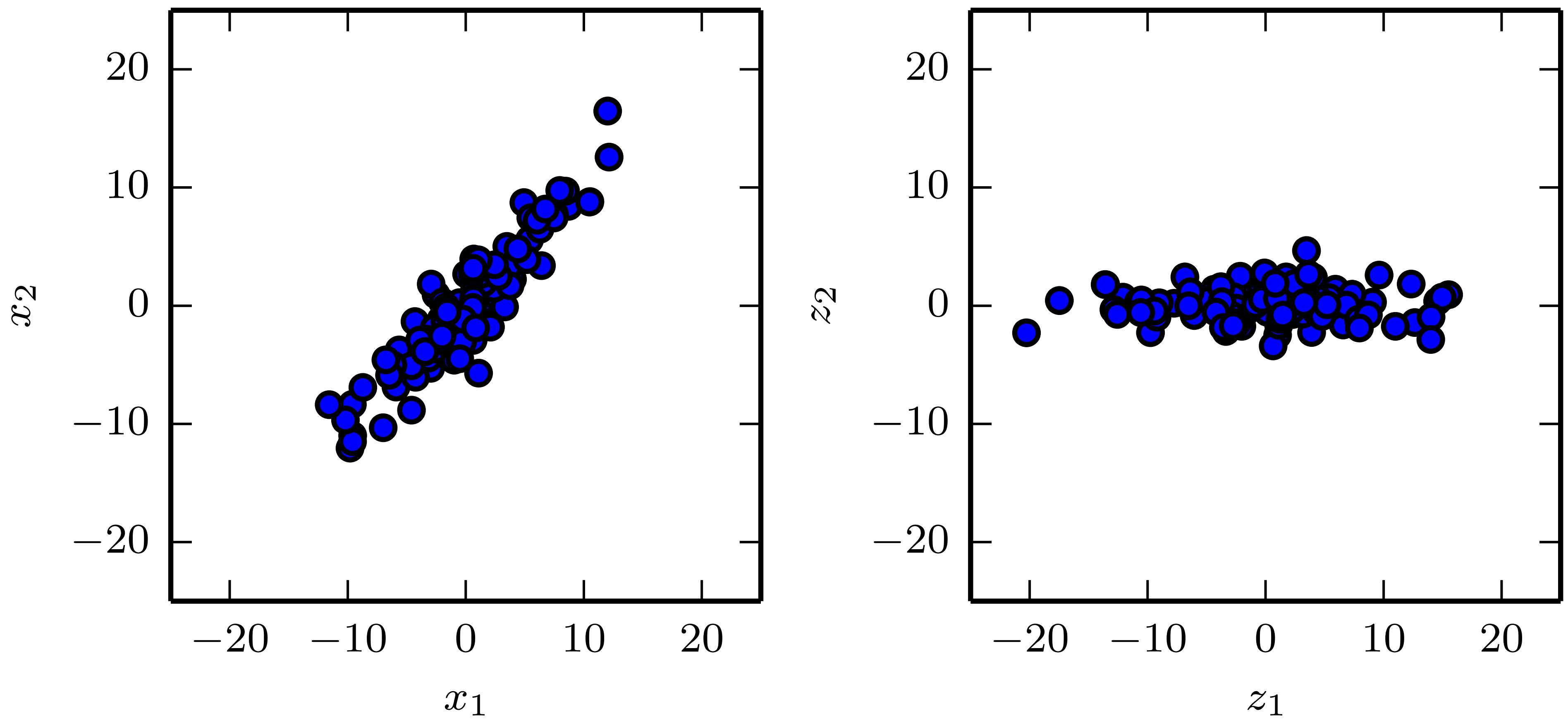


Figure 5.8

(Goodfellow 2016)

PCA变换

1. X的ＳＶＤ分解：

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{W}^\top$$

则：

$$\boldsymbol{X}^\top\boldsymbol{X} = \left(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{W}^\top\right)^\top\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{W}^\top = \boldsymbol{W}\boldsymbol{\Sigma}^2\boldsymbol{W}^\top$$

因此，

$$\begin{aligned}
\mathrm{Var}[\boldsymbol{x}] &= \frac{1}{m-1}\boldsymbol{X}^\top\boldsymbol{X} \\
&= \frac{1}{m-1}\left(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{W}^\top\right)^\top\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{W}^\top \\
&= \frac{1}{m-1}\boldsymbol{W}\boldsymbol{\Sigma}^\top\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{W}^\top \\
&= \frac{1}{m-1}\boldsymbol{W}\boldsymbol{\Sigma}^2\boldsymbol{W}^\top,
\end{aligned}$$

PCA变换

1. X的ＳＶＤ分解：

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{W}^\top$$

则：

$$\boldsymbol{X}^\top\boldsymbol{X} = \left(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{W}^\top\right)^\top\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{W}^\top = \boldsymbol{W}\boldsymbol{\Sigma}^2\boldsymbol{W}^\top$$

因此，

$$\begin{aligned}
\mathrm{Var}[\boldsymbol{x}] &= \frac{1}{m-1}\boldsymbol{X}^\top\boldsymbol{X} \\
&= \frac{1}{m-1}\left(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{W}^\top\right)^\top\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{W}^\top \\
&= \frac{1}{m-1}\boldsymbol{W}\boldsymbol{\Sigma}^\top\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{W}^\top \\
&= \frac{1}{m-1}\boldsymbol{W}\boldsymbol{\Sigma}^2\boldsymbol{W}^\top,
\end{aligned}$$

PCA变换

令 z ＝W x， 则

$$
\begin{aligned}
\mathrm{Var}[z] &= \frac{1}{m-1}\boldsymbol{Z}^{\top}\boldsymbol{Z} \\
&= \frac{1}{m-1}\boldsymbol{W}^{\top}\boldsymbol{X}^{\top}\boldsymbol{X}^{\top}\boldsymbol{W} \\
&= \frac{1}{m-1}\boldsymbol{W}^{\top}\boldsymbol{W}\boldsymbol{\Sigma}^{2}\boldsymbol{W}^{\top}\boldsymbol{W} \\
&= \frac{1}{m-1}\boldsymbol{\Sigma}^{2},
\end{aligned}
$$

k－mean聚类

1.数据的one－hot表示

2.与任务无关，无法描述真实的情况

随机梯度下降算法（ＳＧＤ）

１．问题：当样本大时，计算代价函数的梯度低效耗时

$$J(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x},\mathrm{y}\sim\hat{p}_{\mathrm{data}}} L(\boldsymbol{x}, y, \boldsymbol{\theta}) = \frac{1}{m}\sum_{i=1}^{m} L(\boldsymbol{x}^{(i)}, y^{(i)}, \boldsymbol{\theta}),$$

计算梯度：

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \frac{1}{m}\sum_{i=1}^{m} \nabla_{\boldsymbol{\theta}} L(\boldsymbol{x}^{(i)}, y^{(i)}, \boldsymbol{\theta})$$

２．方法：从m个样本中随机选择一定固定的小样本集

$$\boldsymbol{g} = \frac{1}{m'} \nabla_{\boldsymbol{\theta}} \sum_{i=1}^{m'} L(\boldsymbol{x}^{(i)}, y^{(i)}, \boldsymbol{\theta})$$

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \epsilon \boldsymbol{g}.$$

维度灾难：

1．传统的学习方法都是局部学习方法，利用局部样本和局部平滑性假设，进行学习和训练。

2．学习到的局部区域不能超过样本数。

3．高维空间数据不足

4．能学习到超过样本的局部区域吗？

维度灾难

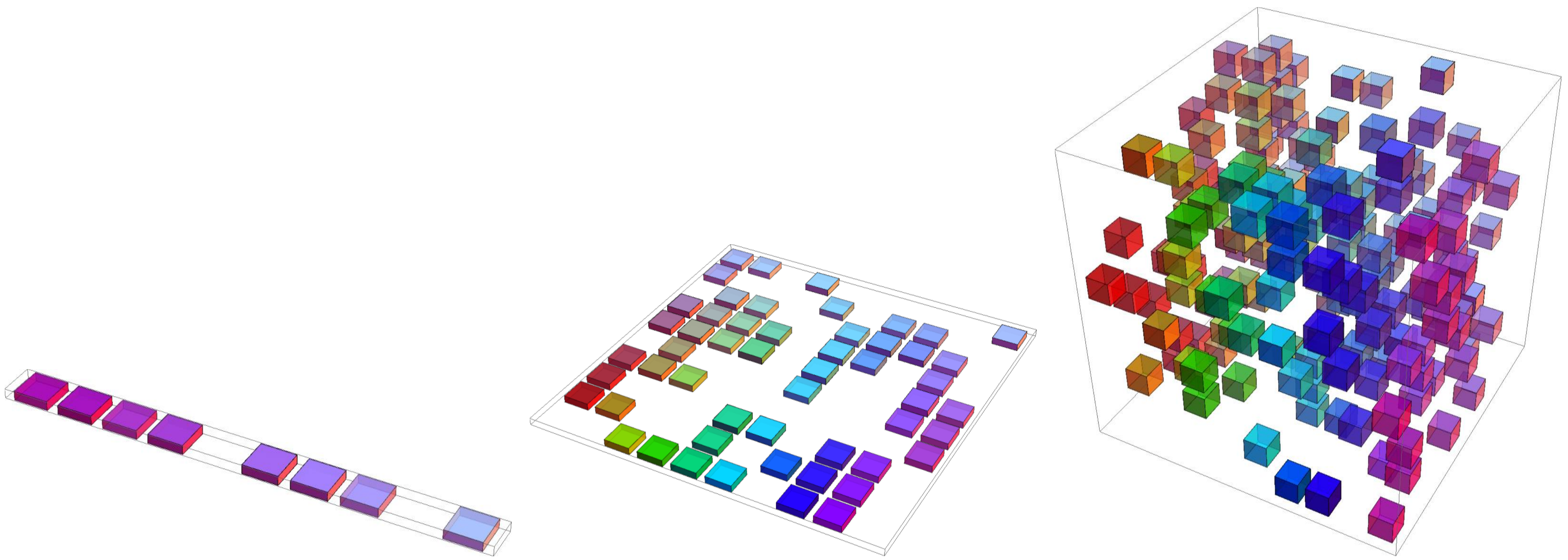


Figure 5.9

(Goodfellow 2016)

Manifold Learning

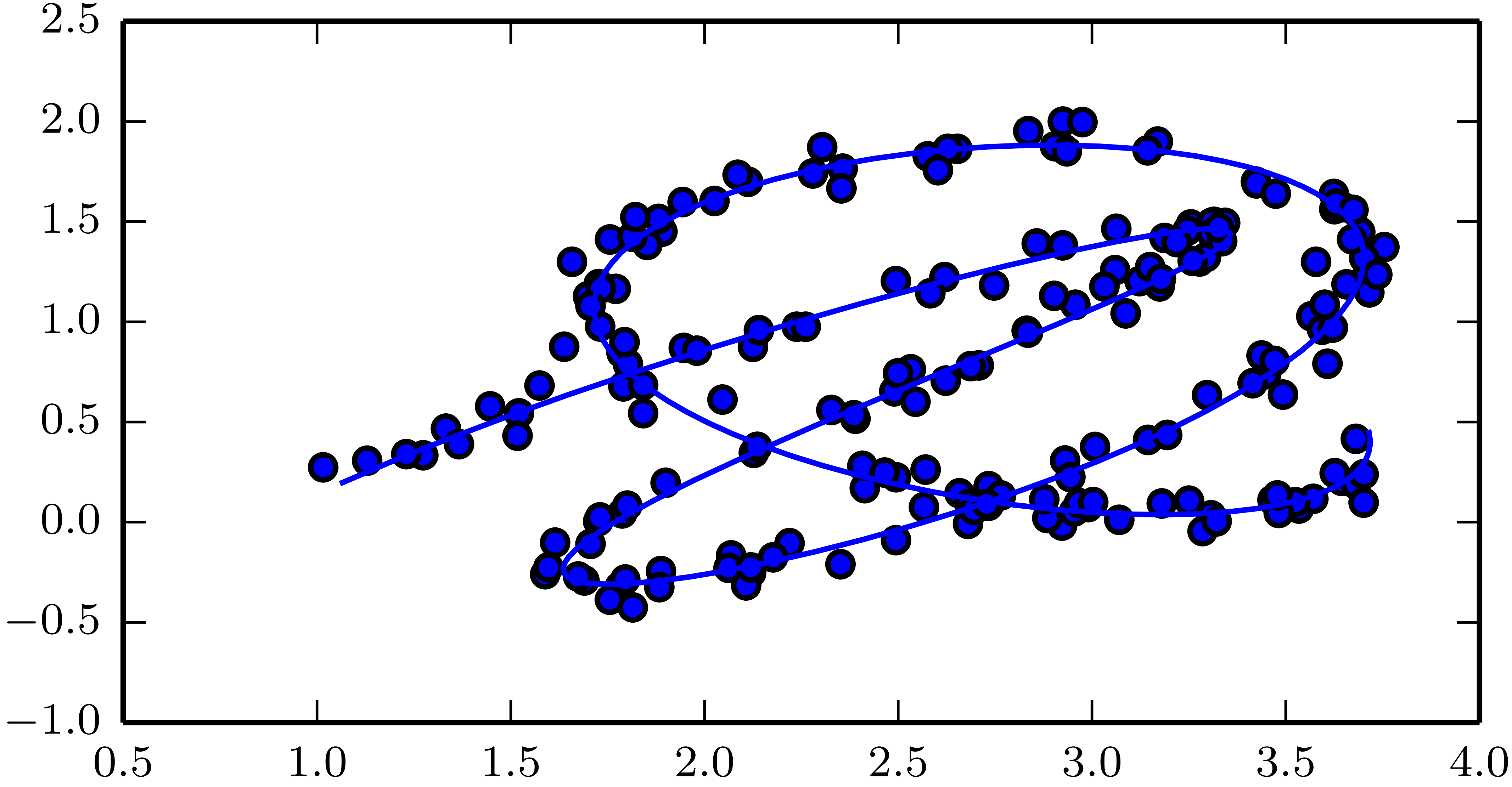


Figure 5.11

(Goodfellow 2016)

流形学习：

1. 输入空间不是整体有效，输入位于一个低维的流形上面。

2. 通过学习输入的低维的流形

均匀采样的图像

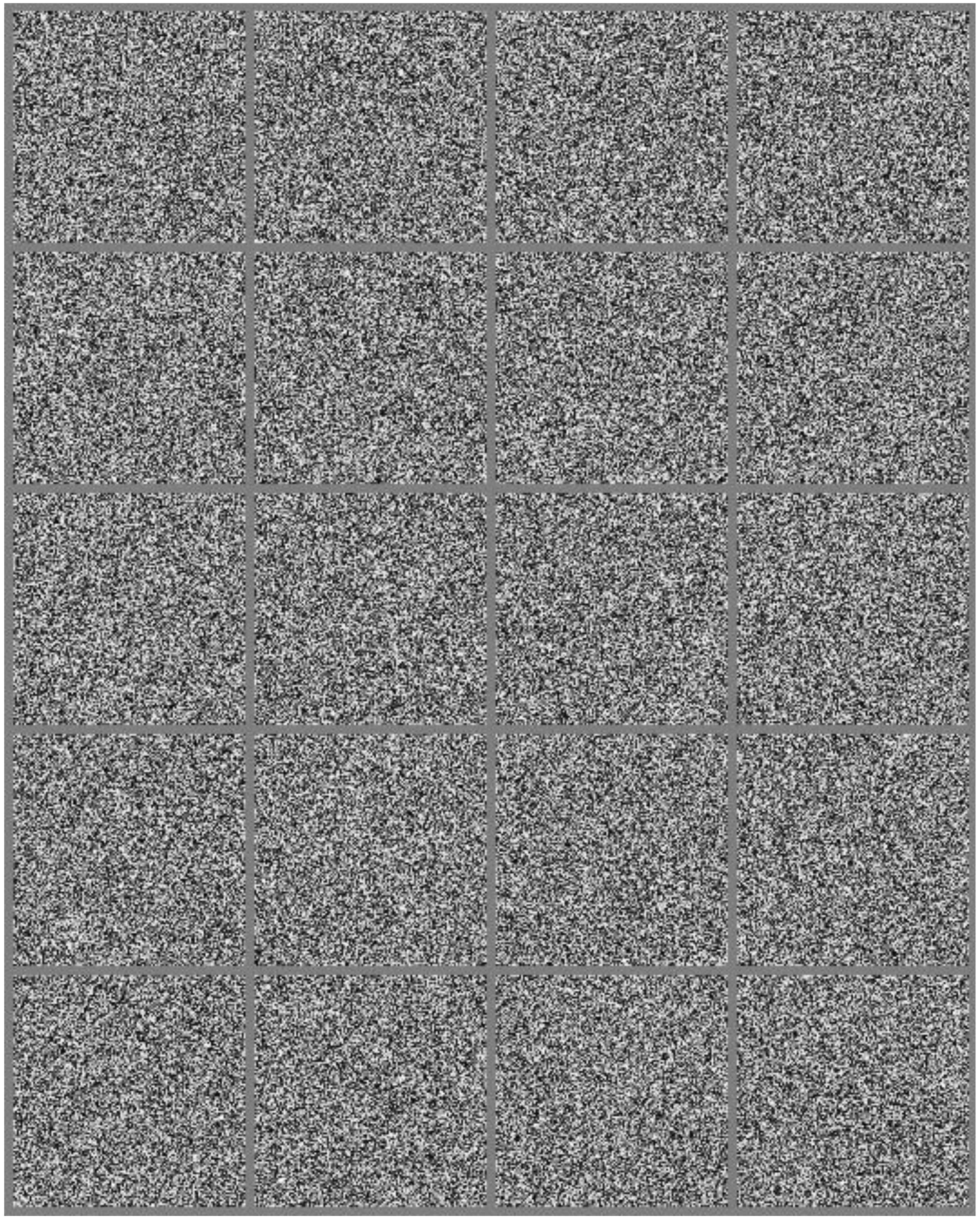


Figure 5.12

(Goodfellow 2016)

数据在流形证明：QMUL Dataset

Figure 5.13

作业

以线性回归为例子，设计不同的多项式模型，评估估计均值和估计方差，由此分析模型的复杂性、过拟合、欠拟合等与估计的均值和估计方差的关系。

(Goodfellow 2016)