# 第五章 机器学习基础

# 5.1 学习算法

- **学习的定义：对于某类任务 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 衡量的性能有所提升。**
- **5.1.1 任务 T**
- 通常**机器学习任务**定义为机器学习系统应该如何处理**样本**（example）。
- **样本**是指我们从某些希望机器学习系统处理的对象或事件中收集到的已经量化的**特征（feature）**的集合。
- 我们通常会将样本表示成一个**向量 x∈Rn**，其中向量的每一个元素 **xi** 是一个特征。
- 例如，一张图片的特征通常是指这张图片的像素值。

# 5.1 学习算法

- **常见的机器学习任务**
- **分类：**根据某些输入指定输出结果属于k类中的哪一类，如：
  - y = f(x)
- 分类模型将向量 x 所代表的输入生成结果y ,y可能是以下几种情况：

➢ 类别，所属，比如手写数字的识别

➢ 概率，比如通过sigmoid函数处理得到的（0，1）之间地 取值

# 5.1 学习算法

- **常见的机器学习任务**
- **缺失输入分类：**对于输入向量x，当一些输入可能丢失时，学习算法必须学习一组函数，而不是 单个分类函数。每个函数对应着分类具有不同缺失输入子集的 x。
- **回归：**计算机程序需要对给定输入预测数值，如房价预测
- **转录：**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。
- **机器翻译：**将一种语言翻译成另一种语言。如自然语言处理中将英语翻译成法语

# 5.1 学习算法

- **常见的机器学习任务**
- **结构化输出：**输出是向量或者其他包含多个值的数据结构， 并且构成输出的这些不同元素间具有重要关系。如计算机程序观察到一幅图，输出描述这幅图的自然语言句子。
- **异常检测**：计算机程序在一组事件或对象中筛选，并标记不正常或非典型的个体。如信用卡欺诈检测。通过对你的购买习惯建模，信用卡公司可以检测到你的卡是否被滥用。
- **合成和采样：**机器学习程序生成一些和训练数据相似的新样本。如在媒体应用中，视频游戏可以自动生成大型物体或风景的纹理，而不是让艺术家手动标记每个像素。

# 5.1 学习算法

- **常见的机器学习任务**
- **缺失值填补：**机器学习算法给定一个新样本 $x \in \mathbb{R}^n$，x 中某些 元素 xi 缺失。算法必须填补这些缺失值。
- **去噪：**干净样本 $x \in \mathbb{R}^n$ 经过未知损 坏过程后得到的损坏样本 $\tilde{x} \in \mathbb{R}^n$。算法根据损坏后的样本 $\tilde{x}$ 预测干净的样本 x，或者更一般地预测条件概率分布 $p(x|\tilde{x})$。
- **密度估计或概率质量函数估计：**算法需要学习观测到的数据的结构，知道什么情况下样本聚集出现，什么情况下不太可能出现，显式地捕获该分布并得到概率分布 p(x)，反过来可以用于缺失值的填补问题 。

# 5.1 学习算法

- **5.1.2 性能度量P**
- 通常性能度 量 P 是特定于系统执行的任务 T 而言的
- **对于诸如分类、缺失输入分类和转录任务：**通常度量模型的准确率或者错误率。
- **对于密度估计这类任务：**度量准确率，错误率是没有意义的。必须使用不同的性能度量，使模型对每 个样本都输出一个连续数值的得分，最常用的方法是输出模型在一些样本上概率对数的平均值。
- **结论：**性能度量的选择或许看上去简单且客观，但是选择一个与系统理想表现对应 的性能度量通常是很难的。

# 5.1 学习算法

- **5.1.3 经验E**
- 根据学习过程中的不同经验，机器学习算法可以大致分类为无监督（unsupervised）算法和监督（supervised）算法。
- 监督学习：训练含有许多特征的数据集，每个数据样本有对应的标签，如
- 无监督学习：训练含有许多特征的数据集，但是没有标签，需要自己学习出数据中有用的结构性质
- 传统地，人们将回归、分类或者结构化输出问题称为监督学习。支持其他任务的密度估计通常被称为无监督学习

# 5.1 学习算法

- **5.1.3 经验E**
- 监督学习与无监督学习的区别与联系
- 无监督学习涉及到观察随机向量 x 的好几个样本，试图显式或隐式 地学习出概率分布 p(x)，或者是该分布一些有意思的性质；而监督学习包含观察随 机向量 x 及其相关联的值或向量 y，然后从 x 预测 y，通常是估计 p(y|x)
- 无监督学习和监督学习不是严格定义的术语。它们之间界线通常是模糊的。很 多机器学习技术可以用于这两个任务。例如，概率的链式法则表明对于向量 x∈Rn， 联合分布可以分解成

$$p(\mathbf{x}) = \prod_{i=1}^{n} p(\mathrm{x}_i \mid \mathrm{x}_1, \ldots, \mathrm{x}_{i-1}).$$

# 5.1 学习算法

- **5.1.3 经验E**
- 监督学习与无监督学习的区别与联系
- 该分解意味着我们可以将其拆分成 n 个监督学习问题，来解决表面上的无监督学习 p(x)。另外，我们求解监督学习问题 p(y |x) 时，也可以使用传统的无监督学习策略 学习联合分布 p(x,y)，然后推断

$$p(y \mid \mathbf{x}) = \frac{p(\mathbf{x}, y)}{\sum_{y'} p(\mathbf{x}, y')}.$$

- 半监督学习： 一些样本有监督目标（数据标签），但其他样本没有。

# 5.1 学习算法

- **5.1.4 示例：线性回归**
- 定义任务 T：通过输出 $\hat{y} = w^\top x$ 从 x 预测 y。
- 定义性能度量P：均方误差

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

$$\|\boldsymbol{x}\|_p = \left(\sum_i |x_i|^p\right)^{\frac{1}{p}} \qquad \mathrm{MSE}_{\mathrm{test}} = \frac{1}{m}\left\|\hat{\boldsymbol{y}}^{(\mathrm{test})} - \boldsymbol{y}^{(\mathrm{test})}\right\|_2^2,$$

# 5.1 学习算法

- **5.1.4 示例：线性回归**

$$\nabla_w \mathrm{MSE}_{\mathrm{train}} = 0$$

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

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

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

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

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

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

# 5.1 学习算法

- **5.1.4 示例：线性回归**
- 截距项b:
- 术语线性回归（linear regression）通常用来指稍微复杂一些， 附加额外参数（截距项 b）的模型。在这个模型中，

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

- 但是通常在实际过程中，将b添加到向量w中，而x添加一项为1的元素

# 5.2 容量、过拟合和欠拟合

- **泛化：**在先前未观测到的输入上表现良好的能力
- **训练误差：**训练集上计算数据与预测目标值的误差
- **泛化误差（测试误差）：**在测试集计算数据与预测目标值的误差
- 在实际过程中，一般测试误差会大于训练误差，所以决定机器学习算法效果是否好的因素

➢ 降低训练误差

➢ 缩小训练误差和测试误差的差距

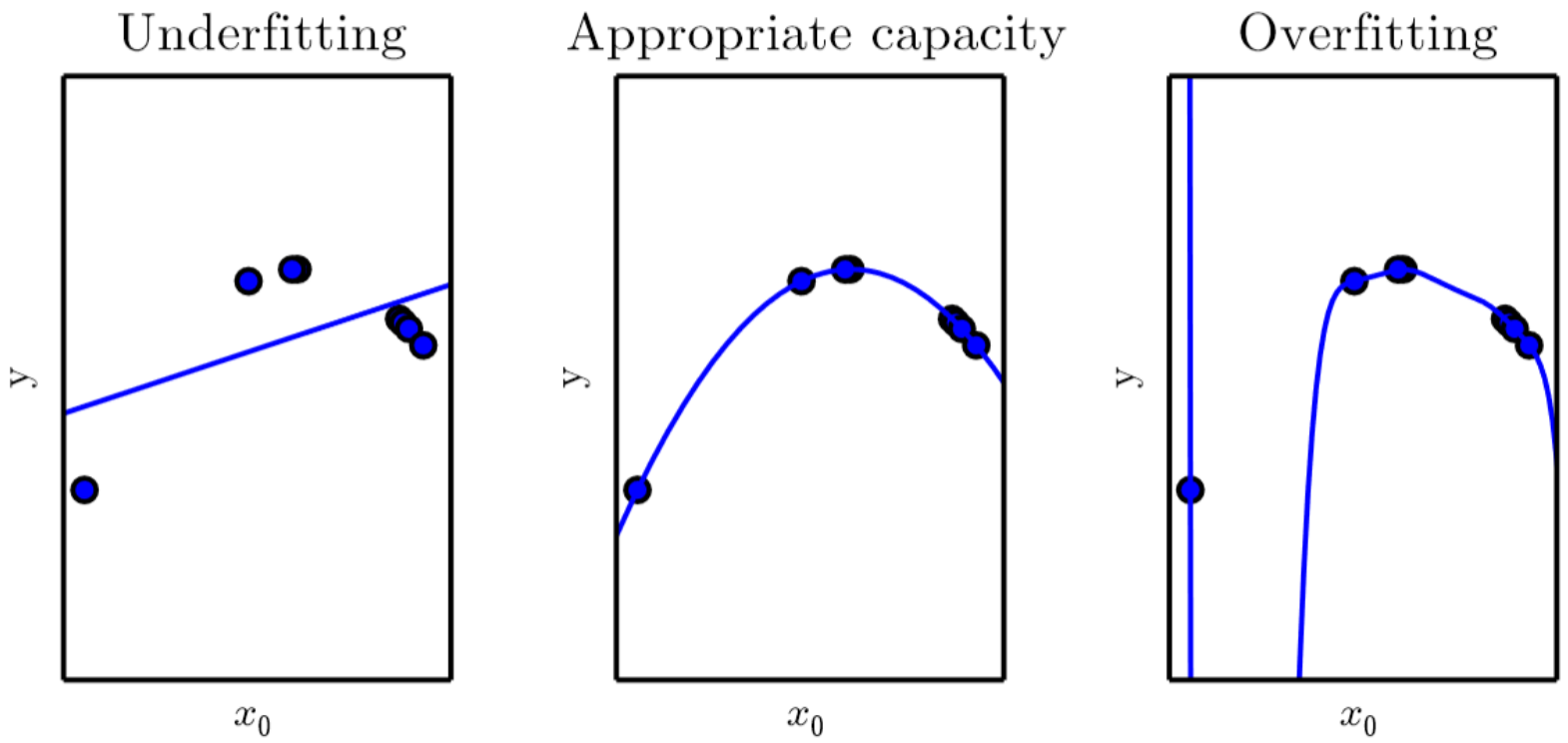
- **挑战：欠拟合和过拟合**
- 欠拟合是指模型不能在训练集上获得足够低的误差。而过拟合是指训 练误差和和测试误差之间的差距太大。

# 5.2 容量、过拟合和欠拟合

- **解决方式：**通过调整模型的**容量**（capacity），可以控制模型是否偏向于过拟合或者欠 拟合。
- 模型的**容量**是指其拟合各种函数的能力。容量低的模型可能很难拟合训练集。容量高的模型可能会过拟合，因为记住了不适用于测试集的训练集性质。
- 具体方法：选择假设空间，即学习算法选择为解决方案的函数集
- 例子：线性回归算法中，广义线性回归的假设空间包括多项式函数，而非仅有线性函数。
- 定义三个不同的模型：一次多项式、二次多项式、九次多项式

$$\hat{y} = b + wx. \qquad \hat{y} = b + w_1 x + w_2 x^2. \qquad \hat{y} = b + \sum_{i=1}^{9} w_i x^i.$$

# 5.2 容量、过拟合和欠拟合



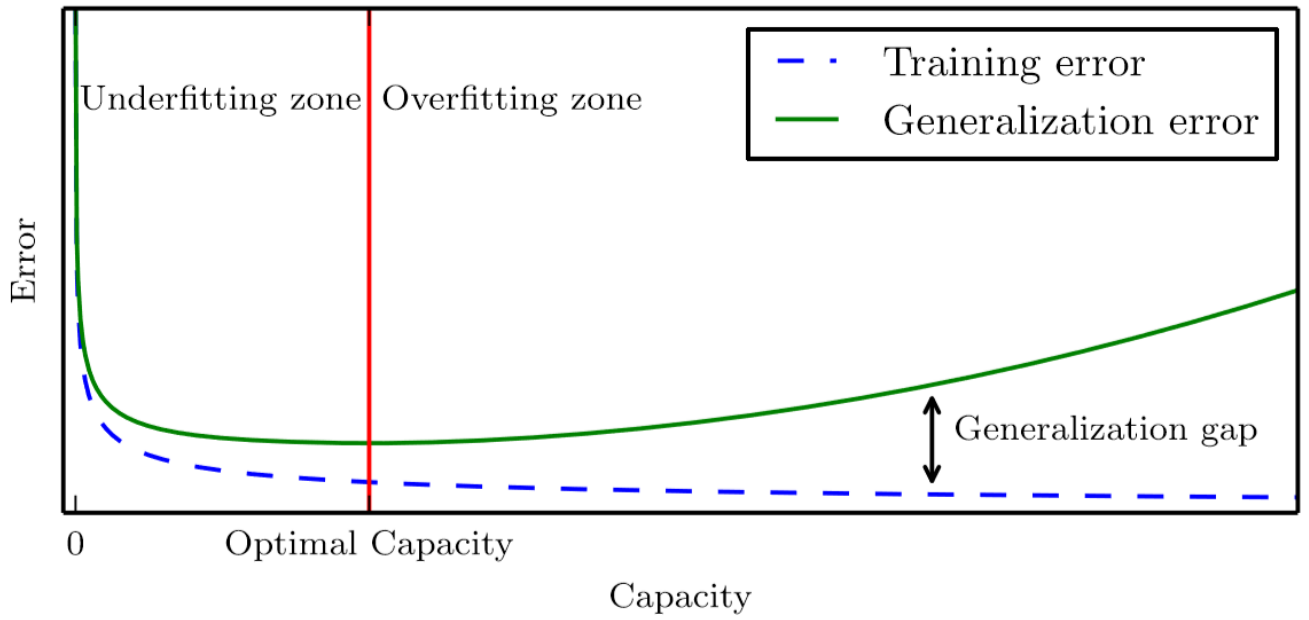

- 在这个问题中，一次多项式欠拟合，九次多项式过拟合，二次模型非常符合任务的真实结构，因此它可以很好地泛化到新数据上。

# 5.2 容量、过拟合和欠拟合

- **表示容量与有效容量**
- 表示容量：在多种函数中选择的最优函数所具有的容量。
- 在实际过程中，很多机器学习案例不像线性回归那么简单，函数也更加复杂，寻找一个最佳的优化函数是很困难的，在找不到最优函数的情况下，可以找一个降低训练误差的函数
- 有效容量：实际过程中找到的降低训练误差的函数所具有的容量，如果该函数效果不佳，则有效容量小于表示容量
- **容量的量化**
- 最经典的方法--VC维：度量二元分类器的容量，可以简单地理解成：当存在m个x的训练样本，正确预测的个数k，k的最大值就是该分类器的容量

# 5.2 容量、过拟合和欠拟合

- 量化模型的容量使得统计学习理论可以进行量化预测。统计学习理论中最重要的结论阐述了**训练误差和泛化误差之间差异的上界随着模型容量增长而增长，但随着训练样本增多而下降 。**
- 训练误差和泛化误差关于模型容量的函数



# 5.2 容量、过拟合和欠拟合

- **非参数模型**
- 概念：相较于线性回归中有类似w和b等固定参数的情况，非参数模型没有参数的严格限制
- 非参数模型有时候可以是一个不可能实现的理论抽象，但是也可以设计一些实际的非参数模型，如最近邻回归。
- 最近邻回归：预先存储所有的X和y，当为测试点x分类时，模型会查询训练集中离该点最近的点，并返回相关的回归目标。

$$\hat{y} = y_i \text{ 其中 } i = \arg\min \|\boldsymbol{X}_{i,:} - \boldsymbol{x}\|_2^2$$

# 5.2 容量、过拟合和欠拟合

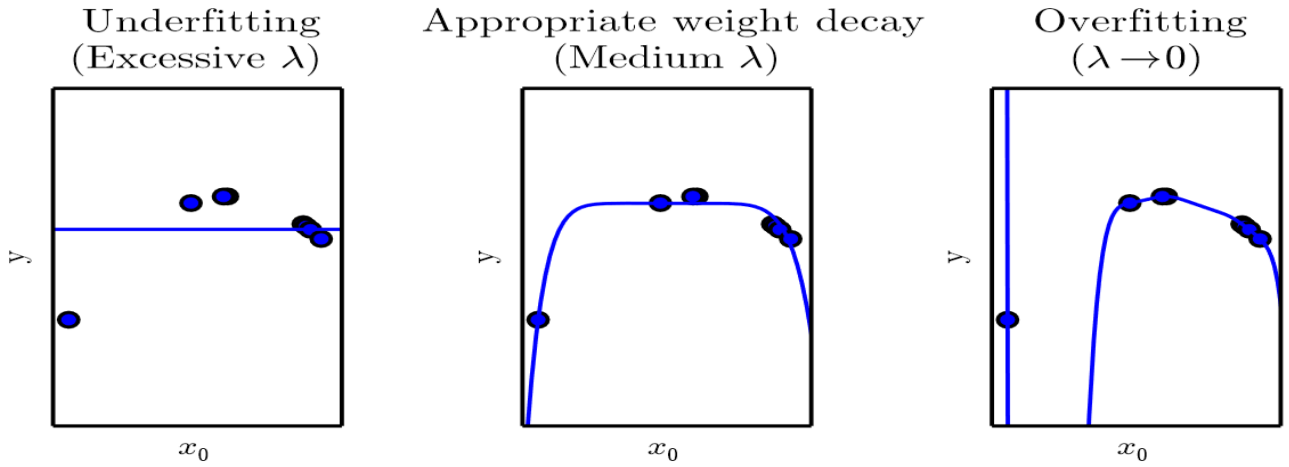
- **5.2.1 没有免费午餐定理**
- 传统的逻辑推导：逻辑地推断一个规则去描述集合中的元素，我们必须具有集合中每个元素的信息
- 机器学习推导：从一组有限的样本中推断一般的规则，似乎违背传统的逻辑原则。但是机器学习还是能够在大多数样本上找到可能正确的规则。
- 没有免费午餐定理：即使机器学习能够找到相关规则，但是没有一个机器学习算法总是比其他的要好，即在当前情况下效果最优的机器学习模型，在另一种情况下效果并不是最优。
- 结论：机器学习研究的目标不是找一个通用学习算法或是绝对最好的学习算法。而是寻找一个相关的学习算法，在我们关注的数据生成分布上效果最好

# 5.2 容量、过拟合和欠拟合

- **5.2.2 正则化**
- 正则化例子：可以加入权重衰减（weight decay）来修改线性回归的训练标准。 wTw称为正则化项：

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

- 其中 λ 是提前挑选的值，控制我们偏好小范数权重的程度。当 λ = 0，我们没有任何偏好。越大的 λ 偏好范数越小的权重。（九次多项式正则化例子）



# 5.3 超参数和验证集

- **超参数：**不是算法本身学习出来的参数。如上述正则化线性回归中的参数λ
- **验证集：**从训练集中分离出来的，用于挑选超参数的数据子集：验证不同超参数，对比不同的效果，挑选最好的模型。
- **5.3.1 交叉验证**
- 原因：单从一个小规模的数据集上进行误差测试存在随机性和不确定性
- k-折交叉验证：将数据集分成 k 个 不重合的子集。测试误差可以估计为 k 次计算后的平均测试误差。在第 i 次测试时， 数据的第 i 个子集用于测试集，其他的数据用于训练集。

# 5.4 估计、偏差和方差

- **5.4.1 点估计**
- **定义：**为一些感兴趣的量提供单个 “最优” 预测（最优的单个显示值）
- **函数估计**：点估计的一种，指输入和目标变量之间**关系**的估计（最优的关系表达式）
- **5.4.2 偏差**
- 估计的偏差被定义为：

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

- 结果为0：无偏
- 结果在无穷处趋近于零：渐进无偏

# 5.4 估计、偏差和方差

- **伯努利分布的样本均值估计**
- 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),…,x(m)}:

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

- 判断以下估计是有偏还是无偏:

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

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

# 5.4 估计、偏差和方差

- **高斯分布的样本均值估计**
- 考虑一组独立同分布的样本 {x(1),...,x(m)} 服

从高斯分布 p(x(i)) = N(x(i);μ,σ2)，其中 i ∈

{1,...,m}。

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

- 判断以下估计是有偏还是无偏:

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

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

# 5.4 估计、偏差和方差

- **高斯分布的样本方差估计**
- 考虑样本方差:

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

- 判断以下时有偏还是无偏:

$$\mathrm{bias}(\hat{\sigma}_m^2) = \mathbb{E}[\hat{\sigma}_m^2] - \sigma^2.$$

$$E(S^2) = E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = \frac{1}{n}E\left[\sum_{i=1}^{n}(X_i^2 - 2X_i\bar{X} + \bar{X}^2)\right]$$

$$= \frac{1}{n}E[\sum_{i=1}^{n}(X_i^2)] - \frac{1}{n}E(2\bar{X} \times n\bar{X} - n\bar{X}^2) = \frac{1}{n}\sum_{i=1}^{n}E(X_i^2) - E(\bar{X}^2)$$

$$E(X_i^2) = D(X_i) + E^2(X_i) = \sigma^2 + \mu^2$$

$$E(\bar{X}^2) = D(\bar{X}) + E^2(\bar{X}) = \frac{\sigma^2}{n} + \mu^2$$

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

$$= \frac{m-1}{m}\sigma^2$$

# 5.4 估计、偏差和方差

- **高斯分布的无偏样本方差估计**
- 考虑无偏样本方差:

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

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

- 判断以下时有偏还是无偏:

$$\text{bias}(\hat{\sigma}_m^2) = \mathbb{E}[\hat{\sigma}_m^2] - \sigma^2.$$

# 5.4 估计、偏差和方差

- **5.4.3 方差和标准差**
- **表示方式：** $\mathrm{Var}(\hat{\theta})$ $\mathrm{SE}(\hat{\theta})$
- **均值的标准差被记作**

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

- 例子：伯努利分布

$\hat{\theta}_m$ = $\frac{1}{m}\sum_{i=1}^{m} x^{(i)}$ 的方差

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

# 5.4 估计、偏差和方差

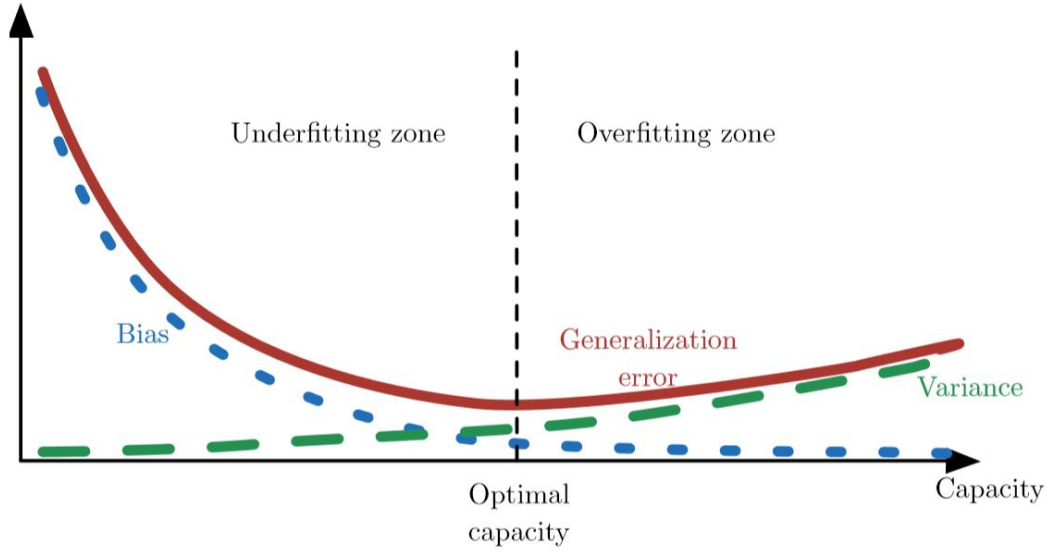
- **5.4.4 权衡偏差和方差以最小化均方误差**
- 偏差和方差度量着估计量的两个不同误差来源。偏差度量着偏离真实函数或参数的**误差期望**。而方差度量着数据上任意特定采样可能导致的估计**期望的偏差**。
- 判断这种权衡最常用的方法是交叉验证。另外，我们也可以比较这些估计的均方误差：

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

$$MSE(\hat{\theta}) = E\left\{\left[\hat{\theta} - E(\hat{\theta})\right] + \left[E(\hat{\theta}) - \theta\right]\right\}^2$$

$$= E\left[\hat{\theta} - E(\hat{\theta})\right]^2 + \left[E(\hat{\theta} - \theta)\right]^2 + 2E\left\{\left[\hat{\theta} - E(\hat{\theta})\right]\left[E(\hat{\theta}) - \theta\right]\right\}$$

$$= D(\hat{\theta}) + \left[E(\hat{\theta}) - \theta\right]^2.$$



# 5.4 估计、偏差和方差

- **5.4.5 一致性**
- 训练数据增多后估计量的效果，我们希望点估计会收敛到对应参数的真实值

$$\text{plim}_{m\to\infty}\hat{\theta}_m = \theta.$$

- 符号 plim 表示依概率收敛，即对于任意的 $\epsilon > 0$，当 $m \to \infty$ 时，有 $P(|\hat{\theta}_m - \theta| > \epsilon) \to 0$ 。上式表示的条件被称为一致性（consistency）。强一致性是指几乎必然（almost sure）从 $\hat{\theta}$收敛到 θ。

# 5.5 最大似然估计

- **目标**
- 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。
- 最大似然估计:
- 考虑一组含有 m 个样本的数据集 X = {x(1),…,x(m)}，独立地由未知的真实数 据生成分布 $p_{data}(x)$ 生成。
- $p_{model}(x;\theta)$ 是一族由 θ 确定在相同空间上的概率分布。换言之， $p_{model}(x;\theta)$ 将任意输入 x 映射到实数来估计真实概率$p_{data}(x)$

# 5.5 最大似然估计

- 对 θ 的最大似然估计被定义为:

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

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

- 为了得到一个便于计算的等价优化问题，我们观察到似然对数不会改变其 argmax 但是将乘积转化成了便于计算的求和形式:

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

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

# 5.5 最大似然估计

- 一种解释最大似然估计的观点是将它看作最小化训练集上的经验分布 ˆ pdata 和模型分布之间的差异，两者之间的差异程度可以通过 KL 散度度量。KL 散度被定义为：

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

- 左边一项仅涉及到数据生成过程，和模型无关。这意味着当我们训练模型最小化 KL 散度时，我们只需要最小化

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

# 5.5 最大似然估计

- **5.5.1 条件对数似然和均方误差**
- 最大似然估计很容易扩展到估计条件概率 P(y|x;θ)，从而给定 x 预测 y。如果 X 表示所有的输入，Y 表示我们观测到的目标，那么条件最大似然估计是

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

- 如果假设样本是独立同分布的，那么这可以分解成

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

# 5.5 最大似然估计

- 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 |x)，而不 只是得到一个单独的预测 ˆ y。现在学习算法的目标是拟合分布 p(y |x) 到和 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 $p(y \mid \boldsymbol{x}) = \mathcal{N}(y; \hat{y}(\boldsymbol{x}; \boldsymbol{w}), \sigma^2)$
- 假设样本是独立同分布的，条件对数似然如下

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

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

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

# 5.5 最大似然估计

- **5.5.2 最大似然的性质**
- 在合适的条件下，最大似然估计具有一致性，训练样本数目趋向于无穷大时，参数的最大似然估计会收敛到参数的真实值，这些条件是：

➢ 真实分布 pdata 必须在模型族 pmodel(·;θ) 中。否则，没有估计可以还原 pdata

➢ 真实分布 pdata 必须刚好对应一个 θ 值。否则，最大似然估计恢复出真实分布 pdata 后，也不能决定数据生成过程使用哪个 θ。

# 5.6 贝叶斯统计

- 一般来讲，我们经常采用基于估计单一 值 θ 的方法，然后基于该估计作所有的预测。
- 区别于上述方法，在做预测时会考虑所有可能的 θ，这属于贝叶斯统计的范畴。
- 先验概率分布：在观测数据之前，将 θ 的已知知识表示成先验概率分布（已知的，可能的取值）
- 现在假设我们有一组数据样本 {x(1),...,x(m)}。通过贝叶斯规则结合数据似然 p(x(1),...,x(m) | θ) 和先验，我们可以恢复数据对我们关于 θ 信念的影响：

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

# 5.6 贝叶斯统计

- **相较于最大似然估计的两个区别**
- 不像最大似然方法预 测时使用 θ 的点估计，贝叶斯方法使用 θ 的全分布。例如，在观测到 m 个样本后， 下一个数据样本 x(m+1) 的预测分布如下：

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

- 这里，每个具有正概率密度的 θ 的值有助于下一个样本的预测，其中贡献由后验密度本身加权。在观测到数据集 {x(1),...,x(m)} 之后，如果我们仍然非常不确定 θ 的值，那么这个不确定性会直接包含在我们所做的任何预测中。
- 贝叶斯方法和最大似然方法的第二个最大区别是由贝叶斯先验分布造成的。先验能够影响概率质量密度朝参数空间中偏好先验的区域偏移。实践中，先验通常表现为偏好更简单或更光滑的模型。

# 5.6 贝叶斯统计

- **例子：贝叶斯线性回归**
- 给定一组 m 个训练样本 (X(train),y(train))，我们可以表示整个训练集对 y 的预测：

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

- 表示为 y(train) 上的高斯条件分布，我们得到

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

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

- 其中，定义方差为1

# 5.6 贝叶斯统计

- 为确定模型参数向量 w 的后验分布，我们首先需要指定一个先验分布。在实践中我们通常假设一个相当广泛的分布来表示 θ 的高度不 确定性。实数值参数通常使用高斯作为先验分布：

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

- 其中，μ0 和 Λ0 分别是先验分布的均值向量和协方差矩阵。确定好先验后，我们现在可以继续确定模型参数的后验分布。

$$\begin{aligned} p(\boldsymbol{w} \mid \boldsymbol{X}, \boldsymbol{y}) &\propto p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{w}) p(\boldsymbol{w}) \\ &\propto \exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w})^\top(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w})\right) \exp\left(-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{\mu}_0)^\top \boldsymbol{\Lambda}_0^{-1}(\boldsymbol{w}-\boldsymbol{\mu}_0)\right) \\ &\propto \exp\left(-\frac{1}{2}\left(-2\boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{w} + \boldsymbol{w}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w} + \boldsymbol{w}^\top \boldsymbol{\Lambda}_0^{-1}\boldsymbol{w} - 2\boldsymbol{\mu}_0^\top \boldsymbol{\Lambda}_0^{-1}\boldsymbol{w}\right)\right). \end{aligned}$$

# 5.6 贝叶斯统计

- 现在我们定义 $\boldsymbol{\Lambda}_m = (\boldsymbol{X}^\top \boldsymbol{X} + \boldsymbol{\Lambda}_0^{-1})^{-1}$ 和 $\boldsymbol{\mu}_m = \boldsymbol{\Lambda}_m(\boldsymbol{X}^\top \boldsymbol{y} + \boldsymbol{\Lambda}_0^{-1}\boldsymbol{\mu}_0)$.
- 使用上述新的变量，改写高斯分布：

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

- 检查此后验分布可以让我们获得贝叶斯推断效果的一些直觉。大多数情况下， 我们设置 μ0 = 0。如果我们设置 Λ0 = $\frac{1}{\alpha} I$，那么 μm 对 w 的估计就和频率派带权重 衰减惩罚 $\alpha w^\top w$ 的线性回归的估计是一样的

# 5.6 贝叶斯统计

- 5.6.1 最大后验估计
- 目标：让先验影响点估计的选择来利用贝叶斯方法的优点，而不是简单 地回到最大似然估计
- 最大后验（Maximum A Posteriori, MAP）点估计。MAP估计选择后验概率最大的点（或在 θ 是连续值 的更常见情况下，概率密度最大的点）：

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

- 我们可以认出上式右边的 logp(x| θ) 对应着标准的对数似然项，logp(θ) 对应着先 验分布。