Machine Learning Basics -1

Weiliang Teng

20161204

OverView

- 5.1 Learning Algorithms
- 5.2 Capacity, Overfitting and Underfitting
- 5.3 Hyperparameters and Validation Sets
- 5.4 Estimators, Bias and Variance
- 5.5 Maximum Likelihood Estimation
- 5.6 Bayesian Statistics

5.1 Learning Algorithms

- 定义： 对于某类任务T和性能度量P，如果一个计算机程序在T上以P衡量的性能随着经验E而自我完善，那么我们称这个计算机程序在从经验E中学习

5.1.1 The Task, T

- 分类:
- 有缺失输入数据的分类:
- 回归: $f : \mathbb{R}^n \rightarrow \mathbb{R}$　　算法交易
- 转录，改写: OCR, 语音识别
- 机器翻译
- 结构化输出： Parsing, 图像分割
- 异常检测:
- 合成与采样： 语音合成
- 填充：缺失数据
- 去噪
- 概率密度函数估计

5.1.2 The Performance Measure, P

- 准确度，错误率
- expected 0-1 loss
- average log-probability
- 测试集
- 有时候不好评价，例如语音识别任务或者某些回归任务

5.1.3 The Experience, E

- 无监督学习
  - a dataset containing many features,then learn useful properties of the structure of this dataset
  - 学习 p(x)
- 有监督学习
  - a dataset containing features, but each example is also associated with a label or target
  - 学习 p(y|x)

5.1.4 Example: Linear Regression

$$\boldsymbol{x} \in \mathbb{R}^n \qquad y \in \mathbb{R} \qquad \boldsymbol{w} \in \mathbb{R}^n$$

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

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

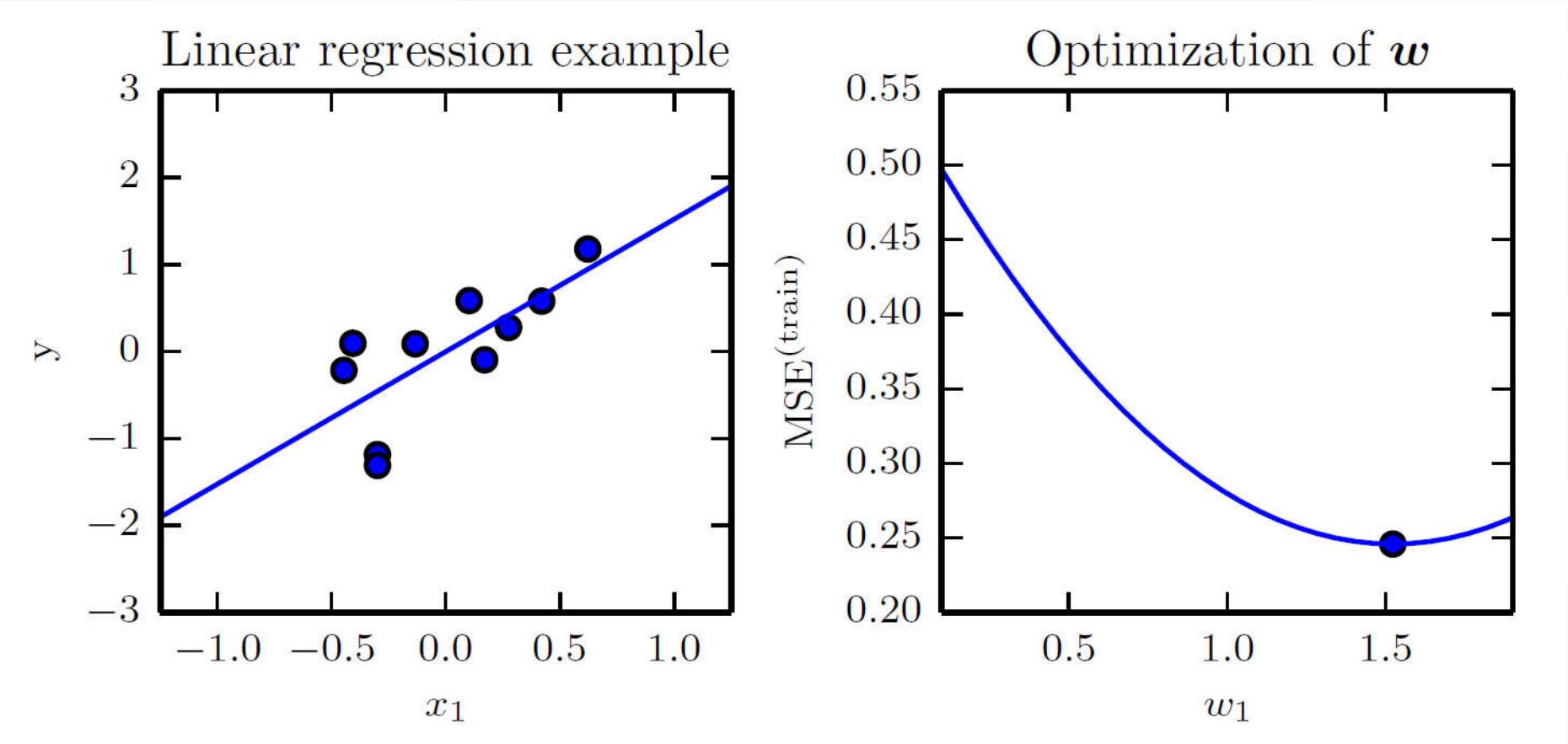
$$\text{MSE}_{\text{test}} = \frac{1}{m} ||\hat{\boldsymbol{y}}^{(\text{test})} - \boldsymbol{y}^{(\text{test})}||_2^2$$

$$\nabla_{\boldsymbol{w}} \text{MSE}_{\text{train}} = 0$$

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

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

5.1.4 Example: Linear Regression



5.1.4 Example: Linear Regression

$$\Rightarrow \nabla_{\boldsymbol{w}} \left(\boldsymbol{X}^{(\text{train})} \boldsymbol{w} - \boldsymbol{y}^{(\text{train})} \right)^{\top} \left(\boldsymbol{X}^{(\text{train})} \boldsymbol{w} - \boldsymbol{y}^{(\text{train})} \right) = 0 \tag{5.9}$$

$$\Rightarrow \nabla_{\boldsymbol{w}} \left(\boldsymbol{w}^{\top} \boldsymbol{X}^{(\text{train})\top} \boldsymbol{X}^{(\text{train})} \boldsymbol{w} - 2\boldsymbol{w}^{\top} \boldsymbol{X}^{(\text{train})\top} \boldsymbol{y}^{(\text{train})} + \boldsymbol{y}^{(\text{train})\top} \boldsymbol{y}^{(\text{train})} \right) = 0 \tag{5.10}$$

$$\Rightarrow 2\boldsymbol{X}^{(\text{train})\top} \boldsymbol{X}^{(\text{train})} \boldsymbol{w} - 2\boldsymbol{X}^{(\text{train})\top} \boldsymbol{y}^{(\text{train})} = 0 \tag{5.11}$$

$$\Rightarrow \boldsymbol{w} = \left(\boldsymbol{X}^{(\text{train})\top} \boldsymbol{X}^{(\text{train})} \right)^{-1} \boldsymbol{X}^{(\text{train})\top} \boldsymbol{y}^{(\text{train})} \tag{5.12}$$

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

5.2 Capacity, Overfitting and Underfitting

- 泛化

  The ability to perform well on previously unobserved inputs

- 训练误差
- 泛化误差 测试误差

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

$$\frac{1}{m^{(\text{test})}}||\boldsymbol{X}^{(\text{test})}\boldsymbol{w}-\boldsymbol{y}^{(\text{test})}||_2^2.$$

5.2 Capacity, Overfitting and Underfitting

- 统计学习理论
- 数据生成过程 i.i.d. 独立同分布

  要求训练集与测试集中的数据i.i.d.来自同一个数据生成分布
- 训练目标 p_{data}
  - Make the training error small
  - Make the gap between training and test error small

5.2 Capacity, Overfitting and Underfitting

- 欠拟合
  - 模型的训练集上的误差无法降得充分低
- 过拟合

  模型的训练误差与测试误差有较大差别
- capacity

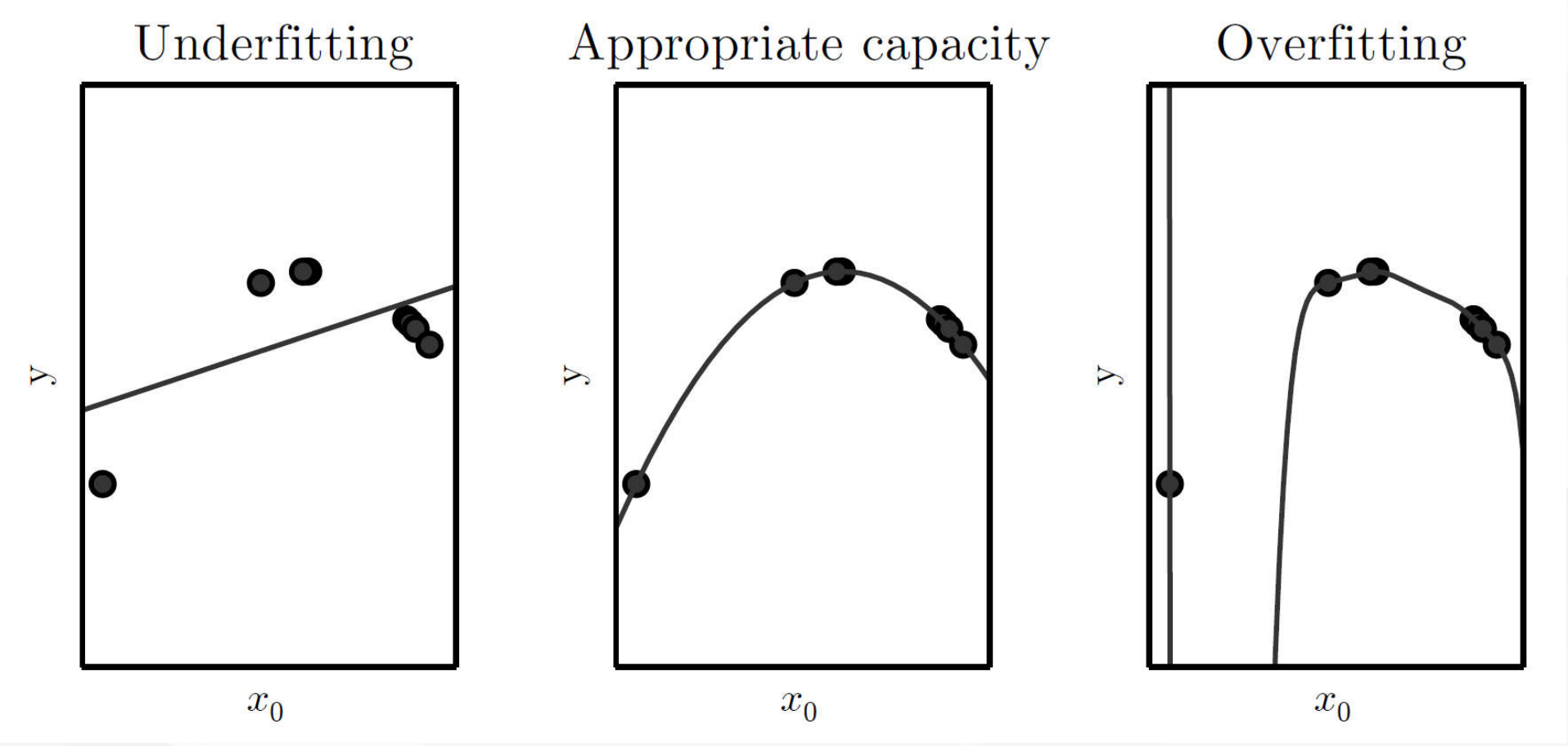
ability to fit a wide variety of functions

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

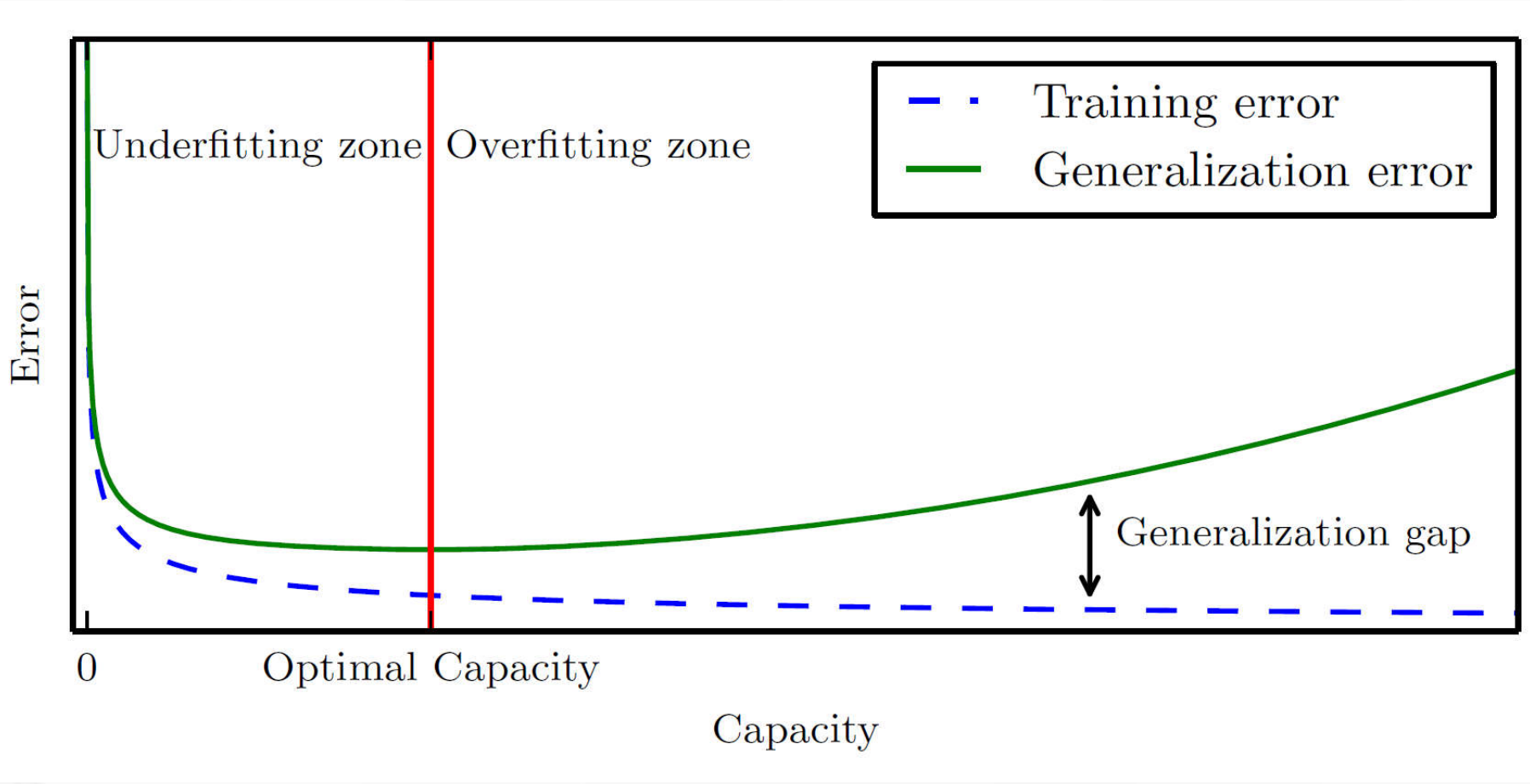
$$\hat{y} = b + w_1 x + w_2 x^2 .$$

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

5.2 Capacity, Overfitting and Underfitting



5.2 Capacity, Overfitting and Underfitting



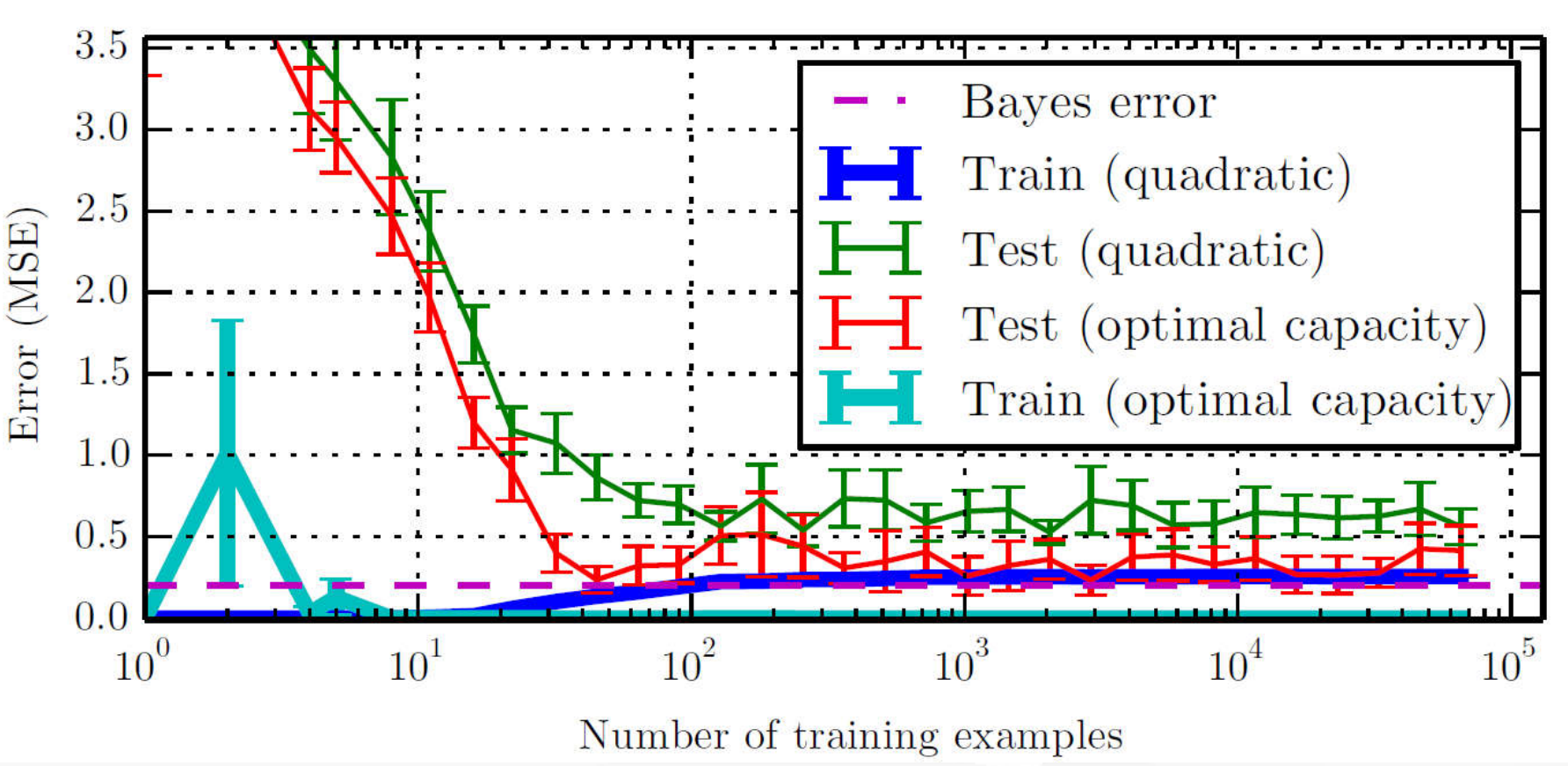
5.2 Capacity, Overfitting and Underfitting

- 奥卡姆剃刀法则

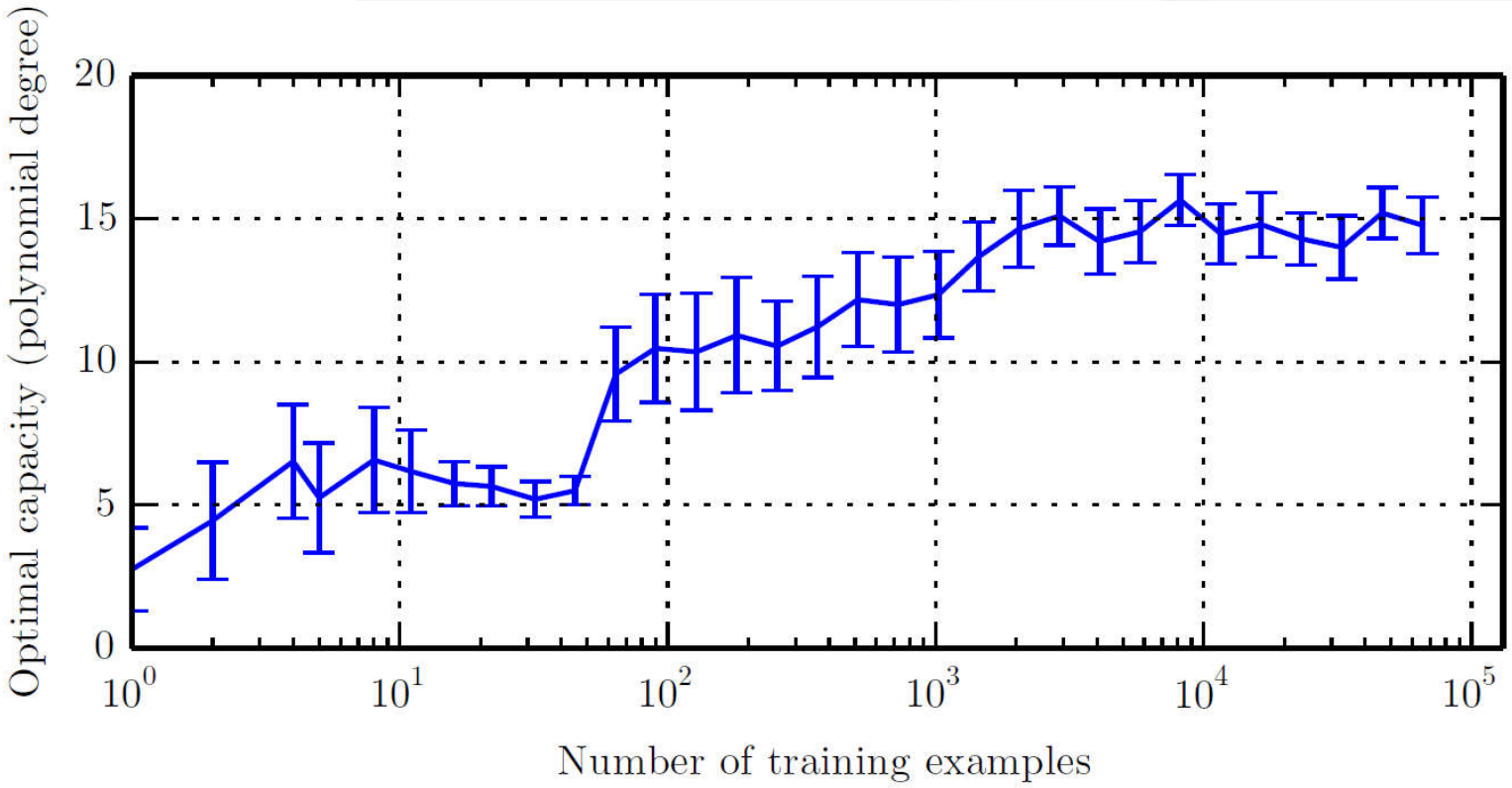
  如无必要，勿增实体

  among competing hypotheses that explain known observations equally well, one should choose the “simplest” one.

5.2.1 The No Free Lunch Theorem

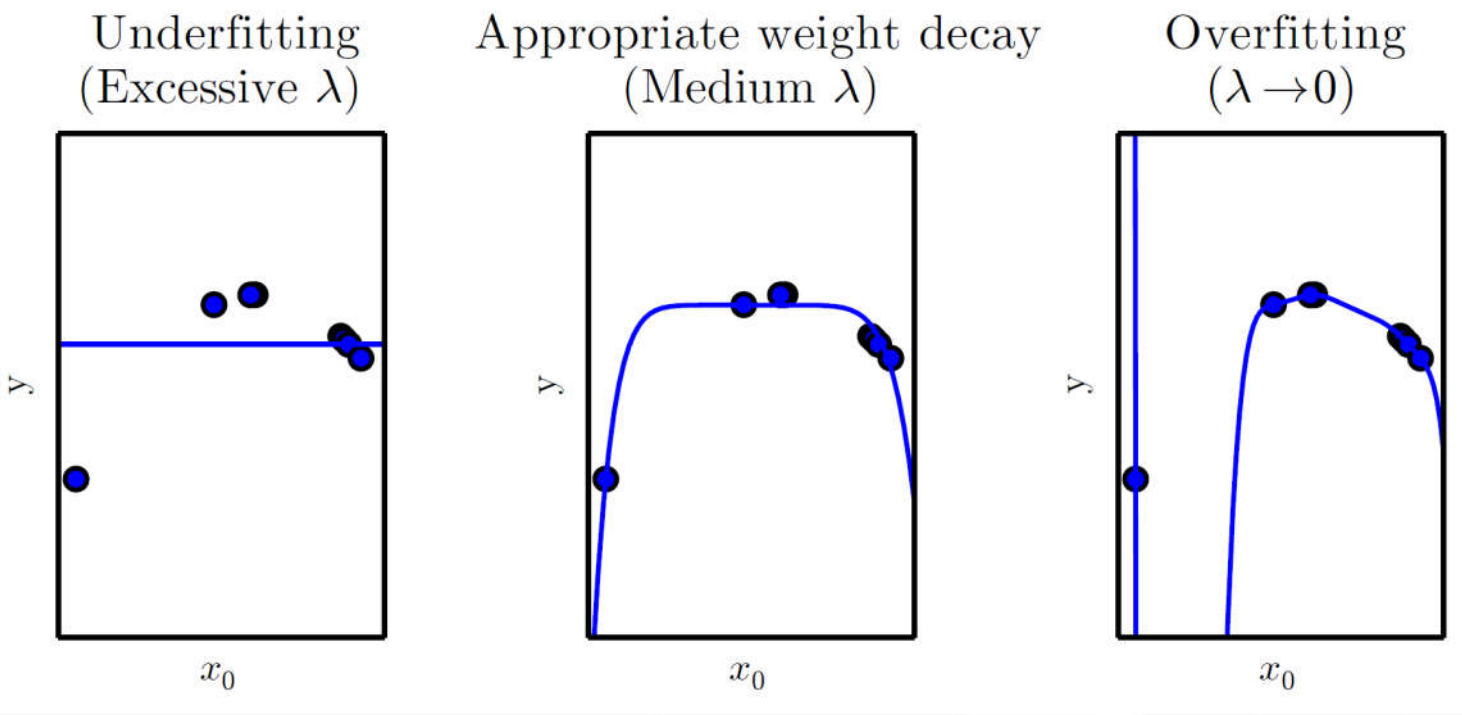


5.2.1 The No Free Lunch Theorem



5.2.2 Regularization

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



5.3 Hyperparameters and Validation Sets

- 超参数：机器学习中模型里面的框架参数
  - 神经网络的层数，每层的神经元个数
  - 聚类中的类别个数
  - 多项式回归中最高项的幂次数
  - 正则项系数 λ

- 超参数不好设定，靠人工经验设定，或者网格搜索

5.3 Hyperparameters and Validation Sets

- Validation Sets：验证集

  Training Sets, Validation Sets, Test Sets

  The validation set is used for model selection (调超参数）

  The test set for final model (the model which was selected by selection process) prediction error

  一般来说 Training data中80%用于training, 20%用于validation

5.3.1 Cross-Validation

- 数据集很小的时候，为了充分利用所有数据，采用交叉验证法，代价是增加计算量

“交叉验证法”(cross validation)先将数据集 D 划分为 k 个大小相似的互斥子集, 即 $D = D_1 \cup D_2 \cup \ldots \cup D_k$, $D_i \cap D_j = \varnothing$ $(i \neq j)$. 每个子集 D_i 都尽可能保持数据分布的一致性, 即从 D 中通过分层采样得到. 然后, 每次用 $k-1$ 个子集的并集作为训练集, 余下的那个子集作为测试集; 这样就可获得 k 组训练/测试集, 从而可进行 k 次训练和测试, 最终返回的是这 k 个测试结果的均值. 显然, 交叉验证法评估结果的稳定性和保真性在很大程度上取决于 k 的取值, 为强调这一点, 通常把交叉验证法称为“ k 折交叉验证”(k-fold cross validation). k 最常用的取值是 10, 此时称为 10 折交叉验证; 其他常用的 k 值有 5、20 等. 图 2.2 给出了 10 折交叉验证的示意图.

5.3.1 Cross-Validation

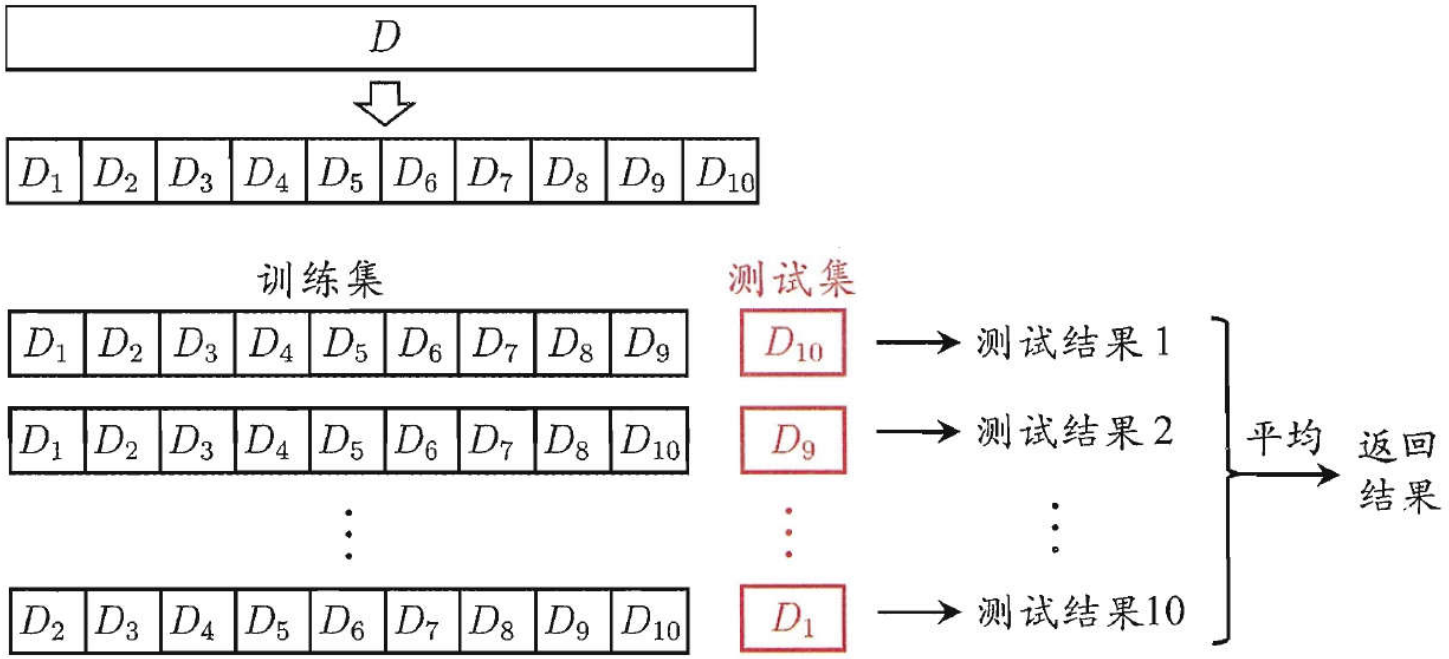


图 2.2 10 折交叉验证示意图

5.4 Estimators, Bias and Variance

- 5.4.1 Point Estimation

$$\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$$

是一组来自独立同分布的m个数据样本点

点估计器 $\hat{\boldsymbol{\theta}}_m = g(\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)})$.

- 函数估计：函数空间上的点估计
- $\hat{f}$ 估计 f

$$\boldsymbol{y} = f(\boldsymbol{x}) + \boldsymbol{\epsilon}$$

5.4.2 Bias

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

- 无偏估计：$\text{bias}(\hat{\boldsymbol{\theta}}_m) = \mathbf{0}$
- 渐进无偏估计：$\lim_{m\to\infty} \text{bias}(\hat{\boldsymbol{\theta}}_m) = \mathbf{0}$

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

5.4.2 Bias

- 高斯分布的均值估计

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

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

5.4.2 Bias

- 高斯分布的方差估计

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

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

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

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

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

5.4.3 Variance and Standard Error

点估计的方差与标准差

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

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

均值估计的95%置信区间

$$(\hat{\mu}_m - 1.96\mathrm{SE}(\hat{\mu}_m), \hat{\mu}_m + 1.96\mathrm{SE}(\hat{\mu}_m))$$

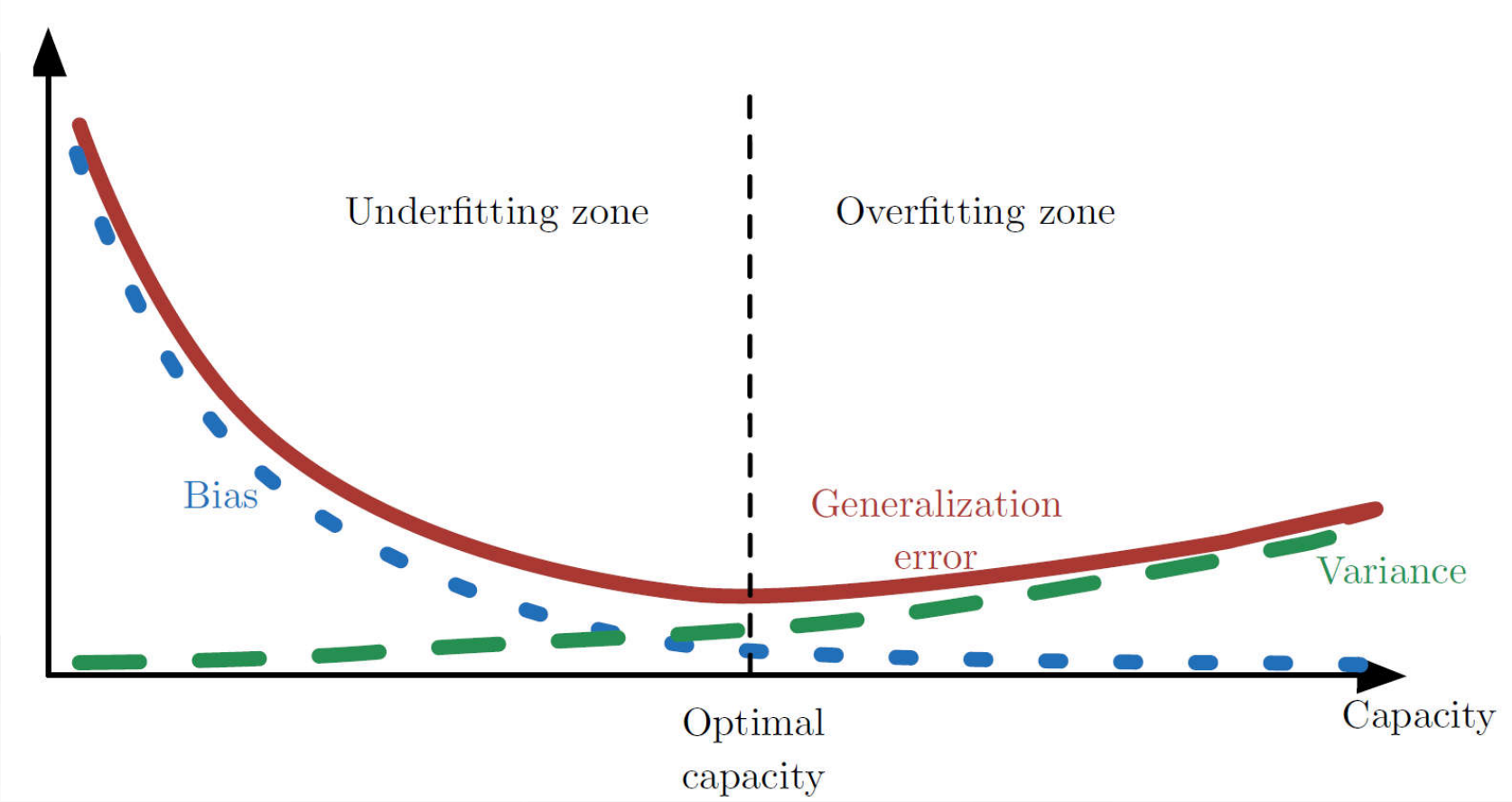
5.4.4 Trading off Bias and Variance to Minimize Mean Squared Error

均方误差（Mean Squared Error, MSE）包含了估计器偏差和方差，是一种较方便的衡量方法

一个好的估计器要使得均方误差减小

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

5.4.4 Trading off Bias and Variance to Minimize Mean Squared Error



5.4.5 Consistency

弱一致性

$$\lim_{m\to\infty} \hat{\theta}_m \xrightarrow{p} \theta.$$

强一致性　　几乎处处收敛

一致性保证了随着数据的增长， 估计器的偏差将逐渐消失

5.5 Maximum Likelihood Estimation

- 数据生成概率分布 $p_{\text{data}}(\mathbf{x})$
- 参数化的数据概率估计函数族 $p_{\text{model}}(\boldsymbol{x}; \boldsymbol{\theta})$
- 极大似然估计

$$\begin{aligned}\boldsymbol{\theta}_{\text{ML}} &= \arg\max_{\boldsymbol{\theta}} p_{\text{model}}(\mathbb{X}; \boldsymbol{\theta}) \\ &= \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{m} p_{\text{model}}(\boldsymbol{x}^{(i)}; \boldsymbol{\theta})\end{aligned}$$

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

5.5 Maximum Likelihood Estimation

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

最小化 K L 熵

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

等价于最小化 $-\mathbb{E}_{\mathbf{x}\sim\hat{p}_{\mathrm{data}}}\left[\log p_{\mathrm{model}}(\boldsymbol{x})\right]$

5.5.1 Conditional Log-Likelihood and Mean Squared Error

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

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

极大似然线性估计器

$$p(y \mid \boldsymbol{x}) = \mathcal{N}(y; \hat{y}(\boldsymbol{x}; \boldsymbol{w}), \sigma^2)$$

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

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

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

5.6 Bayesian Statistics

贝叶斯方法　用概率分布表征预先知道的信息

先验概率分布　p(θ)

某个区间上的均匀分布，高斯分布

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

序列预测

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

5.6 Bayesian Statistics

- 贝叶斯方法在训练数据有限时，泛化性能好
- 缺点是如果训练数据量大时，需要大量的运算

5.6.1 Maximum A Posteriori (MAP) Estimation

- 贝叶斯方法有些时候难以实现，点估计方法比较好实现。两者结合，最大后验方法MAP

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

带正则项的极大似然方法，可以部分解释为MAP

Q&A

Thanks !