机器学习
第10节

涉及知识点：
学习理论、可学性、假设空间ERM边界、VC维、
模型选择、偏差-方差分解、交叉验证、特征选择

学习理论与模型选择

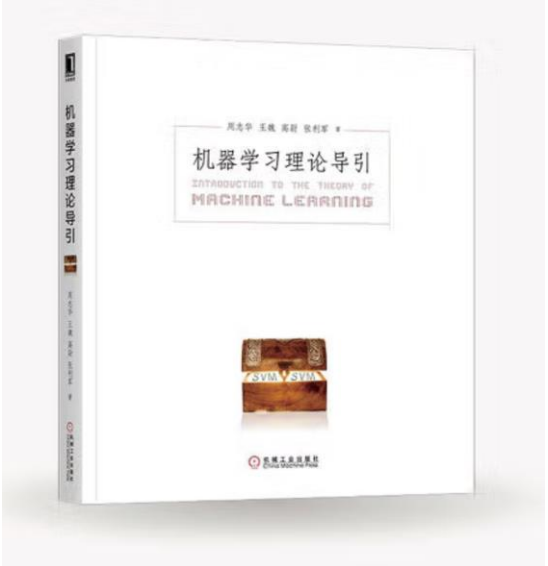
张伟楠 – 上海交通大学

学习理论

张伟楠 – 上海交通大学

学习理论

- 学习理论
  - 可学性
  - 样本复杂度
  - 有限/无限假设空间ERM边界
  - VC维



PAC可学（Probably Approximately Correct Learnable）

- 概念类C，即所有概念的集合$\{c\}$
  - 例如{三角形，四边形，五边形}

- PAC辨识（PAC Identify）：给定$0 < \epsilon, \delta < 1$，对所有概念$c \in C$和分布p，若存在学习算法L，其输出假设$h \in \mathcal{H}$满足

$$P(E(h) \leq \epsilon) \geq 1 - \delta$$

则称学习算法L能从假设空间$\mathcal{H}$中PAC辨识概念类C。

- PAC可学：令D表示从分布p中独立同分布地采样得到的数据集，其大小为N，给定$0 < \epsilon, \delta < 1$，对所有分布p，若存在学习算法 L 和多项式函数 $\text{poly}(\cdot,\cdot,\cdot,\cdot)$，使得对于任何$N \geq \text{poly}\left(\frac{1}{\epsilon}, \frac{1}{\delta}, \text{size(x)}, \text{size}(c)\right)$，$L$ 能从假设空间$\mathcal{H}$中PAC辨识概念类C，则称概念类C对假设空间$\mathcal{H}$而言是PAC可学的，简称概念类C是PAC可学的。
  - size(x)表示数据本身的复杂度，$\text{size}(c)$表示概念的复杂度
  - 满足PAC学习算法L所需的最小数据集大小N，称为学习算法L样本复杂度

学习理论

- 在计算复杂性或样本复杂性方面刻画具体的学习问题或具体算法的定理的总和
  - 例如，学习给定准确性的假设所必需的或者足够的训练实例的数量

$$\epsilon(d, N, \delta) = \sqrt{\frac{1}{2N}\left(\log d + \log \frac{1}{\delta}\right)}$$

误差　训练样本　假设空间　正确的概率

学习理论

- 学习问题的复杂性取决于：
  - 假设空间的大小或表达性
  - 目标概念必须近似的准确性
  - 学习器一定产生成功假设的可能性
  - 训练样本的方式，例如随机训练或向oracle查询

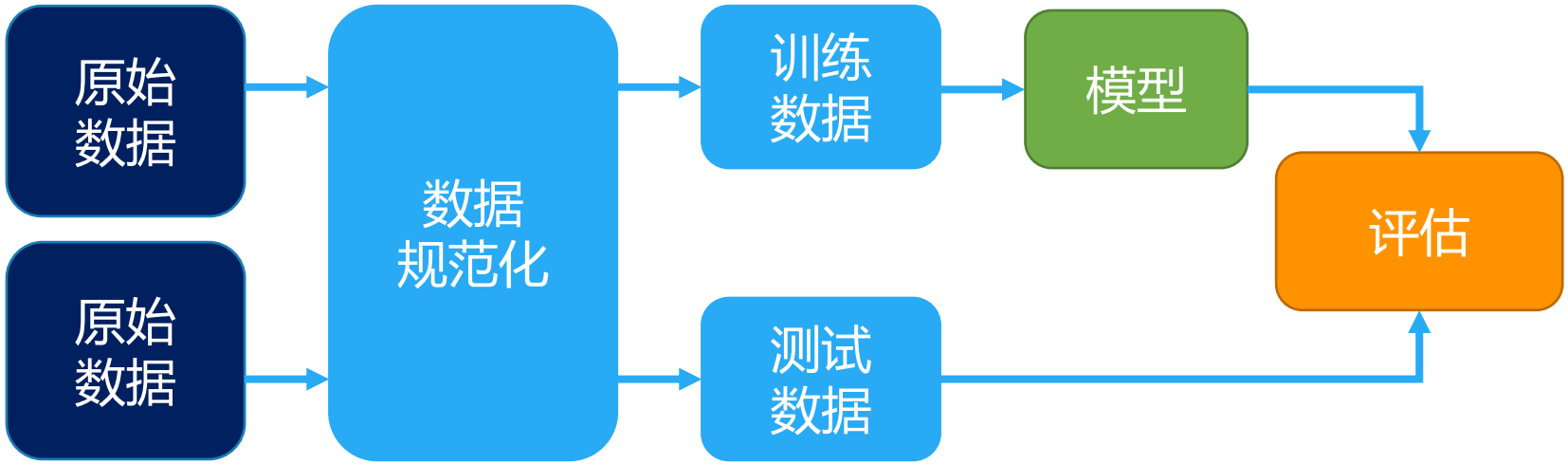
$$\epsilon(d, N, \delta) = \sqrt{\frac{1}{2N}\left(\log d + \log \frac{1}{\delta}\right)}$$

误差　训练样本　假设空间　正确的概率

假设空间ERM边界

张伟楠 – 上海交通大学

机器学习过程



1. 选择了“好的”超参数
2. 对整个训练数据训练模型
3. 用测试数据对模型进行测试

泛化能力

- 泛化能力指的是模型对未观测数据的预测能力
  - 可以通过泛化误差来评估，定义如下：

$$R(f) = \mathbb{E}\left[\mathcal{L}\left(Y, f(X)\right)\right] = \int_{X \times Y} \mathcal{L}\left(y, f(x)\right) p(x, y) dx dy$$

  - $p(x, y)$是潜在的（可能是未知的）联合数据分布
- 在训练数据集上对泛化能力的经验估计是：

$$\hat{R}(f) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}\left(y_i, f(x_i)\right)$$

泛化误差

- 有限假设集 $\mathcal{F} = \{f_1, f_2, \dots, f_d\}$
- 泛化误差约束定理：

对任意函数$f \in \mathcal{F}$，以不小于$1 - \delta$的概率满足下式：

$$R(f) \leq \hat{R}(f) + \epsilon(d, N, \delta)$$

其中

$$\epsilon(d, N, \delta) = \sqrt{\frac{1}{2N}(\log d + \log\frac{1}{\delta})}$$

- N：训练实例个数
- d：假设集的函数个数

引理： Hoeffding不等式

- 令$X_1, X_2, \ldots, X_n$为独立同分布的随机变量，其中$X_i \in [a, b]$，那么平均变量Z为：

$$Z = \frac{1}{n}\sum_{i=1}^{n} X_i$$

那么下述不等式成立：

$$P(Z - \mathbb{E}[Z] \geq t) \leq \exp\left(\frac{-2nt^2}{(b-a)^2}\right)$$

$$P(\mathbb{E}[Z] - Z \geq t) \leq \exp\left(\frac{-2nt^2}{(b-a)^2}\right)$$

http://cs229.stanford.edu/extra-notes/hoeffding.pdf

泛化误差约束定理证明

- 对二值分类，误差率$0 \leq R(f) \leq 1$
- 基于Hoeffding不等式，对$\epsilon > 0$，有

$$P\left(R(f) - \hat{R}(f) \geq \epsilon\right) \leq \exp(-2N\epsilon^2)$$

- 由于$\mathcal{F} = \{f_1, f_2, \ldots, f_d\}$为有限集，满足

$$\begin{aligned} P\left(\exists f \in \mathcal{F}: R(f) - \hat{R}(f) \geq \epsilon\right) &= P\left(\bigcup_{f \in \mathcal{F}} \left\{R(f) - \hat{R}(f) \geq \epsilon\right\}\right) \\ &\leq \sum_{f \in \mathcal{F}} P\left(R(f) - \hat{R}(f) \geq \epsilon\right) \\ &\leq d \exp(-2N\epsilon^2) \end{aligned}$$

泛化误差约束定理证明

- 下述不等式等价：

$$P(\exists f \in \mathcal{F}: R(f) - \hat{R}(f) \geq \epsilon) \leq d\exp(-2N\epsilon^2)$$
$$\Updownarrow$$
$$P(\forall f \in \mathcal{F}: R(f) - \hat{R}(f) < \epsilon) \geq 1 - d\exp(-2N\epsilon^2)$$

- δ参数化方式如下：

$$\delta = d\exp(-2N\epsilon^2) \Leftrightarrow \epsilon = \sqrt{\frac{1}{2N}\log\frac{d}{\delta}}$$

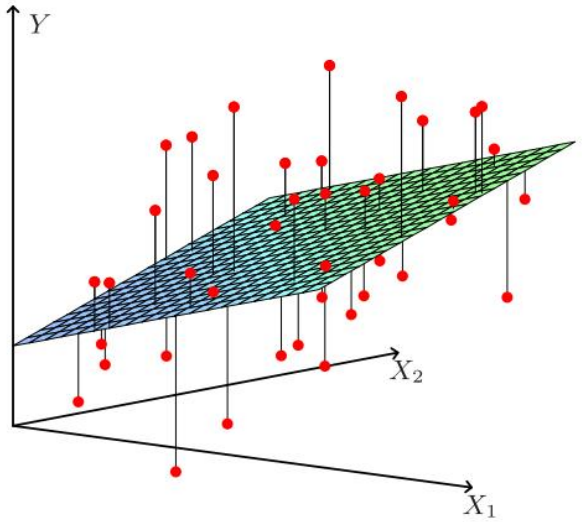
那么泛化误差以$1-\delta$概率被约束

$$P(R(f) < \hat{R}(f) + \epsilon) \geq 1 - \delta$$

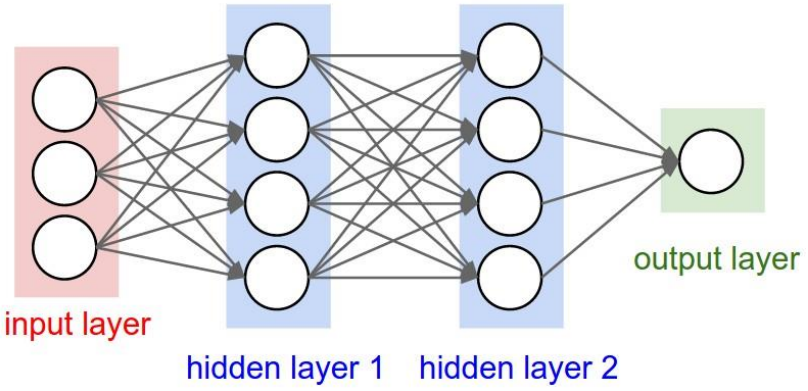
对无限假设空间

- 许多假设类，包括任何由实数参数化的假设类实际上都包含无数个函数
- 例如：线性模型、神经网络

$$f(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$



$$f(x) = \sigma(W_3(W_2 \tanh(W_1 x + b_1) + b_2) + b_3)$$



量化实数

- 假设我们有一个H假设空间，由m个实数参数化
- 在计算机中，每个实数用64位表示（双浮点数）
- 因此，假设类实际上包括至多$d = 2^{64m}$种不同的假设

$$\epsilon(d, N, \delta) = \sqrt{\frac{1}{2N}\left(\log d + \log\frac{1}{\delta}\right)}$$

$$\Rightarrow \epsilon(d, N, \delta) = \sqrt{\frac{1}{2N}\left(64m + \log\frac{1}{\delta}\right)}$$

$$\Rightarrow N = \frac{1}{2\epsilon^2}\left(64m + \log\frac{1}{\delta}\right) = O_{\epsilon,\delta}(m)$$

样本复杂度

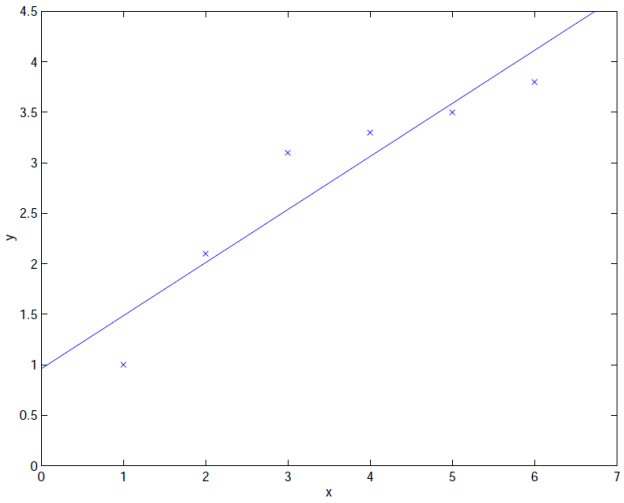
- 对于由m个实数参数化的模型，为了以至少$1-\delta$的概率获得不高于ϵ的泛化误差，我们需要N个训练样本使得

$$N \geq \frac{1}{2\epsilon^2}(64m + \log\frac{1}{\delta}) = O_{\epsilon,\delta}(m)$$

- 它与参数数量线性相关

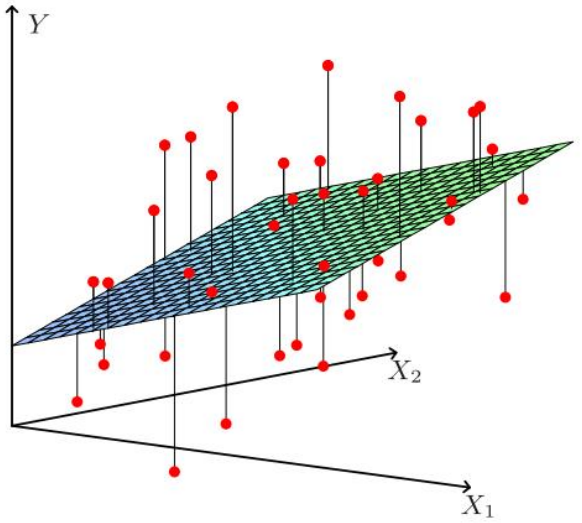
样本复杂度的例子

- 在K维数据上拟合线性回归



$$f(x) = \theta_0 + \theta_1 x$$

对于1维数据线性回归，通常需要大约10个点来以一定置信度拟合直线



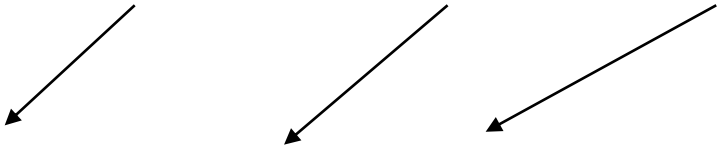
$$f(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

对于2维数据线性回归，通常需要大约20个点来以一定置信度拟合超平面

样本复杂度的例子

- 在K维数据上拟合线性回归
- 一个标准的特征工程范例

x=[工作日=星期五,性别=男性,城市=上海，...]



x=[0,0,0,0,1,0,0 0,1 0,0,1,0...0, ...]

```
1 5:1 9:1 12:1 45:1 154:1 509:1 4089:1 45314:1 988576:1
0 2:1 7:1 18:1 34:1 176:1 510:1 3879:1 71310:1 818034:1
...
```

$$f(x) = \theta_0 + \sum_{i=1}^{10^6} \theta_i x_i$$

对于 100 万维数据线性回归，我们通常需要大约 1000 万条数据来以一定置信度拟合一条直线

工业界训练机器学习模型

- 大公司拥有庞大的数据工程团队和多年的数据
- 模型精度的瓶颈：模型的复杂度 -> 训练时间
  - 不一定要构建很复杂的模型在给定的训练集上尽量提升效果
  - 而是通过扩展数据量和特征集，结合线性模型，来提升效果

特征扩展：构建更多的特征

数据扩展：往前看越早的数据，就能获得更大的数据量，但是早期数据可能偏移

$$\epsilon(d, N, \delta) = \sqrt{\frac{1}{2N}\left(\log d + \log\frac{1}{\delta}\right)}$$

$$\Rightarrow \epsilon(d, N, \delta) = \sqrt{\frac{1}{2N}\left(64m + \log\frac{1}{\delta}\right)}$$

$$\Rightarrow N = \frac{1}{2\epsilon^2}\left(64m + \log\frac{1}{\delta}\right) = O_{\epsilon,\delta}(m)$$

VC维

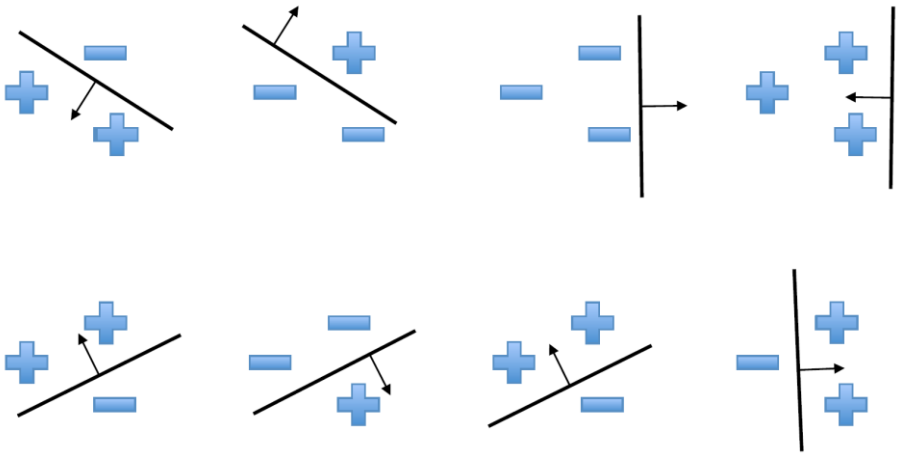
张伟楠 – 上海交通大学

打散（Shattering）

- 定义
  - 一个模型类可以打散成一个数据点集

$$x^{(1)}, x^{(2)}, \ldots, x^{(n)}$$

如果对于这些点上可能存在的每个标签，在该类中存在获得零训练误差的模型



例如：线性模型类打散成3点集

VC维

- 可打散的X子集越大，假设空间越具有表达性，即偏差越小

Vladimir Vapnik

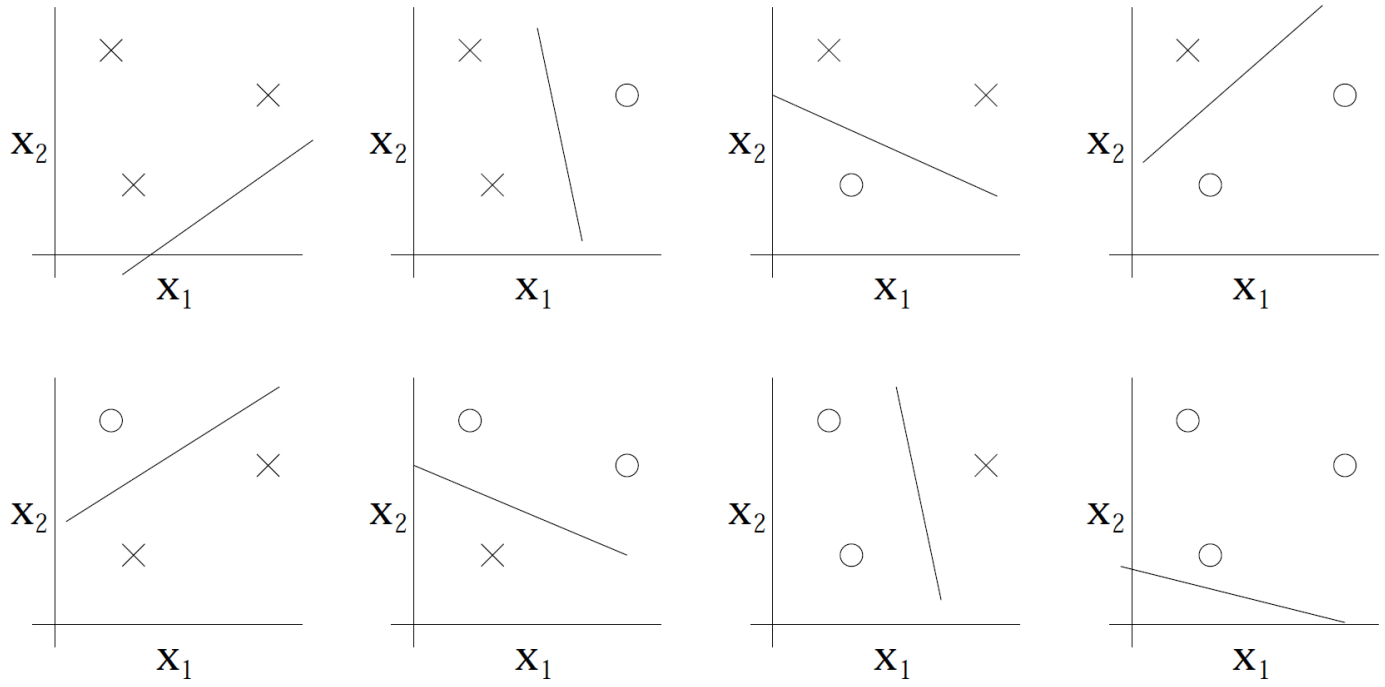
- 在实例空间X上定义的假设空间H的Vapnik-Chervonenkis维数—VC(H)，是由H能打散的X的最大有限子集的大小。
  - 如果X的任意大的有限子集可以被打散，那么$VC(H) = \infty$
  - 如果存在至少一个大小为d的X的子集可以被打散，那么$VC(H) \geq d$
  - 如果没有大小为d的子集可打散，那么$VC(H) < d$
  - 要打散m个实例，我们需要$|H| \geq 2^m$，因此

$$VC(H) = m \leq \log_2 |H|$$

Alexey Chervonenkis

VC维示例

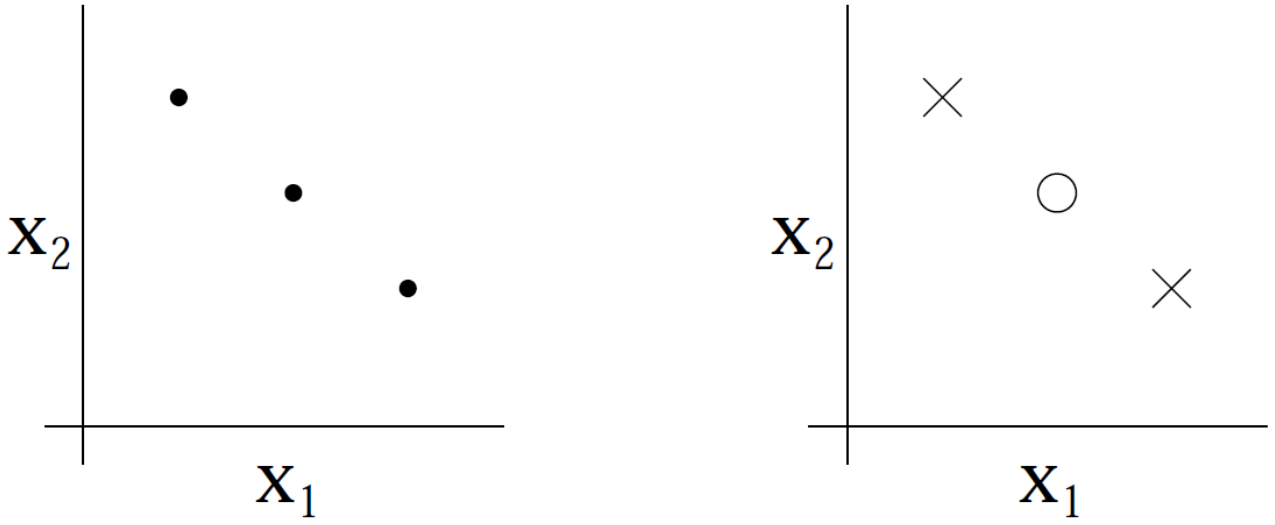
- 考虑实平面中的线性模型
  - 有的 3 个实例可以被打散



所有 8 个可能的标签可以分开

VC维示例

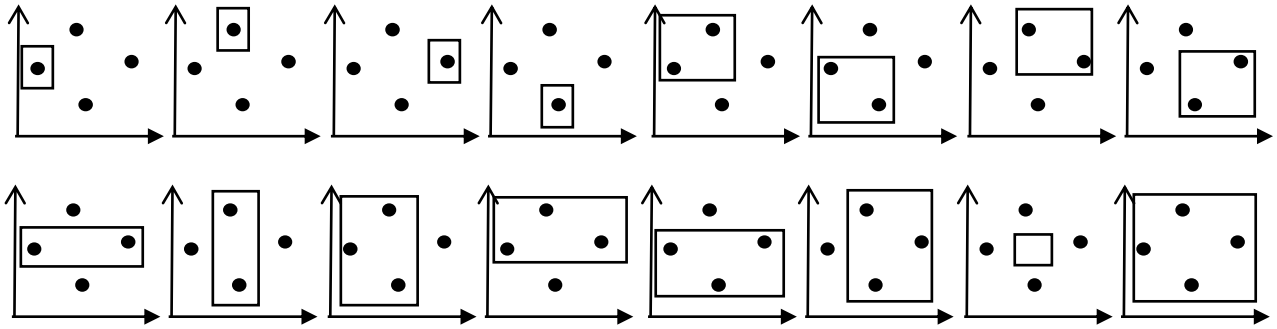
- 有的 3 个实例用一条直线不能打散



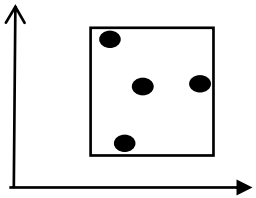
- 由于我们可以找到一个由线性模型打散的3个实例集合，因此线性模型的VC维至少为3

VC维示例

- 考虑实平面上的轴平行矩形，即两个实值特征上的区间连接
  - 有的4个实例可以被打散

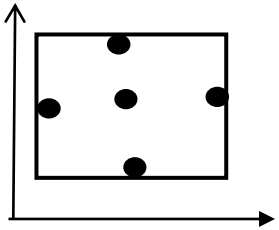


  - 有的4个实例不能被打散



VC维示例

- 没有哪种5个实例可以被打散，因为最多可以有4个不同的极端点(每个维度上的最小值和最大值)，如果不包含任何可能的第5个点，这4个实例就不能被包含



- 因此$VC(H) = 4$
- 推广到轴向平行超矩形(n维区间连接)：$VC(H) = 2n$

使用VC求得样本复杂度的上界

- 使用VC维作为表达性的度量，下面的例子已经被证明足以用于PAC学习(Blumer et al., 1989)

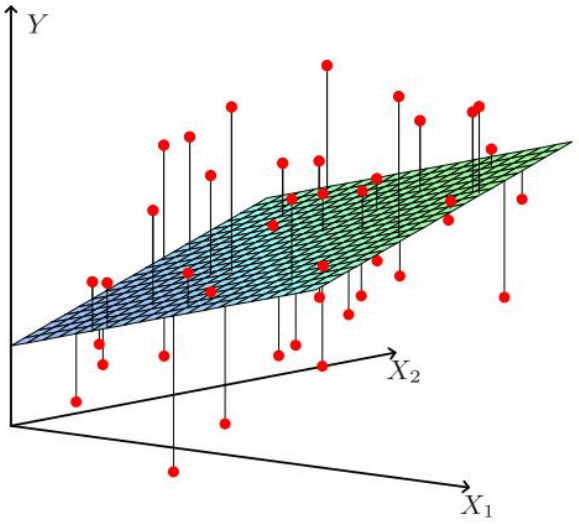
$$N = \frac{1}{\epsilon}\left(4\log_2\left(\frac{2}{\delta}\right) + 8\mathrm{VC}(H)\log_2\left(\frac{13}{\epsilon}\right)\right)$$

- 相比以前使用$\log|H|$的结果，这个边界有一些额外的常量和一个额外的$\log_2(1/\varepsilon)$因子
- 由于$VC(H) \leq \log_2|H|$，而且常是$VC(H) \ll \log_2|H|$，这可以为PAC学习所需的样本数提供一个更严格的上限

$$N = \frac{1}{2\epsilon^2}\left(\log|H| + \log\frac{1}{\delta}\right)$$

一些VC维的例子

- d维超平面的VC维是$d+1$
  - 超平面的VC维巧合地与定义超平面所需的参数数几乎相同

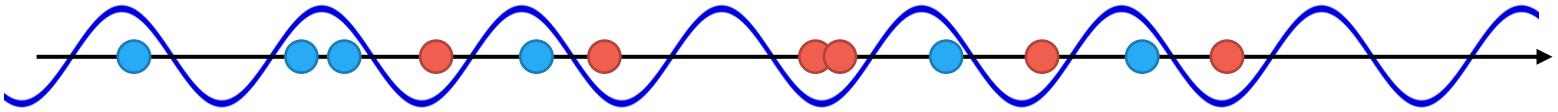


$$f(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

一些VC维的例子

- 正弦波具有无限的VC维，但只有两个参数
  - 通过仔细选择相位和周期，我们可以打碎任意一组一维数据点

$$h(x) = sin(ax + b)$$



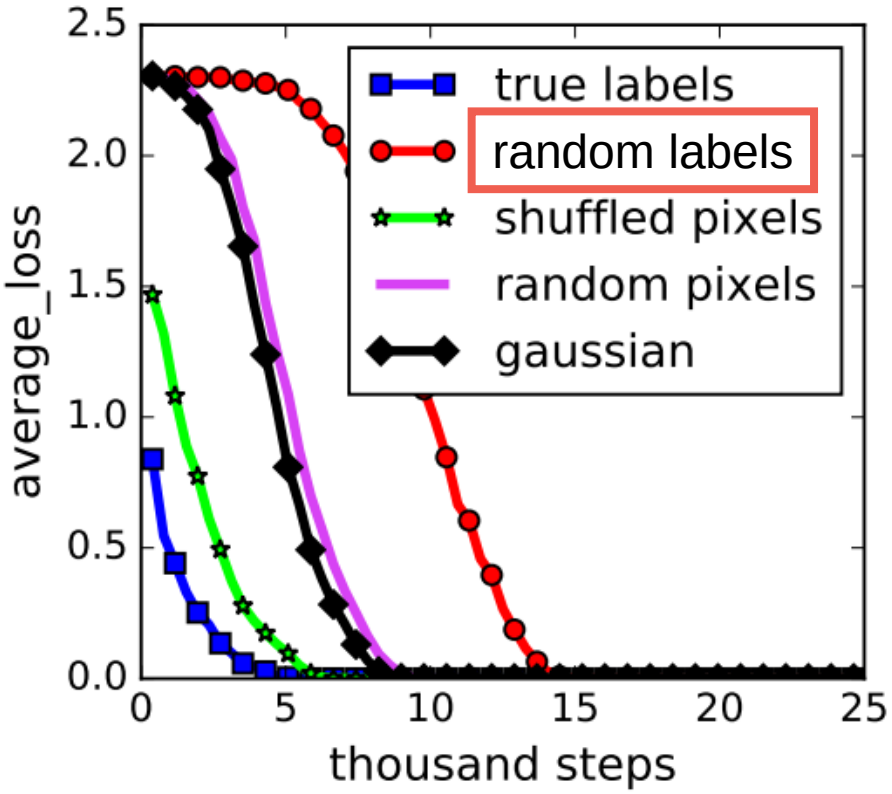
http://mlweb.loria.fr/book/en/VCdiminfinite.html

一些VC维的例子

- 具有某些类型激活函数的神经网络也具有无限的VC维

- 数据集:CIFAR-10
  - 50,000个训练图像
  - 网络：初始模型

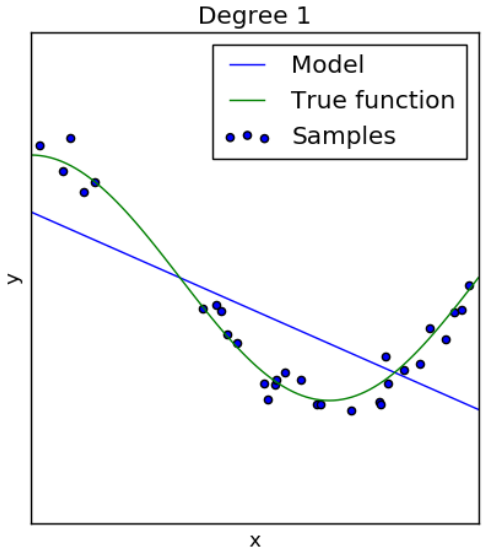
- 在随机标签上，MLP也收敛到零训练损失



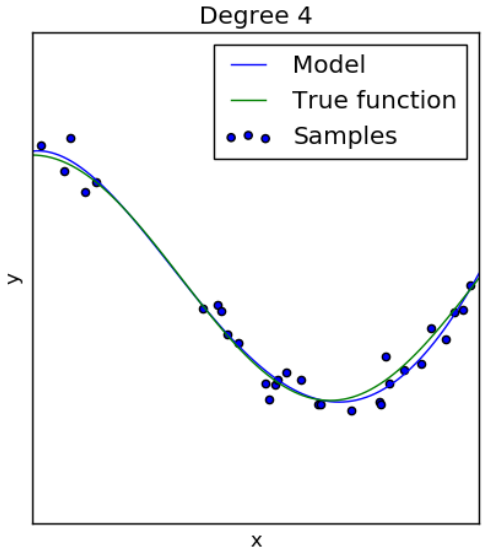
Zhang, Chiyuan, et al. "Understanding deep learning requires rethinking generalization." ICLR 2017.

模型选择

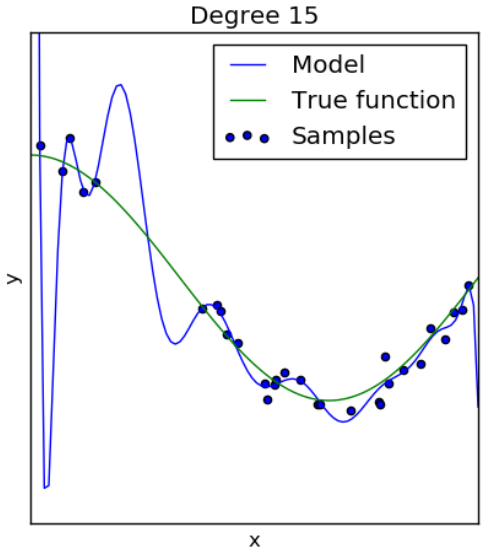
- 下面哪个模型是最好的?



线性模型：欠拟合



四阶模型：合适



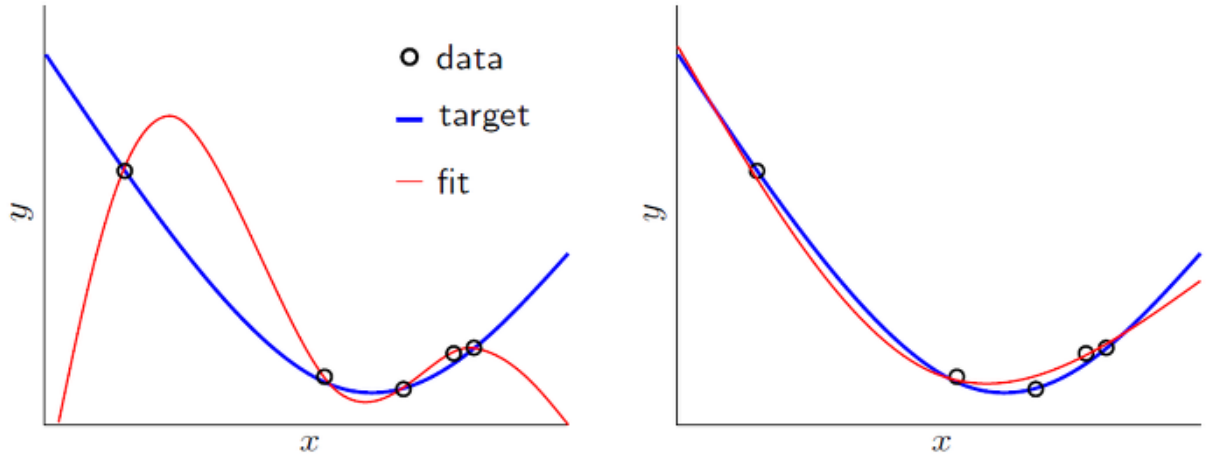
十五阶模型：过拟合

- 当统计模型或机器学习算法无法捕捉数据的底层变化趋势时，就会出现欠拟合。
- 当统计模型把随机误差和噪声也考虑进去，而不仅仅是考虑数据的基础关联时，就会出现过拟合。

正则化

- 添加参数的惩罚项，防止模型对数据过拟合

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}\left(y_i, f_\theta(x_i)\right) + \lambda \Omega(\theta)$$



(a) without regularization　　(b) with regularization

模型选择：
偏差-方差分解

张伟楠 – 上海交通大学

偏差-方差分解

- 偏差-方差分解
  - 假定$Y = f(X) + \epsilon$，其中$\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$
  - 输入点x_0预期预测误差为

$$
\begin{aligned}
\mathrm{Err}(x_0) &= \mathbb{E}[(Y - \hat{f}(X))^2 | X = x_0] \\
&= \mathbb{E}[(\epsilon + f(x_0) - \hat{f}(x_0))^2] \\
&= \mathbb{E}[\epsilon^2] + \underbrace{\mathbb{E}[2\epsilon(f(x_0) - \hat{f}(x_0))]}_{=0} + \mathbb{E}[(f(x_0) - \hat{f}(x_0))^2] \\
&= \sigma_\epsilon^2 + \mathbb{E}[(f(x_0) - \mathbb{E}[\hat{f}(x_0)] + \mathbb{E}[\hat{f}(x_0)] - \hat{f}(x_0))^2] \\
&= \sigma_\epsilon^2 + \mathbb{E}[(f(x_0) - \mathbb{E}[\hat{f}(x_0)])^2] + \mathbb{E}[(\mathbb{E}[\hat{f}(x_0)] - \hat{f}(x_0))^2] \\
&\quad - 2\mathbb{E}[(f(x_0) - \mathbb{E}[\hat{f}(x_0)])(\mathbb{E}[\hat{f}(x_0)] - \hat{f}(x_0))] \\
&= \sigma_\epsilon^2 + \mathbb{E}[(f(x_0) - \mathbb{E}[\hat{f}(x_0)])^2] + \mathbb{E}[(\mathbb{E}[\hat{f}(x_0)] - \hat{f}(x_0))^2] \\
&\quad - 2\underbrace{(f(x_0)\mathbb{E}[\hat{f}(x_0)] - f(x_0)\mathbb{E}[\hat{f}(x_0)] - \mathbb{E}[\hat{f}(x_0)]^2 + \mathbb{E}[\hat{f}(x_0)]^2)}_{=0} \\
&= \sigma_\epsilon^2 + (\mathbb{E}[\hat{f}(x_0)] - f(x_0))^2 + \mathbb{E}[(\hat{f}(x_0) - \mathbb{E}[\hat{f}(x_0)])^2] \\
&= \sigma_\epsilon^2 + \mathrm{Bias}^2(\hat{f}(x_0)) + \mathrm{Var}(\hat{f}(x_0))
\end{aligned}
$$

The expectation ranges over different choices of the training set all sampled from the same joint distribution $p(X, Y)$

偏差-方差分解

- 偏差-方差分解
  - 假定$Y = f(X) + \epsilon$，其中$\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$
  - 输入点x_0预期预测误差为

$$\begin{aligned}\mathrm{Err}(x_0) &= \sigma_\epsilon^2 + (\mathbb{E}[\hat{f}(x_0)] - f(x_0))^2 + \mathbb{E}[(\hat{f}(x_0) - \mathbb{E}[\hat{f}(x_0)])^2] \\ &= \sigma_\epsilon^2 + \mathrm{Bias}^2(\hat{f}(x_0)) + \mathrm{Var}(\hat{f}(x_0))\end{aligned}$$

观测噪声（不可约误差）

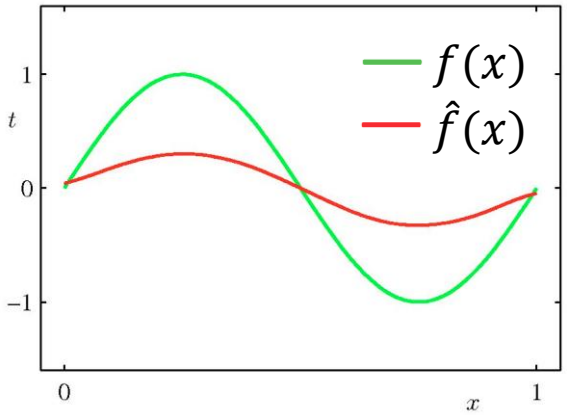
预期预测和真实值的差值

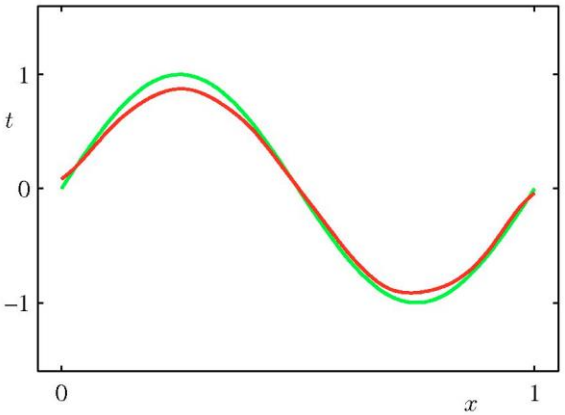
预测的不确定（给定不同的训练设置，例如数据和初始化）

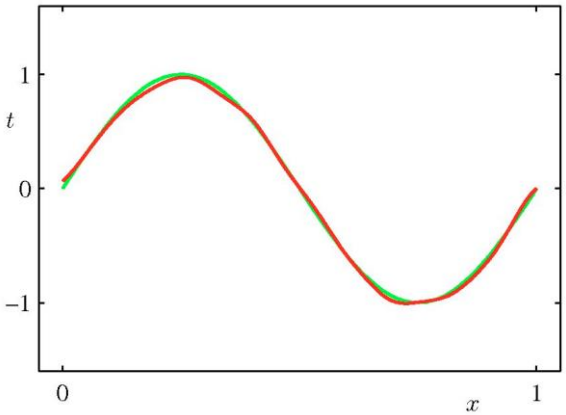
- 这其中，$\hat{f}(x_0)$计算的随机性包括：
  - 有限采样训练集的随机性
  - 标签y的随机性
  - 训练过程的随机性

偏差-方差插图

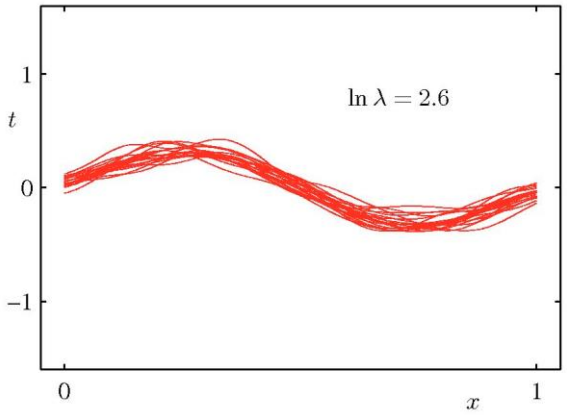
高 ← 偏差 → 低

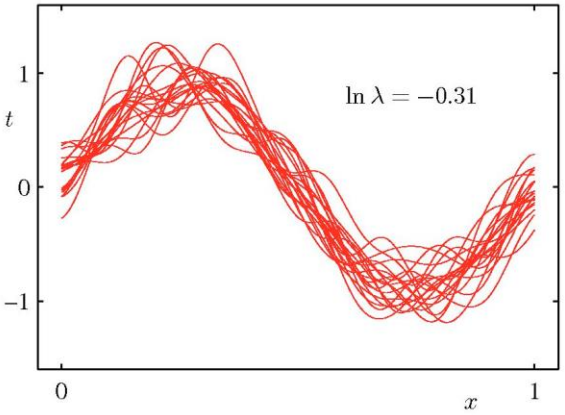


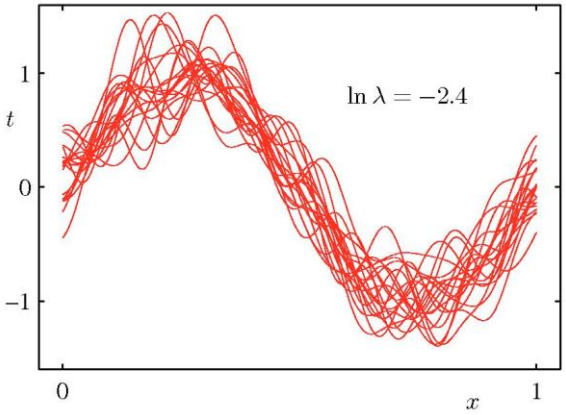




高 ← 正则化 → 低



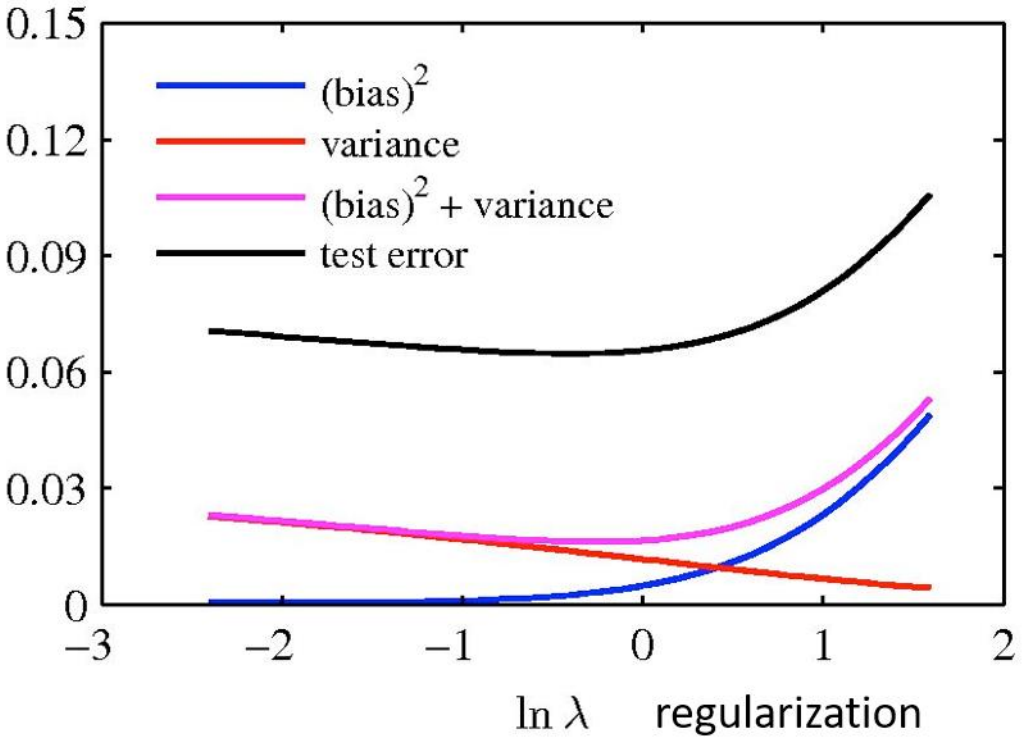




低 ← 方差 → 高

Figures provided by Max Welling

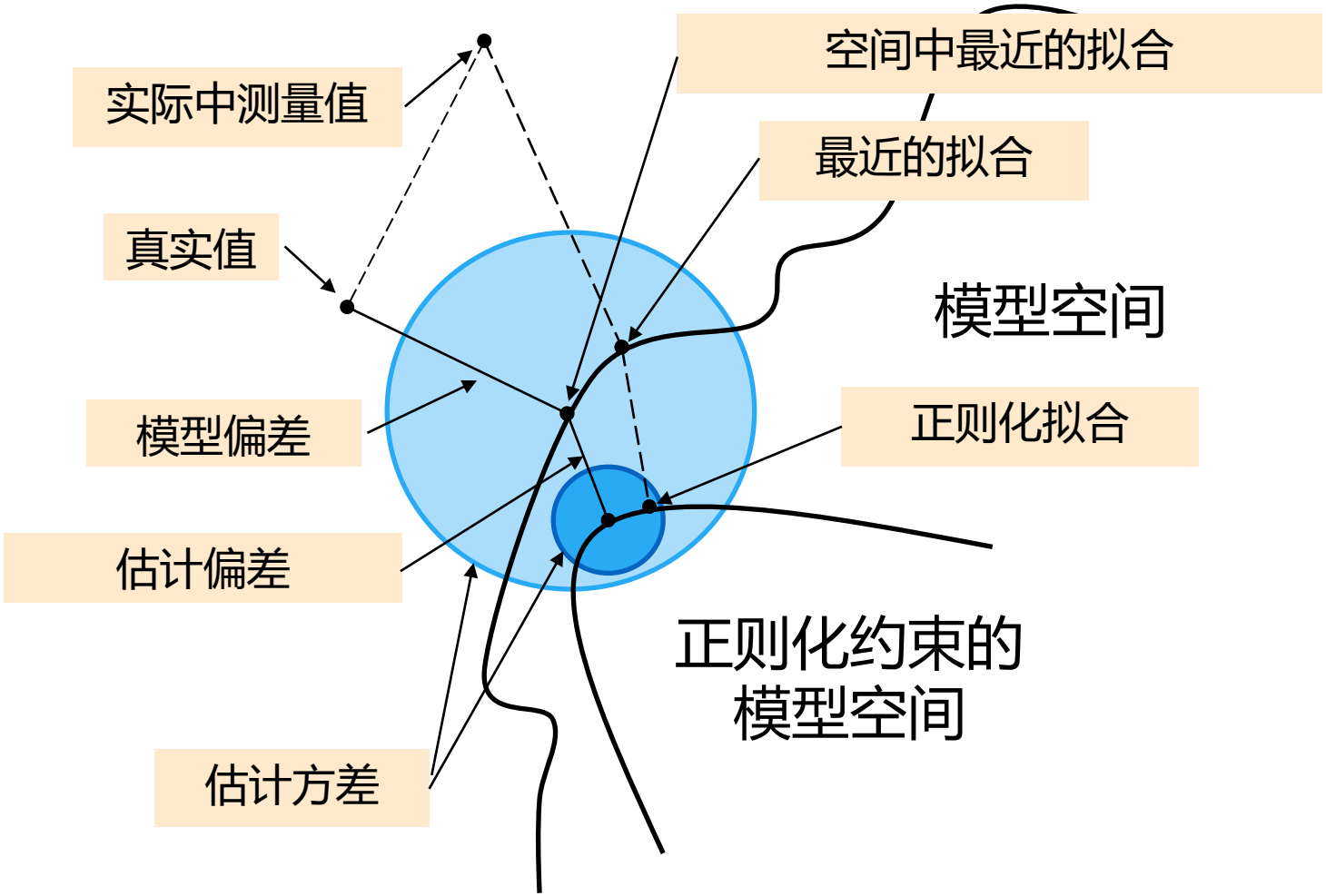
偏差-方差插图



- 训练误差测量偏差bias，但忽略方差variance
- 测试误差/交叉验证误差测量偏差bias和方差variance

偏差-方差分解

- 偏差和方差行为的原理图

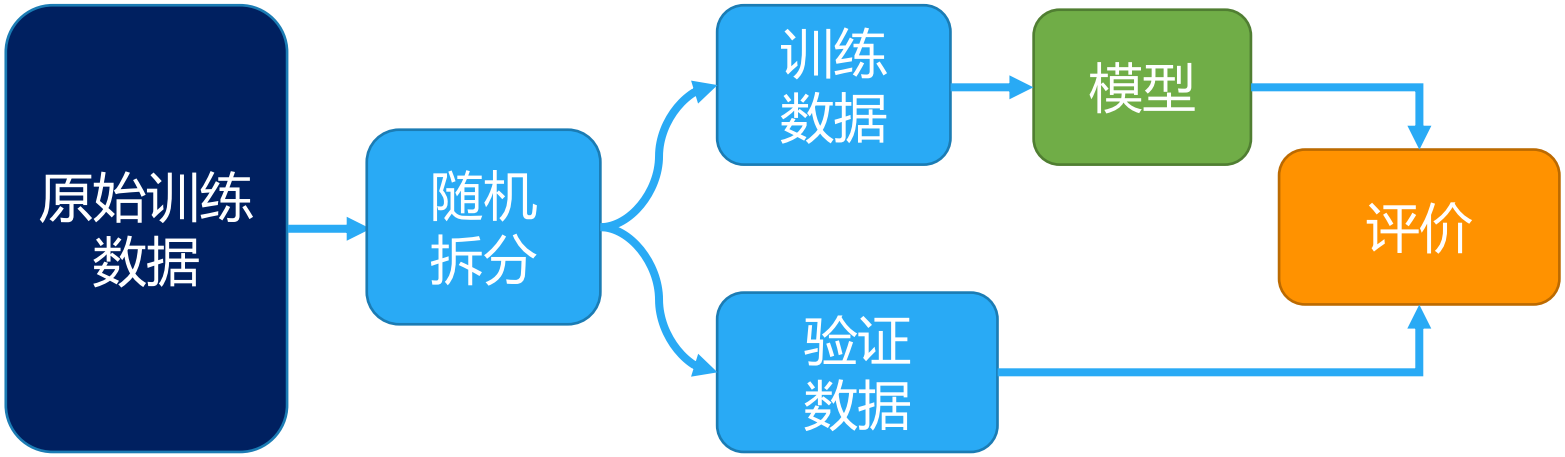


Slide credit Liqing Zhang

交叉验证

张伟楠 – 上海交通大学

交叉验证

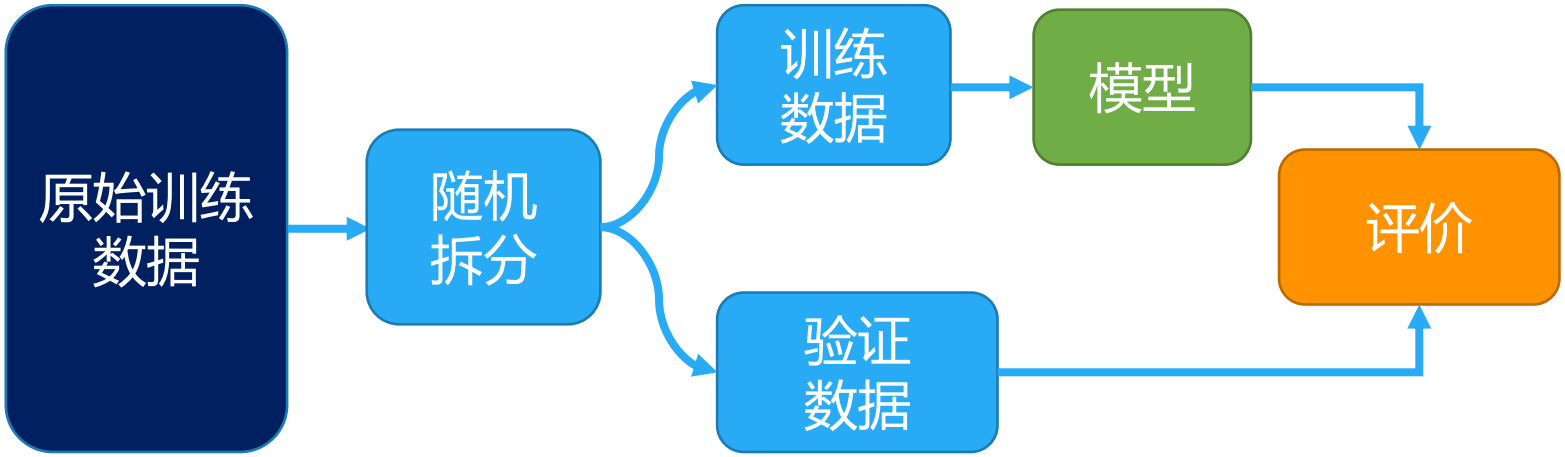


例如，5-fold交叉验证：

- 将数据拆分为5份

- 交叉验证1：在1,2,3,4份数据上训练模型，在第5份数据上进行验证
- 交叉验证2：在2,3,4,5份数据上训练模型，在第1份数据上进行验证
- …

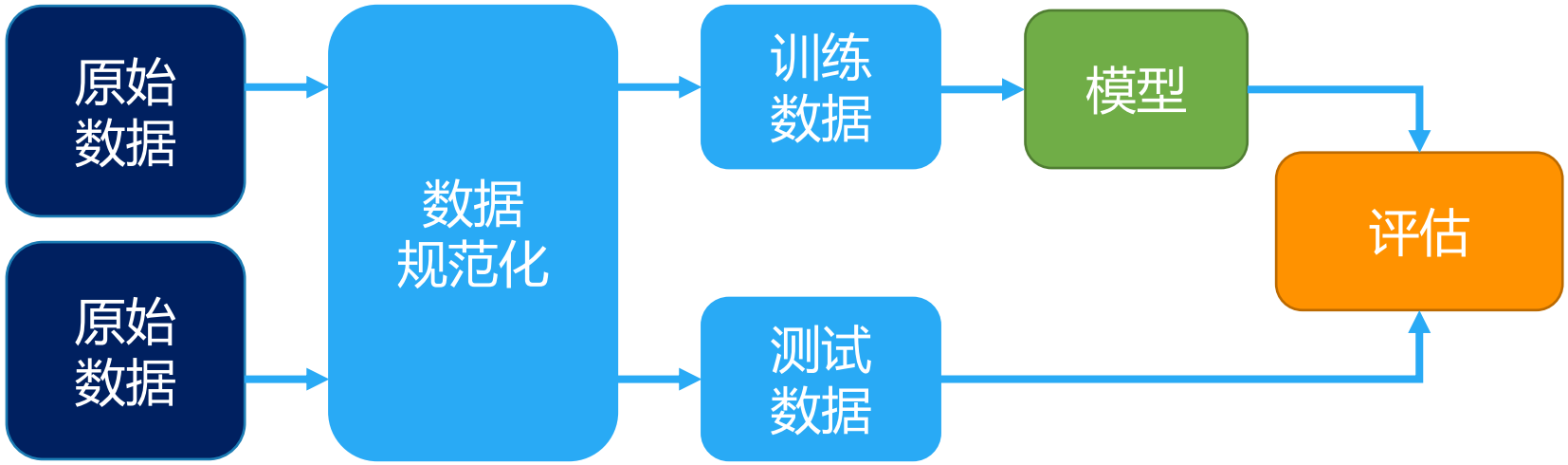
交叉验证



K-fold交叉验证

1. 设置超参数
2. 将原始训练数据随机拆分为K份
3. 重复K次:
   - 若当前为第i次重复（i=1,...,K），选择第i份数据作为验证数据集，其余K-1份作为训练数据集
   - 对训练数据进行建模，并在验证数据上对其进行评估，从而获得评估分数
4. 对K个评估分数取平均作为模型性能

机器学习过程

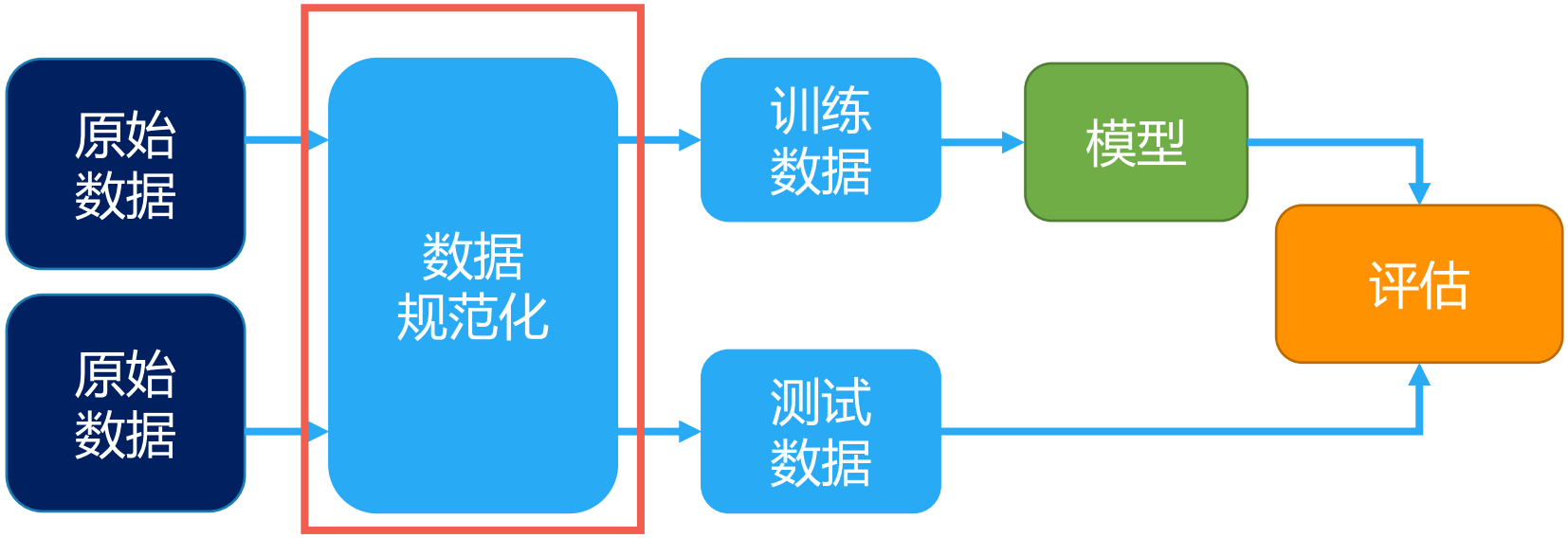


1. 选择了“好的”超参数
2. 对整个训练数据训练模型
3. 用测试数据对模型进行测试

特征选择

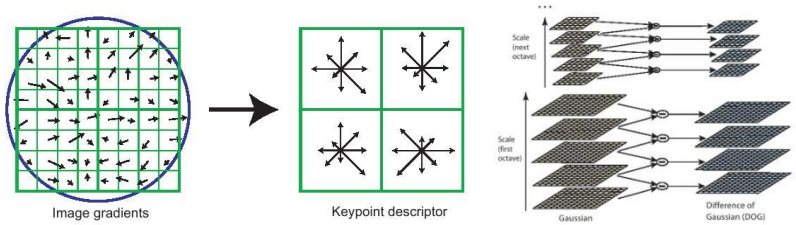
张伟楠 – 上海交通大学

数据表示

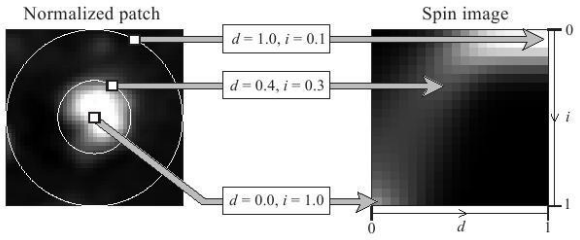


- 将数据规范为特征表示
  - 如何选择一些“好的”特征来改进模型的表现（即提高模型的泛化能力）？

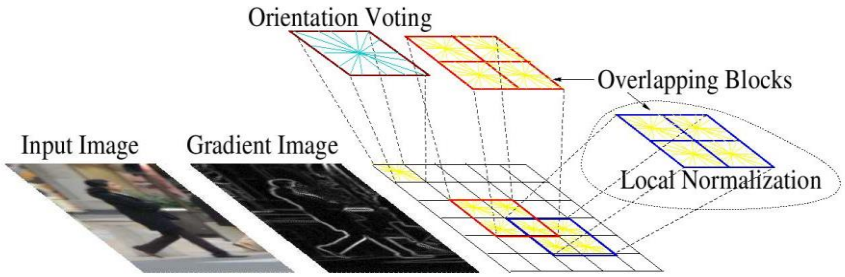
计算机视觉中的特征



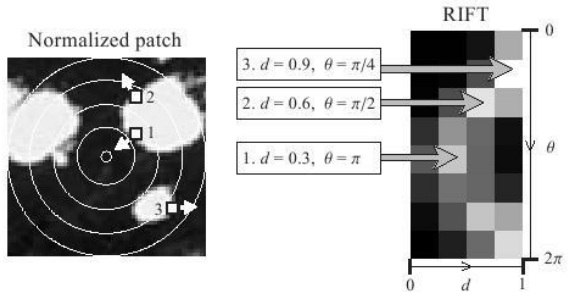
SIFT



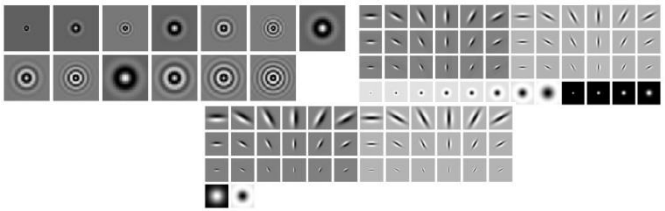
Spin image



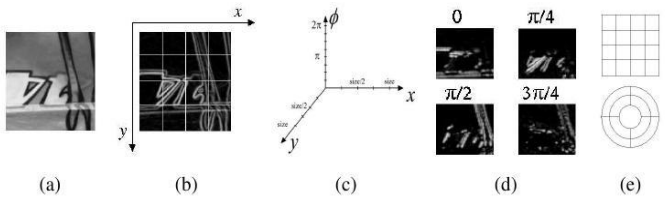
HoG



RIFT



Textons



GLOH

文本分类中的特征

- 输入文本

  SJTU is a public research university in Shanghai, China, established in 1896. Now it is one of C9 universities in China.

- 词袋表示

  SJTU:1, is:2, a:1, public:1, research:1, university:2, in:3, Shanghai:1, China:2, establish:1, 1896:1, now:1, it:1, one:1, of:1, C9:1

- 词汇量大小超过10万

特征选择

- 各种各样的特征表示使得每个数据实例被规范化为一个高维向量
  - 为了使模型可靠（即泛化误差小），需要大量的训练数据

$$N \geq \frac{1}{2\epsilon^2}(64m + \log\frac{1}{\delta}) = O_{\epsilon,\delta}(m)$$

- 泛化误差可以分解为

$$\mathrm{Err}(x_0) = \sigma_\epsilon^2 + \mathrm{Bias}^2(\hat{f}(x_0)) + \mathrm{Var}(\hat{f}(x_0))$$

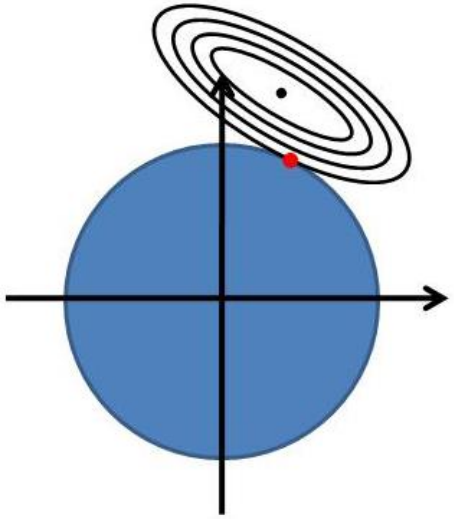
  - 特征太少会增加模型偏差
  - 特征太多会增加模型方差
  - 特征选择：偏差和方差间的权衡

正则化的特征选择作用

- L2正则化（岭回归Ridge）

$$\Omega(\theta) = \| \theta \|_2^2 = \sum_{m=1}^{M} \theta_m^2$$

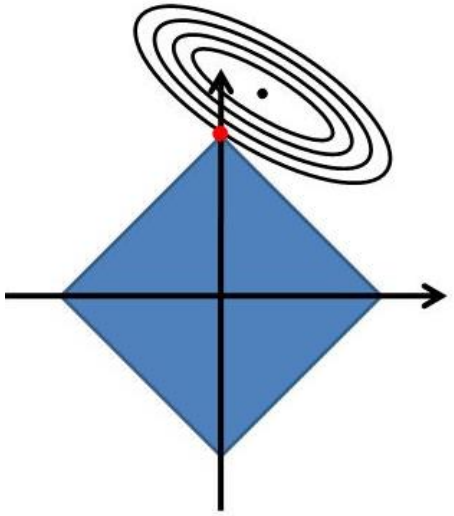
$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}\left(y_i, f_\theta(x_i)\right) + \lambda \| \theta \|_2^2$$



- L1正则化（套索LASSO）

$$\Omega(\theta) = \| \theta \|_1 = \sum_{m=1}^{M} |\theta_m|$$

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}\left(y_i, f_\theta(x_i)\right) + \lambda \| \theta \|_1$$



特征选择方法

- 无监督

	线性	非线性
选择	输入间的相关性	输入间的互信息
投影	主成分分析	流形学习，自组织映射

- 有监督

	线性	非线性
选择	输入和目标间的相关性	输入和目标间的互信息，贪心选择，遗传算法
投影	线性判别分析，偏最小二乘	多层感知机，自编码器，投影追踪

案例分析：特征选择方法学习

- 论文：

A Comparative Study on Feature Selection in Text Categorization

Yiming Yang
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213-3702, USA
yiming@cs.cmu.edu

Jan O. Pedersen
Verity, Inc.
894 Ross Dr.
Sunnyvale, CA 94089, USA
jpederse@verity.com

- 学习任务：文本分类
  - 特征：词袋，每一维表示一个词
  - 数据实例：词的文件
  - 目标：判断文本属于m个类别中的哪一类

Yang, Yiming, and Jan O. Pedersen. "A comparative study on feature selection in text categorization." *ICML*. Vol. 97. 1997.

案例分析：特征选择方法

- 文件频率（Document Frequency，DF）
  - 即包含该特征的文件的数量
  - 选择高DF的特征
    - 假设：低频特征要么包含很少的信息，要么对整体表现不会产生太大的影响

- 信息增益（Information Gain，IG）
  - IG 表示了加入该特征后，为目标预测获得的信息

$$G(t) = -\sum_{i=1}^{m} P(c_i)\log P(c_i)$$
$$+P(t)\sum_{i=1}^{m} P(c_i|t)\log P(c_i|t) + P(\bar{t})\sum_{i=1}^{m} P(c_i|\bar{t})\log P(c_i|\bar{t})$$

Yang, Yiming, and Jan O. Pedersen. "A comparative study on feature selection in text categorization." *ICML*. Vol. 97. 1997.

案例分析：特征选择方法

- 互信息（Mutual Information, MI）
  - 两个随机变量间的MI衡量了两个变量间的相互依赖性

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} \log \frac{p(x,y)}{p(x)p(y)}$$

  - 特征t和目标c(作为两个独立变量)的互信息可以表示为

$$I(t,c) = \log \frac{P(t,c)}{P(t)P(c)} \simeq \log \frac{A \times N}{(A+C) \times (A+B)}$$

    - A：t和c同时出现的文件数
    - B：t出现c不出现的文件数
    - C：t不出现c出现的文件数
    - N：总文件数

Yang, Yiming, and Jan O. Pedersen. "A comparative study on feature selection in text categorization." *ICML*. Vol. 97. 1997.

案例分析：特征选择方法

- 互信息（Mutual Information, MI）
  - 两个随机变量间的MI衡量了两个变量间的相互依赖性

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} \log \frac{p(x,y)}{p(x)p(y)}$$

  - 特征t和目标c(作为两个独立变量)的互信息可以表示为

$$I(t,c) = \log \frac{P(t,c)}{P(t)P(c)} \simeq \log \frac{A \times N}{(A+C) \times (A+B)}$$

  - 测量特征好坏的两种方法

$$I_{\text{avg}}(t) = \sum_{i=1}^{m} P(c_i) I(t,c_i) \qquad I_{\max}(t) = \max_{i=1}^{m} \{I(t,c_i)\}$$

Yang, Yiming, and Jan O. Pedersen. "A comparative study on feature selection in text categorization." *ICML*. Vol. 97. 1997.

案例分析：特征选择方法

- χ^2统计（CHI）
  - 衡量特征t和目标c之间独立性的缺乏

$$\chi^2(t,c) = \frac{N \times (AD - CB)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)}$$

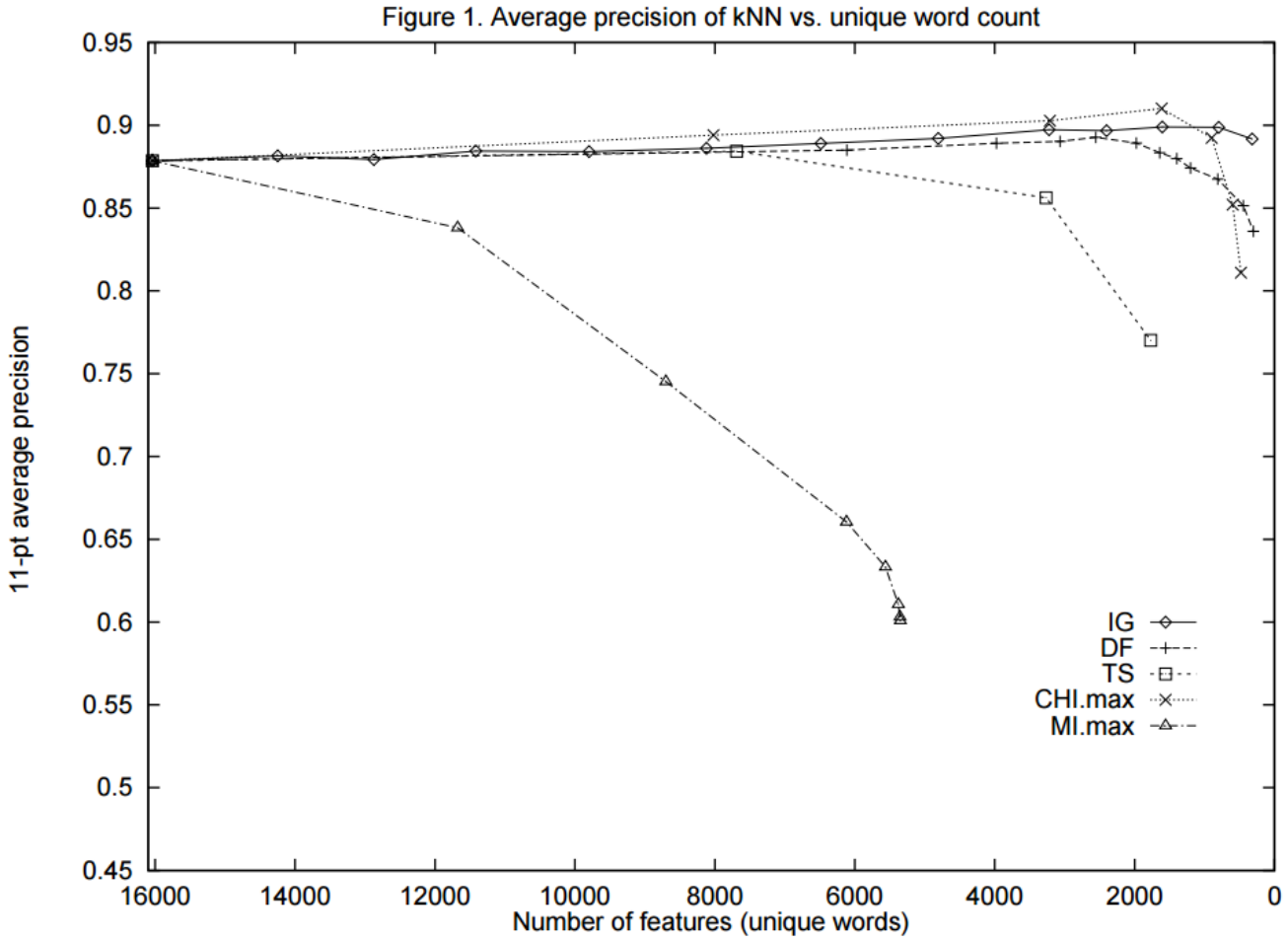
  - A：t和c同时出现的文件数
  - B：t出现c不出现的文件数
  - C：t不出现c出现的文件数
  - D：t和c都不出现的文件数
  - N：总的文件数

  - 测量特征好坏的两种方法

$$I_{\text{avg}}(t) = \sum_{i=1}^{m} P(c_i)\chi^2(t,c_i) \qquad I_{\max}(t) = \max_{i=1}^{m}\{\chi^2(t,c_i)\}$$

Yang, Yiming, and Jan O. Pedersen. "A comparative study on feature selection in text categorization." *ICML*. Vol. 97. 1997.

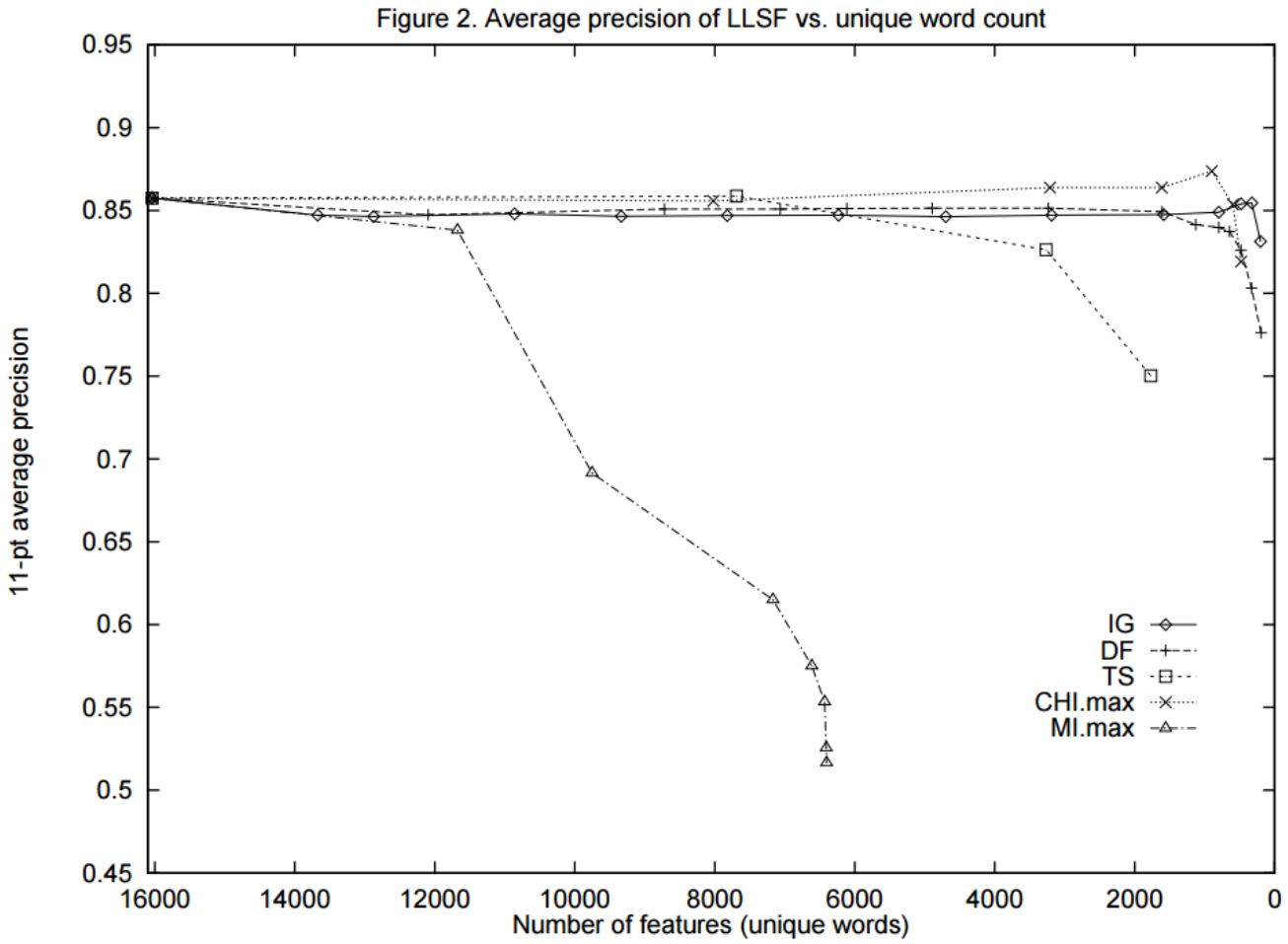
案例分析：实际表现



KNN在Reuters数据集上的表现：9610个训练文档，3662个测试文档

Yang, Yiming, and Jan O. Pedersen. "A comparative study on feature selection in text categorization." *ICML*. Vol. 97. 1997.

案例分析：实际表现



线性模型在Reuters数据集上的表现：9610个训练文档，3662个测试文档

Yang, Yiming, and Jan O. Pedersen. "A comparative study on feature selection in text categorization." *ICML*. Vol. 97. 1997.

总结学习理论和模型选择

- 有监督机器学习一般的流程
  1. 从数据分布$p(x,y)$采样获得有限训练数据集D_{train}
  2. 在D上训练模型$f(x)$用于预测y
  3. 在相同数据分布$p(x,y)$新采样的数据集D_{test}上验证模型的精度

- 学习理论
  - 可学性：基于多项式大小的样本量，L 能从假设空间$\mathcal{H}$中PAC辨识概念类C，则称概念类C是PAC可学的
  - 在有限训练数据上看到的模型精度和在无限测试数据上得到的精度的期望之间的差距
  - 差距越小，模型训练约值得信赖
  - 该差距和模型的精度可能存在tradeoff

- 学习理论指导模型选择和实验过程
  - 偏差-方差的平衡
  - 交叉验证
  - 特征选择

THANK YOU