机器学习导论

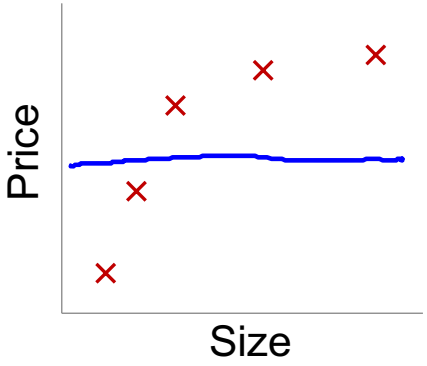
第11章. 机器学习的相关理论

谢茂强

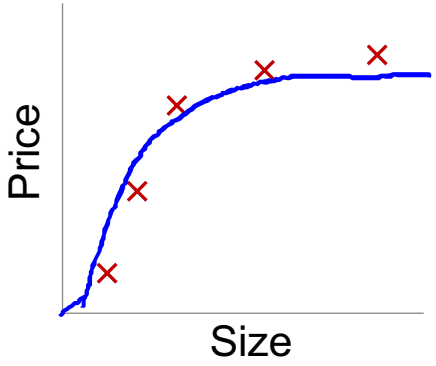
南开大学软件学院

目录

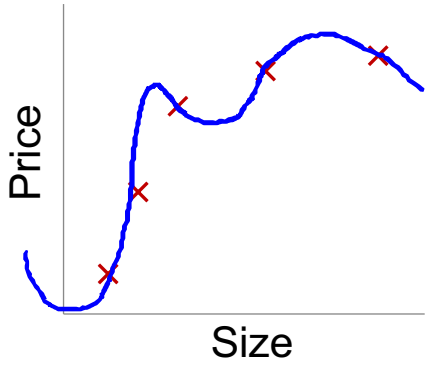
偏差-方差模型背景：模型的复杂程度、过拟合、欠拟合



High bias (underfit)



“Just right”



High variance (overfit)

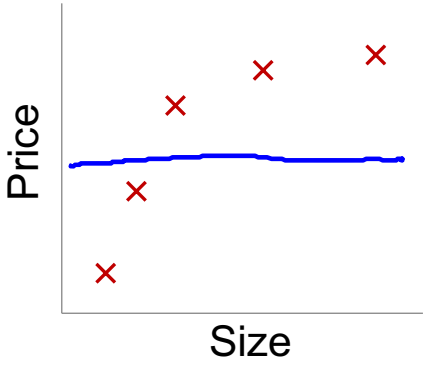
偏差-方差模型：度量过拟合和欠拟合

- 模型在实际预测时表现不佳的两类原因：
  - 过拟合(Overfitting)：高方差(High Variance)
  - 欠拟合(Underfitting)：高偏差(High Bias)
- 导致过拟合和欠拟合的因素
  - 训练样本的数量、样本特征的好坏(数量、表征能力)、特征变换
  - 模型复杂程度、正则化项权重
- 偏差（Bias）:模型在训练样本集的错误率（简称训练错误）
- 方差（Variance）:模型验证错误和训练错误间的差异

正则化：解决过拟合的主要手段

Model: $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \frac{\lambda}{2m}\sum_{j=1}^{m}\theta_j^2$$

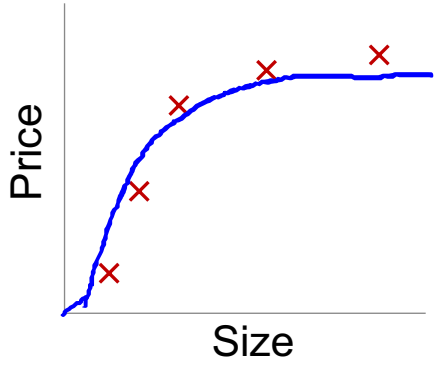


Large λ

High bias (underfit)

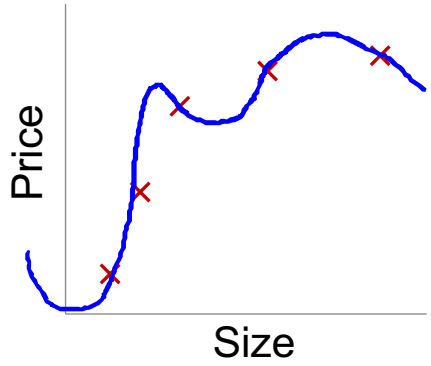
$\lambda = 10000.\ \theta_1 \approx 0, \theta_2 \approx 0, \ldots$

$h_\theta(x) \approx \theta_0$



Intermediate λ

"Just right"



Small λ

High variance (overfit)

不同正则化项的线性回归

- 为避免过拟合，使用正则化方法提升线性回归
  - Ridge Regression

$$J(\theta_{ridge}) = \frac{1}{2m}\sum_{i=1}^{m}\left(y^{(i)} - \theta_0 - \sum_{j=1}^{n}\theta_j x_j^{(i)}\right)^2 + \lambda\sum_{j=1}^{n}\theta_j^2$$

  - LASSO: Least Absolute Shrinkage and Selection Operator (1996)

$$J(\theta_{lasso}) = \frac{1}{2m}\sum_{i=1}^{m}\left(y^{(i)} - \theta_0 - \sum_{j=1}^{n}\theta_j x_j^{(i)}\right)^2 + \lambda\sum_{j=1}^{n}|\theta_j|$$

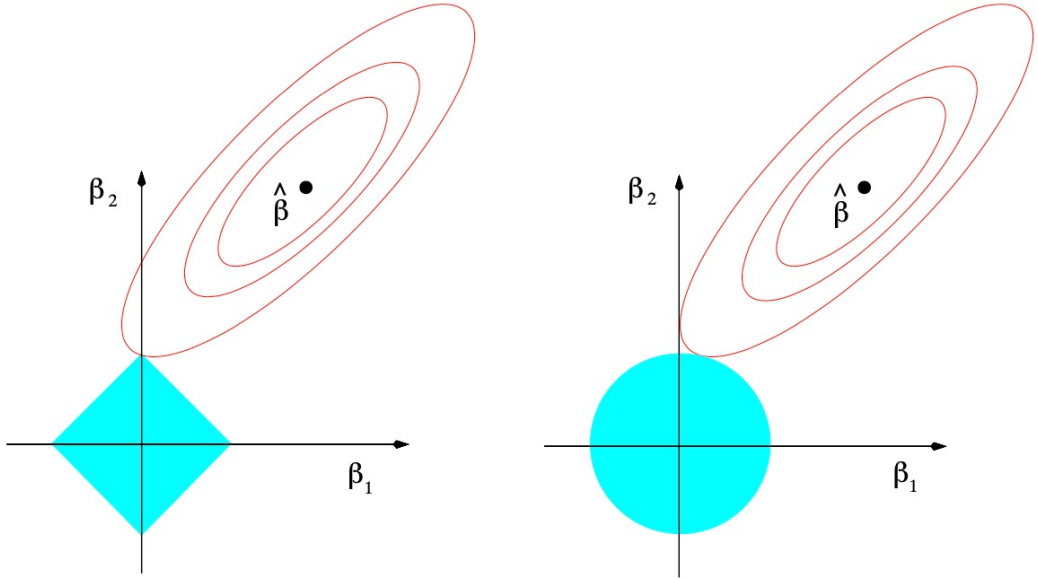


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$*, respectively, while the red ellipses are the contours of the least squares error function.*

图来自T. Hastie, R. Tibshirani, J. Friedman. The Elements of Statistic Learning: Data Mining, Inference, and Prediction. 2017

通过正则化参数 λ 控制模型对训练数据的拟合程度

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \frac{\lambda}{2m}\sum_{j=1}^{m}\theta_j^2$$

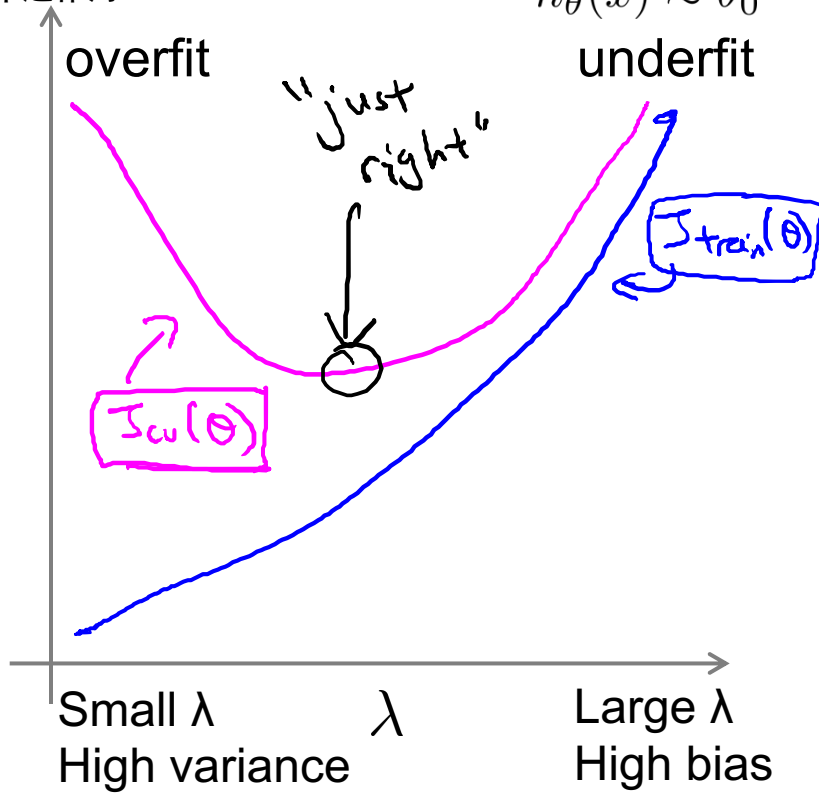
$$J_{train}(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$$

$$J_{cv}(\theta) = \frac{1}{2m_{cv}}\sum_{i=1}^{m_{cv}}(h_\theta(x_{cv}^{(i)}) - y_{cv}^{(i)})^2$$

主要是训练样本起作用，模型惩罚不起作用

$\lambda = 10000.\ \theta_1 \approx 0, \theta_2 \approx 0, \ldots$

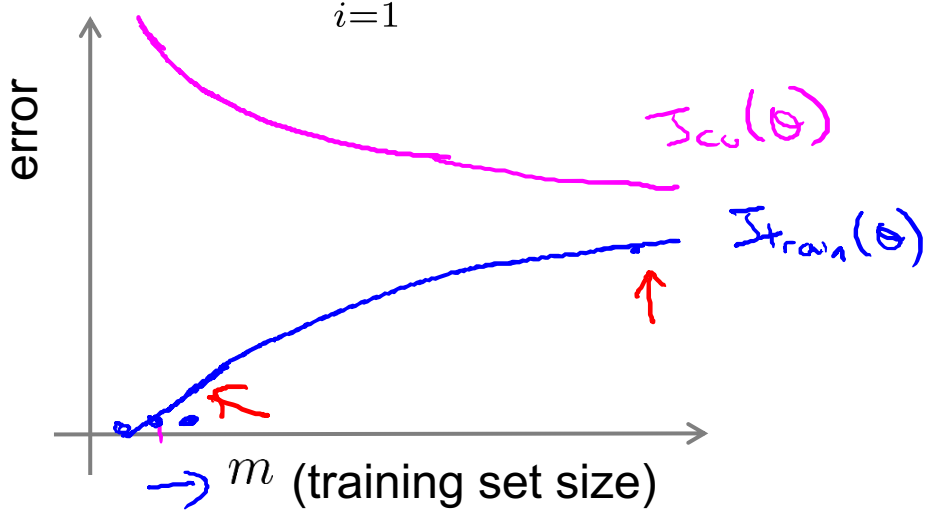
$h_\theta(x) \approx \theta_0$

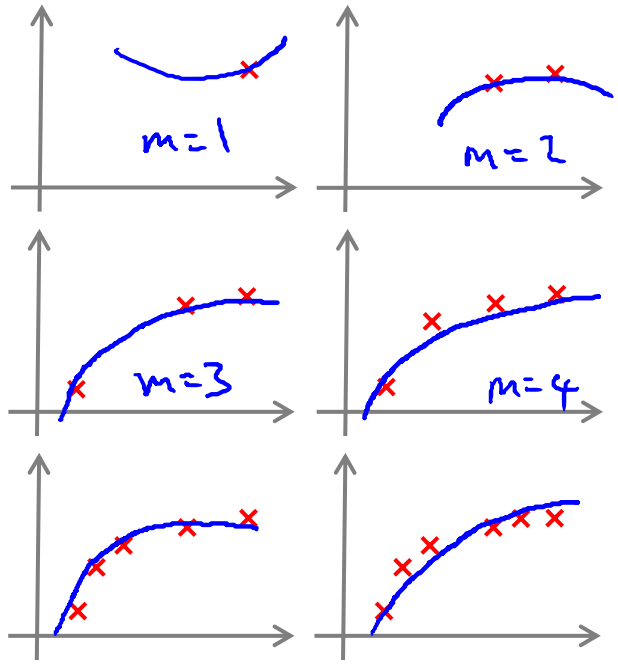


使用学习曲线(Learning curves)来反映模型的学习程度(或拟合训练和验证数据的程度)

$$J_{train}(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$$

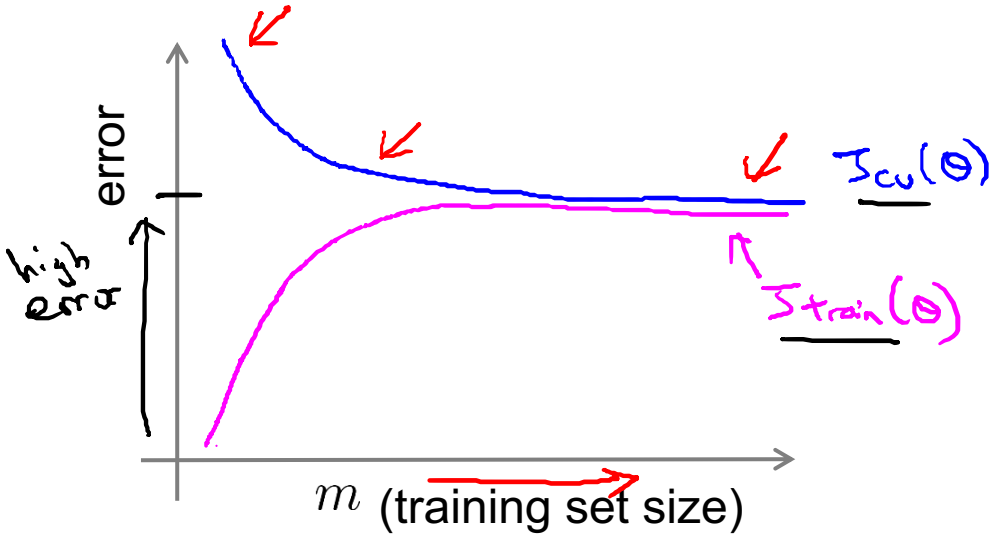
$$J_{cv}(\theta) = \frac{1}{2m_{cv}}\sum_{i=1}^{m_{cv}}(h_\theta(x_{cv}^{(i)}) - y_{cv}^{(i)})^2$$

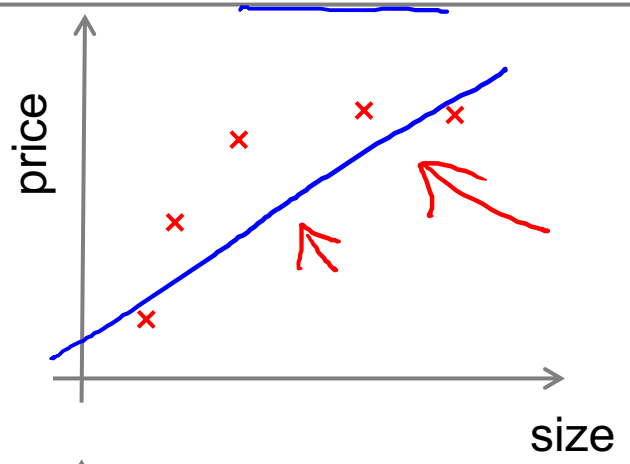


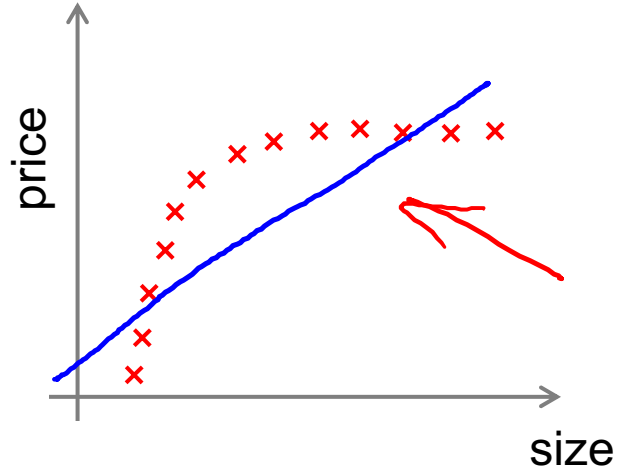


$$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2$$

使用学习曲线展示欠拟合(Underfit, High bias) $h_\theta(x) = \theta_0 + \theta_1 x$





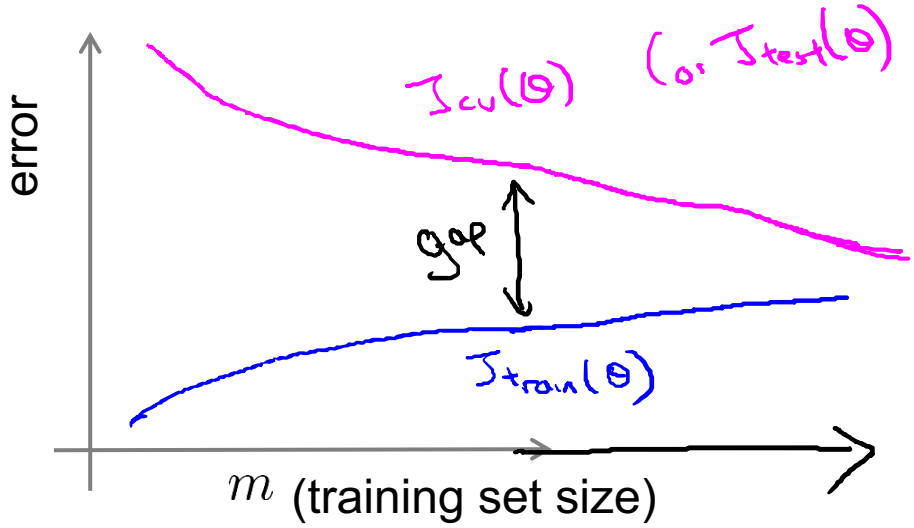


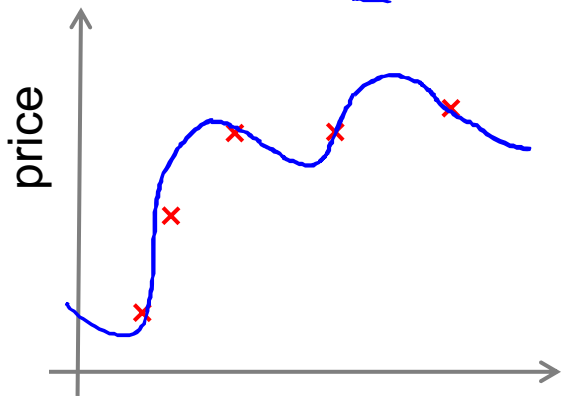
- 模型太简单，在训练数据和验证数据上错误率都高
- 错误率高的情形不会因为训练数据集规模扩大而改善

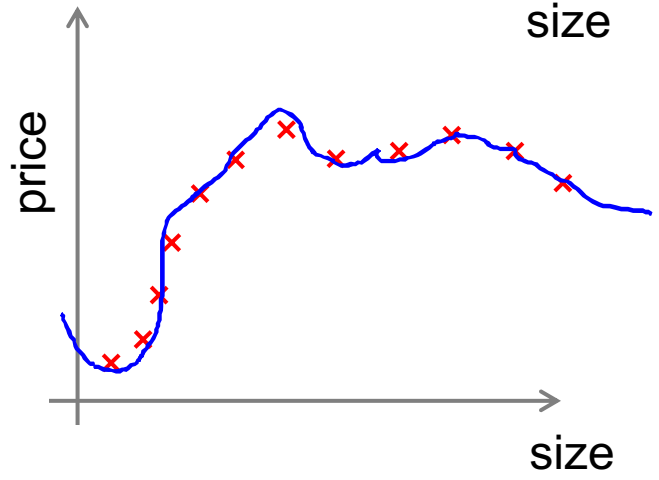
使用学习曲线展示过拟合(高方差)

$h_\theta(x) = \theta_0 + \theta_1 x + \cdots + \theta_{100} x^{100}$

(and small λ)







- 过拟合：模型复杂，且训练数据和验证数据分布差异大
- 扩大训练数据集规模（主要是让其分布更接近验证数据集）对解决过拟合很有帮助

偏差-方差理论的正式定义

- 偏差（Bias）:期望输出与真实标记的差别成为偏差

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

- 方差（Variance）:使用样本数相同的不同训练集产生的方差

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

泛化误差分解为偏差、方差和噪声之和

$$E(f;D) = bias^2(x) + var(x) + \varepsilon^2$$

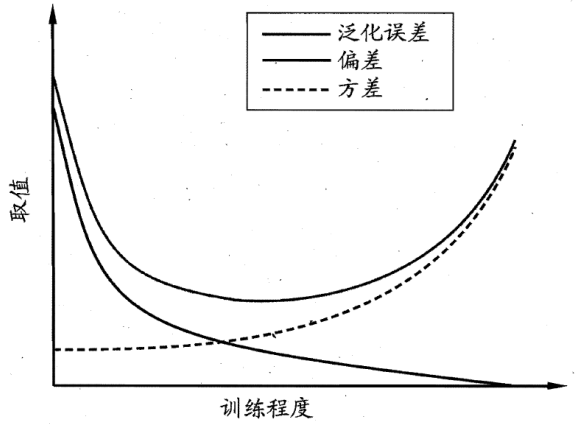
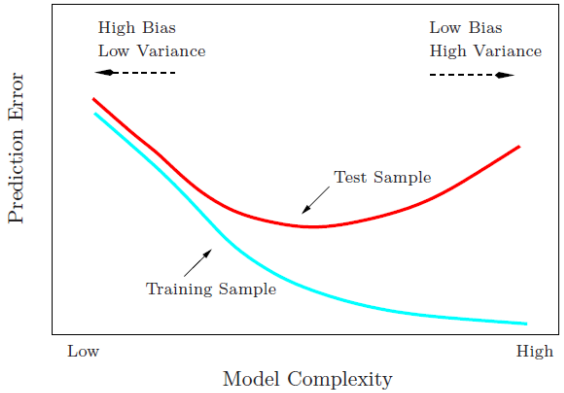
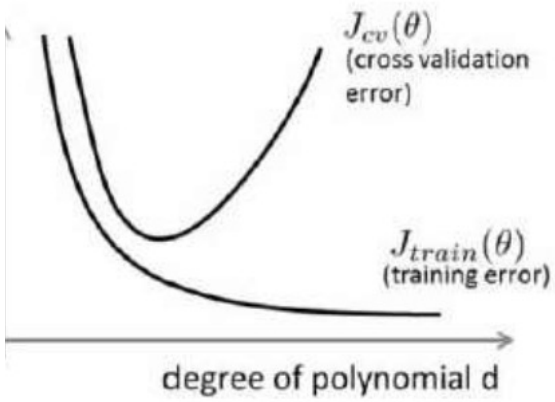
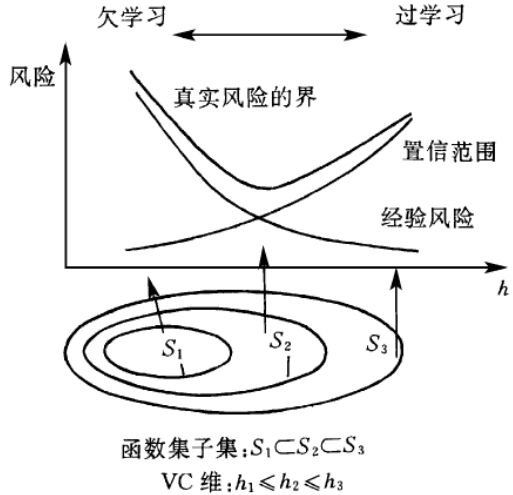


图 2.9 泛化误差与偏差、方差的关系示意图







函数集子集：$S_1 \subset S_2 \subset S_3$

VC 维：$h_1 \leqslant h_2 \leqslant h_3$

FIGURE 2.11. *Test and training error as a function of model complexity.*

2. 统计学习理论

- Vapnik和 Chervonenkis 1962~1973期间提出
- **模型的泛化能力**：对于独立同分布的测试样本推断的能力。（模型能否被generalized to测试样本的推断上）
- 发现学习模型的泛化能力主要取决于学习所基于的假设空间(某模型所有可能解的集合)的复杂性。
- Vapnik 和 Chervonenkis 提出了**VC维**(Vapnik-Chervonenkis Dimension)等著名的用于刻画**假设空间复杂性的度量**,从而来估计和控制学习模型的泛化能力.

Journal of Machine Learning Research 16 (2015) 1677-1681 Published 9/15

Preface to this Special Issue

Alex Gammerman ALEX@CS.RHUL.AC.UK
Vladimir Vovk V.VOVK@RHUL.AC.UK
Computer Learning Research Centre, Department of Computer Science
Royal Holloway, University of London

This issue of JMLR is devoted to the memory of Alexey Chervonenkis. Over the period of a dozen years between 1962 and 1973 he and Vladimir Vapnik created a new discipline of statistical learning theory—the foundation on which all our modern understanding of pattern recognition is based. Alexey was 28 years old when they made their most famous and original discovery, the uniform law of large numbers. In that short period Vapnik and Chervonenkis also introduced the main concepts of statistical learning theory, such as VC-dimension, capacity control, and the Structural Risk Minimization principle, and designed two powerful pattern recognition methods, Generalised Portrait and Optimal Separating Hyperplane, later transformed by Vladimir Vapnik into Support Vector Machine—arguably one of the best tools for pattern recognition and regression estimation. Thereafter Alexey continued to publish original and important contributions to learning theory. He was also active in research in several applied fields, including geology, bioinformatics, medicine, and advertising.

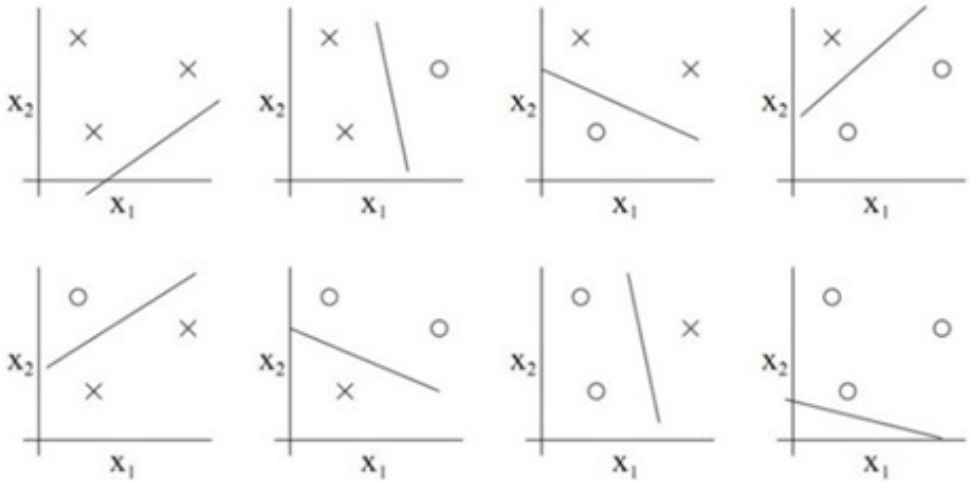
Alexey tragically died in September 2014 after getting lost during a hike in the Elk Island park on the outskirts of Moscow. Vladimir Vapnik suggested to prepare an issue of JMLR to be published at the first anniversary of the death of his long-term collaborator and close friend. Vladimir and the editors contacted a few dozen leading researchers in the fields of

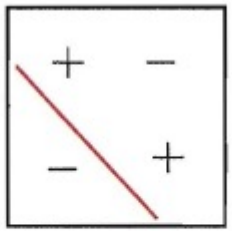
VC维与假设空间打散(Scatter)样本集

- 假设空间打散样本的能力体现模型的能力

定义 12.7 假设空间 $\mathcal{H}$ 的 VC 维是能被 $\mathcal{H}$ 打散的最大示例集的大小, 即

$$\mathrm{VC}(\mathcal{H}) = \max\{m : \Pi_{\mathcal{H}}(m) = 2^m\} . \qquad (12.23)$$





对任何集合，其 $2^4 = 16$ 种对分中至少有一种不能被线性划分实现

(b) 示例集大小为 4

周志华. 机器学习. 12.4节. P273

2. VC维的作用

定理 12.3 若假设空间 $\mathcal{H}$ 的 VC 维为 d, 则对任意 $m > d$, $0 < \delta < 1$ 和 $h \in \mathcal{H}$ 有

$$P\left(E(h) - \widehat{E}(h) \leqslant \sqrt{\frac{8d\ln\frac{2em}{d} + 8\ln\frac{4}{\delta}}{m}}\right) \geqslant 1 - \delta\ . \tag{12.29}$$

证明 令 $4\,\Pi_{\mathcal{H}}(2m)\exp(-\frac{m\epsilon^2}{8}) \leqslant 4(\frac{2em}{d})^d\exp(-\frac{m\epsilon^2}{8}) = \delta$, 解得

$$\epsilon = \sqrt{\frac{8d\ln\frac{2em}{d} + 8\ln\frac{4}{\delta}}{m}}\ ,$$

代入定理 12.2, 于是定理 12.3 得证. ■

由定理 12.3 可知, 式(12.29)的泛化误差界只与样例数目 m 有关, 收敛速率为 $O(\frac{1}{\sqrt{m}})$, 与数据分布 $\mathcal{D}$ 和样例集 D 无关. 因此, 基于 VC 维的泛化误差界是分布无关 (distribution-free)、数据独立 (data-independent) 的.

Rademacher复杂度

定理 12.6 给出了基于 Rademacher 复杂度的泛化误差界. 与定理 12.3 对比可知, 基于 VC 维的泛化误差界是分布无关、数据独立的, 而基于 Rademacher 复杂度的泛化误差界(12.47)与分布 $\mathcal{D}$ 有关, 式(12.48)与数据 D 有关. 换言之, 基于 Rademacher 复杂度的泛化误差界依赖于具体学习问题上的数据分布, 有点类似于为该学习问题“量身定制”的, 因此它通常比基于 VC 维的泛化误差界更紧一些.

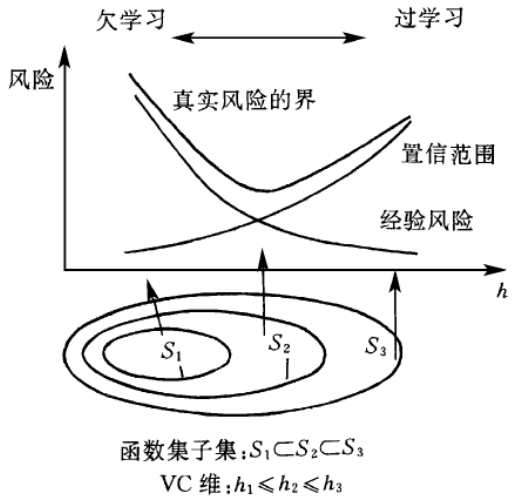
吴新星, 张军平. Rademacher复杂度在统计学习理论中的研究. 自动化学报. 43(1):20-39. 2017.

STL、SVM和SRM

- 结构风险最小化（经验风险+置信风险）

$$\epsilon = \sum_{i=1}^{N} L(p_i) = \sum_{i=1}^{N} L[f(\boldsymbol{X}_i, \boldsymbol{w}), y_i]$$

$$\Omega(f) = \|\boldsymbol{w}\|^p$$



- SVM是以VC维为核心的统计学习理论的代表性方法
  - 最小化泛化误差上界
  - 结构风险最小化
  - 最大化Margin
  - 正则化

对于结构风险有不同的定义，依据上下文判定

2.从SLT看过拟合、偏差、方差

- 从偏差和方差角度看，很难同时追求最优
- 从统计学习理论角度
  - 低偏差代表经验风险最小化
  - 低方差代表结构风险最小化

- SVM的结构风险最小化、最大Margin和正则化

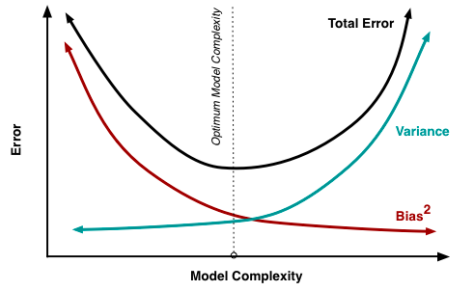
2. 正则化与结构风险最小化

Table 1. The training-set loss functions and the regularizers of eight classifiers

Classifier	training-set loss: $g_1(y_i\vec{x}_i\vec{\beta})$	regularizer: $g_2(\vec{\beta})$
Regularized LLSF	$\sum_{i=1}^{n}(1-y_i\vec{x}_i\vec{\beta})^2$	$\lambda\|\vec{\beta}\|^2$
Regularized LR	$\sum_{i=1}^{n}\log(1+exp(-y_i\vec{x}_i\vec{\beta}))$	$\lambda\|\vec{\beta}\|^2$
Regularized 2-layer NNet	$\sum_{i=1}^{n}(1-\pi(y_i\vec{x}_i\vec{\beta}))^2$	$\lambda\|\vec{\beta}\|^2$
SVM	$\sum_{i=1}^{n}(1-y_i\vec{x}_i\vec{\beta})_+$	$\lambda\|\vec{\beta}\|^2$
Rocchio	$-\sum_{y_i=1}y_i\vec{x}_i\vec{\beta}-\frac{bN_c}{N_{\overline{c}}}\sum_{y_i=-1}y_i\vec{x}_i\vec{\beta}$	$\frac{N_c}{2}\|\vec{\beta}\|^2$
Prototype	$-\sum_{y_i=1}y_i\vec{x}_i\vec{\beta}$	$\frac{N_c}{2}\|\vec{\beta}\|^2$
kNN	$-\sum_{y_i=1\wedge\vec{x}_i\in R_k(\vec{x})}y_i\vec{x}_i\vec{\beta}_x$	$\frac{1}{2}\|\vec{\beta}_x\|^2$
NB without smoothing	$-\sum_{y_i=1}y_i\vec{x}_i\vec{\beta}$	$S_c\|e^{\vec{\beta}}\|_1$
NB with Laplace smoothing	$-\sum_{y_i=1}y_i\vec{x}_i\vec{\beta}$	$(p+S_c)\|e^{\vec{\beta}}\|_1+\|\vec{\beta}\|_1$

Fan Li, YiMing Yang. A Loss Function Analysis for Classification Methods in Text Categorization. ICML 2013.

2. 偏差-方差、泛化误差、结构风险最小化



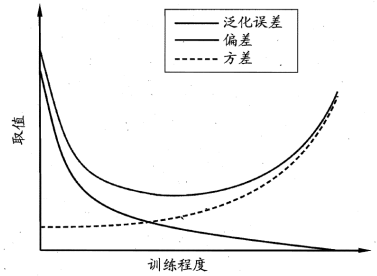
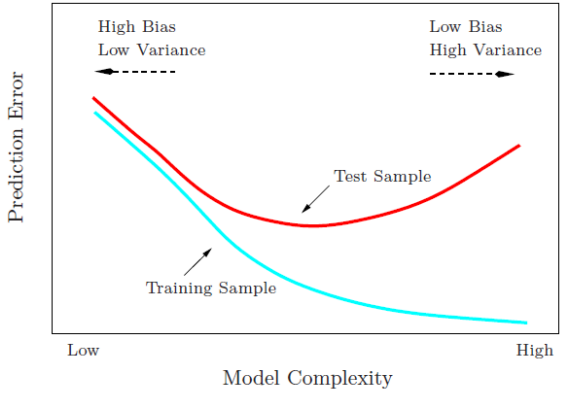


图 2.9 泛化误差与偏差、方差的关系示意图

泛化误差分解为偏差、方差和噪声之和[1]

$$E(f;D) = bias^2(x) + var(x) + \varepsilon^2$$



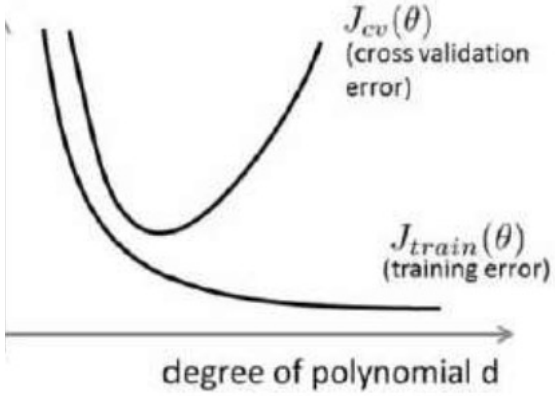
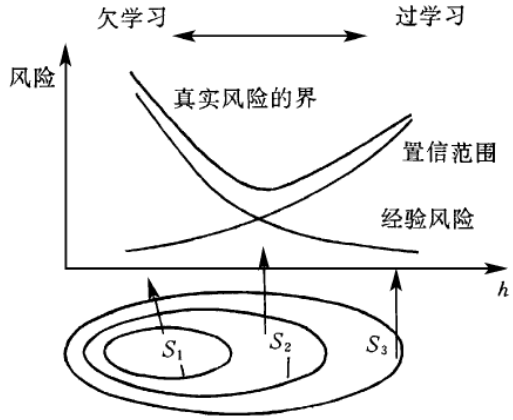


FIGURE 2.11. *Test and training error as a function of model complexity.*



函数集子集：$S_1 \subset S_2 \subset S_3$

VC 维：$h_1 \leqslant h_2 \leqslant h_3$

[1]. 见周志华《机器学习》公式（2.42）. P45

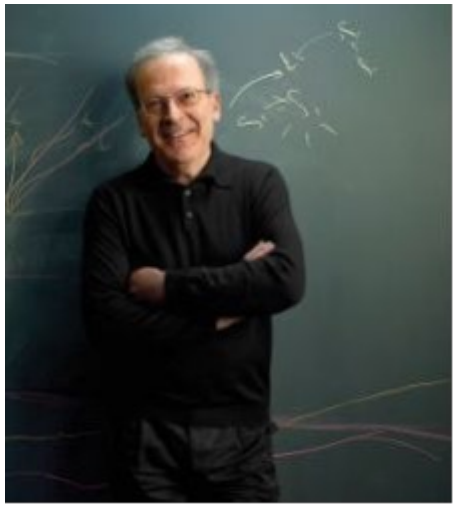
3. 计算学习理论与PAC学习

- 计算学习理论(Computational Learning Theory)
  - 研究关于通过“计算”来进行“学习”的理论，
  - 目的是分析学习任务的困难本质，为学习算法提供理论基础。
- 几个重要的不等式(估计与期望间误差足够小的概率)
  - Jensen不等式
  - Hoeffding不等式
  - McDiarmid不等式

3. PAC学习

- Valiant等人1984年提出，2010年获图灵奖

- PAC：Probably Approximately Correct
- 弱学习算法---识别错误率小于1/2(即准确率仅比随机猜测略高的学习算法)
- 强学习算法---识别准确率很高并能在多项式时间内完成的学习算法



3. PAC学习

- Valiant和Kearns首次提出PAC学习模型中弱学习和强学习算法的等价性问题，即任意给定仅比随机猜测略好的弱学习算法，是否可以将其提升为强学习算法？
- 如果二者等价，那么只需找到一个比随机猜测略好的弱学习算法就可以将其提升为强学习算法，而不必寻找很难获得的强学习算法。
- Schapire在1990年证明了该等价性，97年提出了**AdaBoost**，PAC学习开始得到关注。
- 开启了集成学习有理论支撑的局面。

PAC学习的主要概念

- PAC 辨识(PAC Identity)学习算法L能从假设空间对中辨识概念类C.
- PAC 可学习(PAC Learnable)从假设空间H中PAC 辨识概念类C，则称概念类C 对假设空间H而言是PAC可学习的
- PAC学习算法
- 样本复杂度
- 不可知PAC 可学习(agnostic PAC learnable)

4.其他统计学习问题

1. 假设空间
2. 归纳学习与归纳偏好（或偏置）

4.1 假设空间

- 我们可以把学习过程看做一个在所有假设组成的空间中进行搜索的过程，搜索目标是找到与训练及匹配的假设。也就是说能够将训练集中判断正确的假设。
- 可见的样本空间到不可见的函数族/概念空间

4.2 归纳学习与归纳偏置

- 归纳(Induction)是从特殊到一般的泛化过程。既从具体的事实归结出一般性规律。（区别于演绎Deduction）
- 而从样例中学习，显然是一个归纳的过程，因此也称归纳学习。
- 狭义的归纳学习要求从训练数据中学得概念(数据挖掘)
- 从归纳的角度，假设空间可以是候选概念集合，归纳学习是从概念集合（假设空间）中找到与训练集匹配(fit)的概念或假设。

偏置（bias）：bias的含义要结合上下文环境来判断。

4.2 典型的归纳偏好

- 归纳偏好：评估候选假设或概念的准则，并利用该特性将机器学习工作从枚举变为搜索（带剪枝）。智能也体现在此。
  - 最大条件独立性
  - 最小交叉验证误差
  - 最大间隔（Maximum Margin）
  - 最小描述长度（Minimum description length，奥卡姆剃刀）。
  - 最少特征数
  - 最近邻居

4.2 奥卡姆剃刀

- 奥卡姆剃刀(Occam's Razor)：**若有多个假设与观察一致，则选最简单的那一个**。是一种常用的自然科学研究中最基本的原则
- 我们认为更平滑，意味着更简单。模型更简洁的表达更容易表达，意味着更简单。（统计学习中模型复杂程度的惩罚来源于此。）

5.(统计)机器学习的一些整理

- 机器学习的目标是使得学到的模型能够很好地适用于新样本，而不是仅仅在训练样本上工作的很好。 学得的模型适用于新样本的能力，被称之为泛化能力。
- 我们的训练样本是从样本空间中抽样而来。每个样本都是独立的，从这个分布上采用获得的。
- 目前，提到的机器学习往往是统计机器学习。

5. 统计学习的研究

- 统计学习方法
  - 开发新的学习方法
- 统计学习理论
  - 探究统计学习的基本理论
  - 学习方法的有效性和效率
- 统计学习应用
  - 将现实应用问题建模为学习问题
  - 并解决建模过程中应用问题和模型不匹配的问题。

5.统计学习的三要素

判决函数

条件概率函数

Loss or Cost

ERM v.s. SRM

Gradient Descent

MLE

EM

SDP SMO for SVM

…

统计学习三要素 - 模型

- 条件概率分布、判决函数
- 包含条件概率分布或判决函数的假设空间
- 模型给定后，假设空间可以进一步被表达成函数族或分布族，
- 进而，统计机器学习也可以看做是在这些族的参数空间中找到最优参数的过程。

统计学习三要素 – 策略

- 探究统计学习的基本理论
  - 有了模型的假设空间，需要考虑按照何种准则来学习，或者选择最优模型。
- 损失函数与风险函数
  - 0-1损失函数（0-1 Loss Function）
  - 平方损失函数（Quadratic Loss Function）
  - 对数损失函数或对数似然损失函数（Logarithmic or log-likelihood）
  - 指数损失函数（Exponential Loss Function）

统计学习三要素 – 策略 – 风险函数

- 理论上模型 $f(X)$关于联合分布 $P(X,Y)$ 的损失函数的期望

$$R_{exp}(f) = E_P[L(Y, f(X))] = \int_{\mathcal{X}\times\mathcal{Y}} L(y, f(x))P(x,y)dxdy$$

- 困难： $P(X,Y)$ 是未知的。无法直接得到 $P(Y|X)$ 。
- 统计学习需要联合分布，但联合分布是未知的。因此监督学习就成为病态问题(ill-formed/ill-posed problem)

统计学习三要素 – 策略 – 风险

- 经验风险最小化

$$\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i))$$

- 结构风险最小化

$$\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} \Big(L(y_i, f(x_i)) + \lambda R(f) \Big)$$

统计学习三要素 – 算法

- 算法：学习模型的具体计算方法
  - 根据学习策略，从假设空间中选择最优模型的计算方法
- 通用计算算法
  - 最小二乘法、GD、半正定规划、二次规划、最大似然估计…
- 专用算法
  - SMO for SVM、LARS for LASSO

5. 另一种的三要素

Table 1: The three components of learning algorithms.

Representation	Evaluation	Optimization
Instances	Accuracy/Error rate	Combinatorial optimization
K-nearest neighbor	Precision and recall	Greedy search
Support vector machines	Squared error	Beam search
Hyperplanes	Likelihood	Branch-and-bound
Naive Bayes	Posterior probability	Continuous optimization
Logistic regression	Information gain	Unconstrained
Decision trees	K-L divergence	Gradient descent
Sets of rules	Cost/Utility	Conjugate gradient
Propositional rules	Margin	Quasi-Newton methods
Logic programs		Constrained
Neural networks		Linear programming
Graphical models		Quadratic programming
Bayesian networks		
Conditional random fields		

统计学习的应用

- **将现实应用问题建模为学习问题**
  - 自然语言处理（序列学习）
  - 计算视觉（特征抽取）
  - 信息检索（排序学习）
  - 生物信息中的蛋白质交互作用预测（网络学习）
- **并解决建模过程中应用问题和模型不匹配的问题**
  - 序列消歧、SIFT、特征学习、代价敏感的排序学习、双向网络游走的补齐