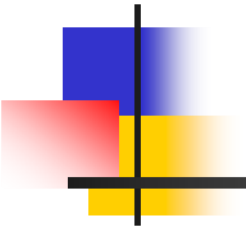
第三章

概率密度函数估计



- 实际应用中，未知的是概率密度函数和$p(x|\omega_i)$先验概率$p(\omega_i)$
- 为了使用贝叶斯决策方法，如何得到概率密度函数？

三个问题

- 如何利用样本集估计概率密度函数和先验概率？
- 估计量的性质如何？
- 如何利用样本集估计分类器的错误率?

概念

- 统计量 如：均值，方差
- 参数空间：未知参数向量的所有可能取值的集合
- 点估计、估计量、估计值
- 区间估计

最大似然估计

假设:

- 待估计的参数 θ 是一个确定的、未知的量
- 概率分布函数的形式已知

最大似然估计

似然函数

$$p(\chi \mid \theta) = p(x_1, x_2, \cdots, x_N \mid \theta) = \prod_{k=1}^{N} p(x_k \mid \theta) = l(\theta)$$

$$\frac{dl(\theta)}{d\theta} = 0$$

$$H(\theta) = \ln l(\theta) = \ln \prod_{k=1}^{N} p(x_k \mid \theta) = \sum_{k=1}^{N} \ln\, p(x_k \mid \theta)$$

$$\frac{dH(\theta)}{d\theta} = 0$$

一个特殊的例子

$$p(x|\theta)=\begin{cases}\dfrac{1}{\theta_2-\theta_1} & \theta_1<x<\theta_2\\ 0 & \text{其他}\end{cases}$$

$$l(\theta)=\begin{cases}1/(\theta_2-\theta_1)^N & \\ 0 & \end{cases}$$

$$H(\theta)=-N\cdot\ln(\theta_2-\theta_1)$$

$$\frac{\partial H}{\partial\theta_1}=N\cdot\frac{1}{\theta_2-\theta_1}$$

$$\frac{\partial H}{\partial\theta_2}=N\cdot\frac{-1}{\theta_2-\theta_1}$$

$$x' = \min\{ x_1, \cdots, x_N \}$$

$$x'' = \max\{ x_1, \cdots, x_N \}$$

$$\hat{\theta}_1 = x' \qquad \hat{\theta}_2 = x''$$

正态分布参数的最大似然估计

- 正态分布 $p(x|\theta)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

$p(x)\sim N(\mu,\sigma^2)$ $\theta=[\mu,\sigma^2]$

$\chi=\{x_1,\cdots,x_N\}\Rightarrow\hat{\mu},\hat{\sigma}^2$

$$\nabla_\theta H(\theta)=\sum_{k=1}^{N}\nabla_\theta \ln p(x_k|\theta)=0$$

正态分布参数的最大似然估计

$$\nabla_\theta H(\theta) = \sum_{k=1}^{N} \nabla_\theta \ln\, p(x_k \mid \theta) = 0 \qquad p(x \mid \theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\ln\, p(x_k \mid \theta) = -\frac{1}{2}\ln(2\pi\theta_2) - \frac{1}{2\theta_2}(x_k - \theta_1)^2$$

$$\nabla_\theta \ln\, p(x_k \mid \theta) = \begin{bmatrix} \dfrac{1}{\theta_2}(x_k - \theta_1) \\ -\dfrac{1}{2\theta_2} + \dfrac{(x_k - \theta_1)^2}{2\theta_2^2} \end{bmatrix}$$

$$\nabla_\theta \ln p(x_k \mid \theta) = \begin{bmatrix} \dfrac{1}{\theta_2}(x_k - \theta_1) \\ -\dfrac{1}{2\theta_2} + \dfrac{(x_k - \theta_1)^2}{2\theta_2^2} \end{bmatrix}$$

$$\Rightarrow \begin{cases} \displaystyle\sum_{k=1}^{N} \frac{1}{\hat{\theta}_2}(x_k - \hat{\theta}_1) = 0 \\ \displaystyle -\sum_{k=1}^{N} \frac{1}{\hat{\theta}_2} + \sum_{k=1}^{N} \frac{(x_k - \hat{\theta}_1)^2}{\theta_2^2} = 0 \end{cases}$$

$$\hat{\theta}_1 = \hat{\mu} = \frac{1}{N}\sum_{k=1}^{N} x_k \qquad \hat{\theta}_2 = \hat{\sigma}^2 = \frac{1}{N}\sum_{k=1}^{N} (x_k - \hat{\mu})^2$$

多元正态分布参数的最大似然估计结果

$$\hat{\theta}_1 = \hat{\mu} = \frac{1}{N}\sum_{k=1}^{N} x_k$$

$$\hat{\Sigma} = \frac{1}{N}\sum_{k=1}^{N} (x_k - \hat{\mu})(x_k - \hat{\mu})^T$$

- 一元情况

$$\hat{\theta}_1 = \hat{\mu} = \frac{1}{N}\sum_{k=1}^{N} x_k \qquad \hat{\theta}_2 = \hat{\sigma}^2 = \frac{1}{N}\sum_{k=1}^{N} (x_k - \hat{\mu})^2$$

估计的性质

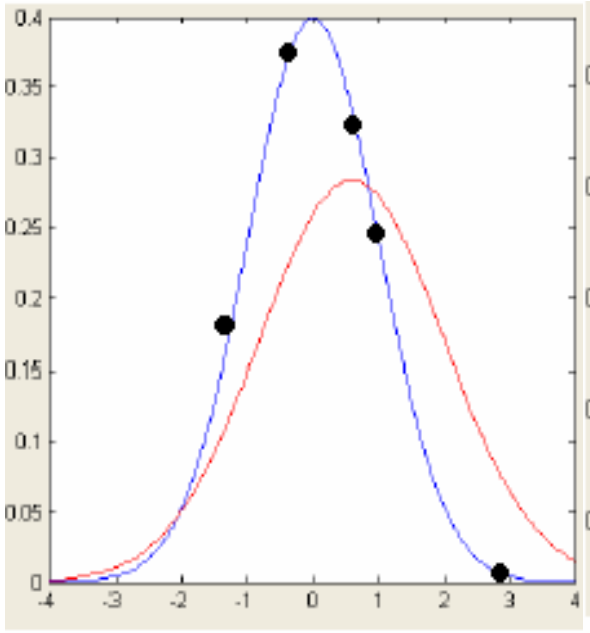
- 无偏估计
- 渐进无偏估计
- 有偏估计

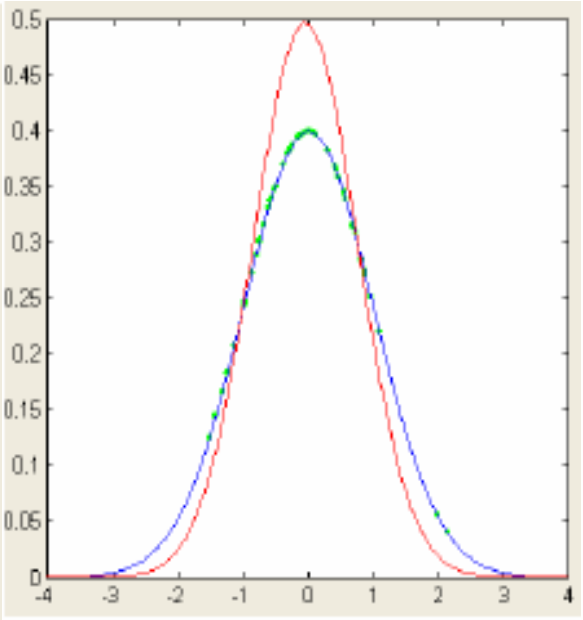
$$E\hat{\theta} = \theta$$

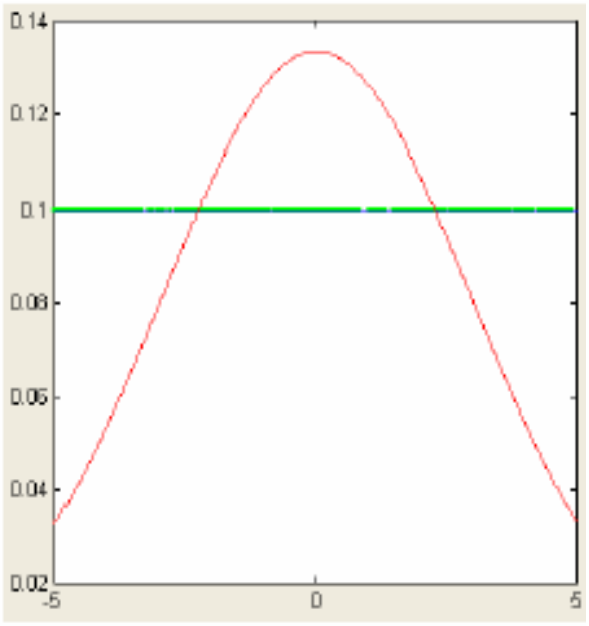
$$\lim_{N \to \infty} E\hat{\theta} = \theta$$

实验

- 从正态分布采样5、50个样本点，然后估计该分布的参数
- 从均匀分布采样500个点，估计一个正态分布的参数







应用举例:背景建模

- 为每一个像素建模为一个正态分布

一个不可识别的例子

x 是取值为0，1的离散变量。

$$p(x|\theta)=\frac{1}{2}\theta_1^x(1-\theta_1)^{1-x}+\frac{1}{2}\theta_2^x(1-\theta_2)^{1-x}$$

$$=\begin{cases}\frac{1}{2}(\theta_1+\theta_2) & x=1\\ 1-\frac{1}{2}(\theta_1+\theta_2) & x=0\end{cases}$$

可识别性问题

- 对于参数空间的两个取值 θ 和 θ'，存在一个 x 有：

$$p(x|\theta) \neq p(x|\theta')$$

则称为可识别的。

可识别性问题

离散随机变量的混合密度函数，往往是不可识别的。

大部分连续随机变量的混合密度函数，往往是可识别的。

贝叶斯估计

假设

- 参数 θ 是一个随机变量
- 已知参数 θ 的先验分布
- 首先确定参数的先验分布，然后计算参数的似然，最后计算参数的后验概率

$$p(\theta \mid x) = \frac{p(x \mid \theta)\, p(\theta)}{\int p(x \mid \theta)\, p(\theta) d\theta} = \frac{p(x \mid \theta)\, p(\theta)}{p(x)}$$

正态分布参数的贝叶斯估计

$p(x|\mu) \sim N(\mu, \sigma^2)$

$p(\mu) \sim N(\mu_0, \sigma_0^2)$

- 表示了对于均值的不确定性

$$p(\mu \mid \chi) = \frac{p(\chi \mid \mu)\, p(\mu)}{\int p(\chi \mid \mu)\, p(\mu) d\mu}$$

$$= \alpha \prod_{k=1}^{N} p(x_k \mid \mu)\, p(\mu)$$

$$= \alpha \prod_{k=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x_k - \mu}{\sigma}\right)^2} \cdot \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{1}{2}\left(\frac{\mu - \mu_0}{\sigma_0}\right)^2}$$

$$= \alpha' e^{-\frac{1}{2}\left[\sum_{k=1}^{N}\left(\frac{\mu - x_k}{\sigma}\right)^2 + \left(\frac{\mu - \mu_0}{\sigma_0}\right)^2\right]}$$

$$= \alpha' e^{-\frac{1}{2}\left[\sum_{k=1}^{N}\left(\frac{\mu - x_k}{\sigma}\right)^2 + \left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right]}$$

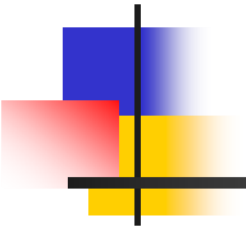
$$= \alpha'' e^{-\frac{1}{2}\left[\left(\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\frac{1}{\sigma^2}\sum_{k=1}^{N} x_k + \frac{\mu_0}{\sigma_0^2}\right)\mu\right]}$$

$$= \frac{1}{\sqrt{2\pi}\sigma_N} e^{-\frac{1}{2}\left(\frac{\mu-\mu_N}{\sigma_N}\right)^2}$$

$$\mu_N = \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2} m_N + \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 \qquad p(\mu \mid \chi) \sim N(\mu_N, \sigma_N^2)$$

$$m_N = \frac{1}{N}\sum_{k=1}^{N} x_k \qquad \sigma_N^2 = \frac{\sigma_0^2\sigma^2}{N\sigma_0^2+\sigma^2}$$

仍然是一个正态分布。



$$p(x \mid X) = \int p(x \mid \mu) p(\mu \mid X)\, d\mu$$

$$= \int \frac{1}{\sqrt{(2\pi)}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} \frac{1}{\sqrt{(2\pi)}\sigma_N} \exp\left\{-\frac{1}{2}\left(\frac{\mu-\mu_N}{\sigma_N}\right)^2\right\} d\mu$$

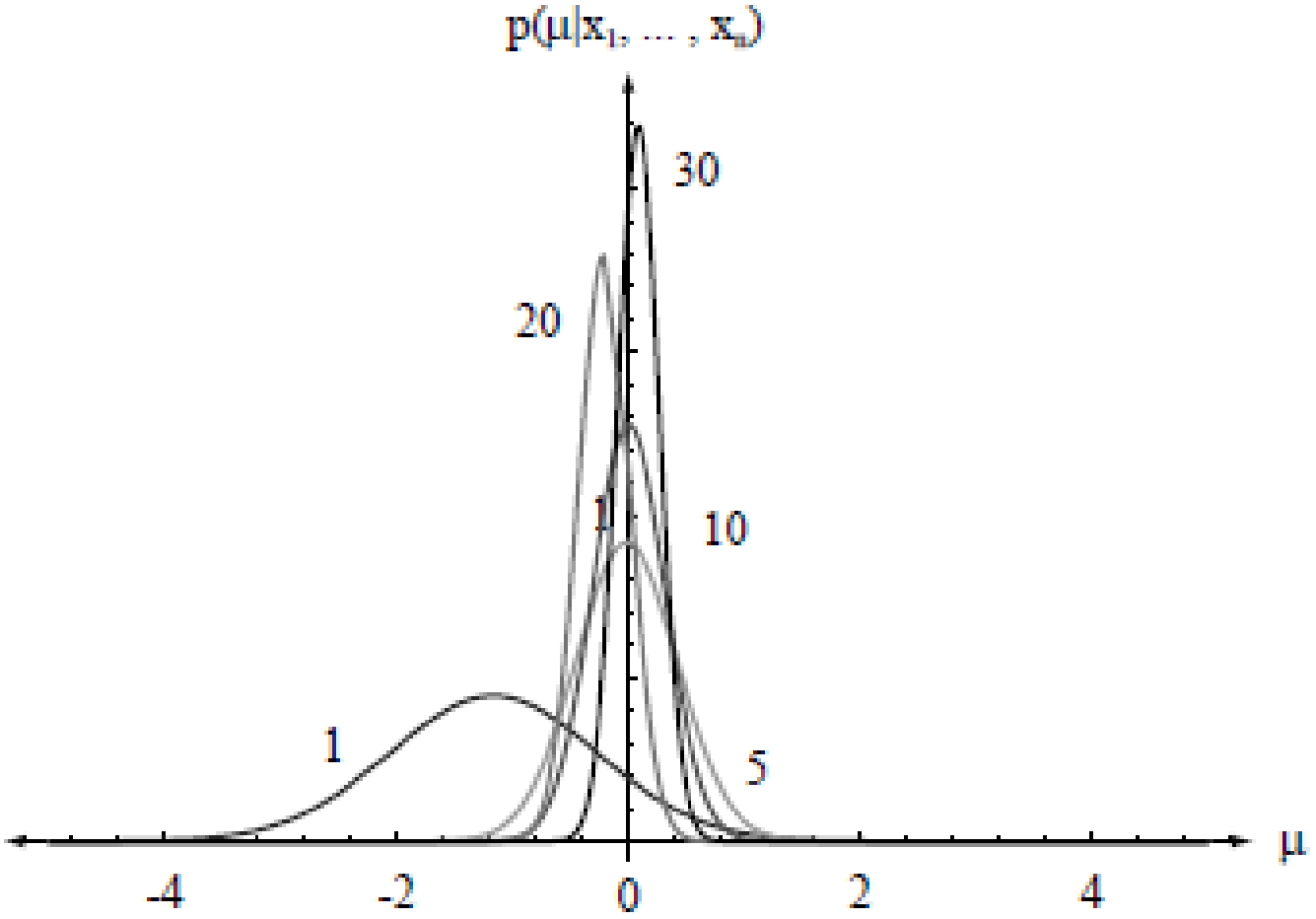
$$= \frac{1}{2\pi\sigma\sigma_N} \exp\left\{-\frac{1}{2}\frac{(x-\mu_N)^2}{\sigma^2+\sigma_N^2}\right\} f(\sigma, \sigma_N)$$

$$p(x \mid \chi) = \int p(x \mid \mu)\, p(\mu \mid \chi) d\mu$$

$$\sim N(\mu_N, \sigma^2 + \sigma_N^2)$$

- 样本数、先验对于估计结果的影响
- σ_N^2 引起了不确定性的增加。

对一维正态分布均值进行贝叶斯学习的过程



最大似然估计与贝叶斯估计

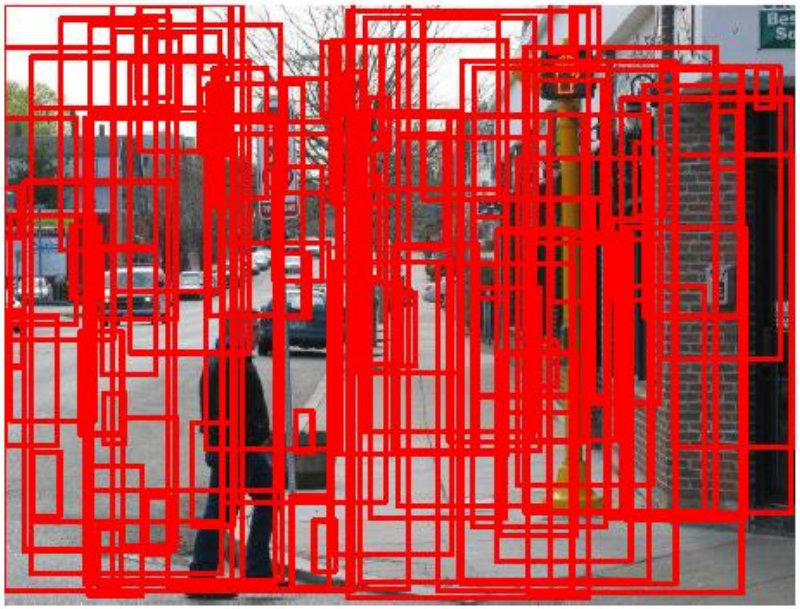
- 当训练样本数无穷多的时候，最大似然估计和贝叶斯估计的结果是一样的
- 贝叶斯估计由于使用了先验概率，利用了更多的信息
- 如果这些信息是可靠的，那么有理由认为贝叶斯估计比最大似然估计的结果更准确

最大似然估计与贝叶斯估计

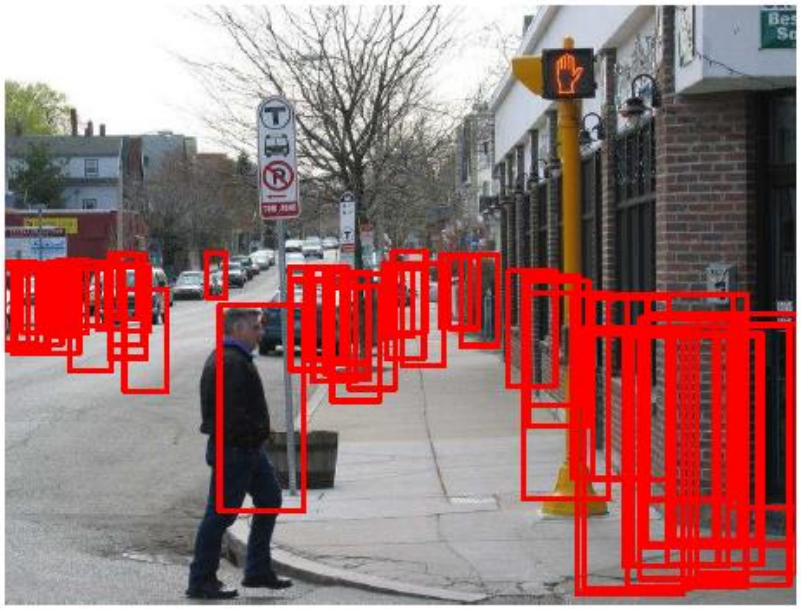
- 有时候先验概率很难设计
- 如：估计一个物体在一个区域出现的位置
- 在没有特别先验知识的时候，取先验概率是这个区域中的均匀分布:无信息先验
- 无信息先验: 最大似然估计结果和贝叶斯估计结果相似

举例

– 先验信息用于修剪检测输出



(b) P(person) = uniform



g) P(person|viewpoint,geometry)

Hoiem, Efros, Herbert, 2006

最大似然估计与贝叶斯估计

- 最大似然估计方法优点
- 计算简单。贝叶斯估计方法通常要计算复杂的积分。
- 易于理解。最大似然估计给出的是参数的一个最佳估计结果。

分布的非参数估计

假设：

- 不知道分布函数形式
- N 个样本服从独立同分布

  $x_1, x_2, \cdots, x_N \sim p(x)$
- P: 单个样本属于R的概率
- N 个样本中有k 个属于R的概率

$$P(k) = C_N^k P^k (1-P)^{N-k} \qquad C_N^k = \frac{N!}{k!(N-k)!}$$

$$E[k] = NP$$

- 众数(mode)： 数集合中出现频率最高的数

$$P_m = \max P_k$$

$$k = m \approx (N+1)\hat{P} \approx N\hat{P}$$

$$\hat{P} \approx \frac{k}{N}$$ 极小区域内的平均密度

$$P = \int_{\Re} p(x)dx = p(x)V$$

$$\frac{k}{N} \approx \hat{P} = \int_{\Re} \hat{p}(x)dx = \hat{p}(x)V$$

$$\hat{p}(x) = \frac{k}{NV}$$

构造：

$R_1, R_2, \cdots, R_N$

1　2　$\cdots$　N

$V_1, V_2, \cdots, V_N$

$k_1, k_2, \cdots, k_N$

构造包含x的序列 $R_1,\cdots,R_N$

$$\hat{p}_N(x)=\frac{k_N/N}{V_N}$$ 收敛至 $p(x)$ 当以下条件成立：

1) $\lim_{N\to\infty} V_N = 0$

2) $\lim_{N\to\infty} k_N = \infty$

3) $\lim_{N\to\infty} \frac{k_N}{N} = 0$

Parzen窗

- R_N 是 d 维的超立方体

$$V_N = h_N^d$$

$$\varphi(u) = \begin{cases} 1, & \text{if } |u_j| \le \frac{1}{2}, \forall j = 1, 2, \cdots, d \\ 0, & \text{其他} \end{cases}$$

- 以点x为中心，体积为 V_N 的区域包含的样本数为：

$$k_N = \sum_{i=1}^{N} \varphi\left(\frac{x - x_i}{h_N}\right)$$

$$\hat{p}_N(x) = \frac{k_N / N}{V_N}$$

$$\hat{p}_N(x) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{V_N}\varphi(\frac{x - x_i}{h_N})$$

其他的函数均可以，只要满足：

$$1)\varphi(u) \geq 0 \qquad 2)\int \varphi(u)du = 1$$

非负性显然

$$\hat{p}_N(x) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{V_N}\varphi(\frac{x-x_i}{h_N})$$

$$\int \hat{p}_N(x)dx = \int \frac{1}{N}\sum_{i=1}^{N}\frac{1}{V_N}\varphi(\frac{x-x_i}{h_N})dx$$

$$= \frac{1}{N}\sum_{i=1}^{N}\int \frac{1}{V_N}\varphi(\frac{x-x_i}{h_N})dx$$

$$= \frac{1}{N}\sum_{i=1}^{N}\int \varphi(u)du = \frac{1}{N}\cdot N = 1$$

其他的窗函数

- 方窗
- 正态窗
- 指数窗
- 。。。。

$\hat{p}_N(x)$ 估计量的性质

收敛的条件：

（1）总体密度函数 $p(x)$ 在 x 点连续。

（2）窗函数满足一下条件：

$$\varphi(u) \geq 0 \qquad \int \varphi(u)du = 1$$

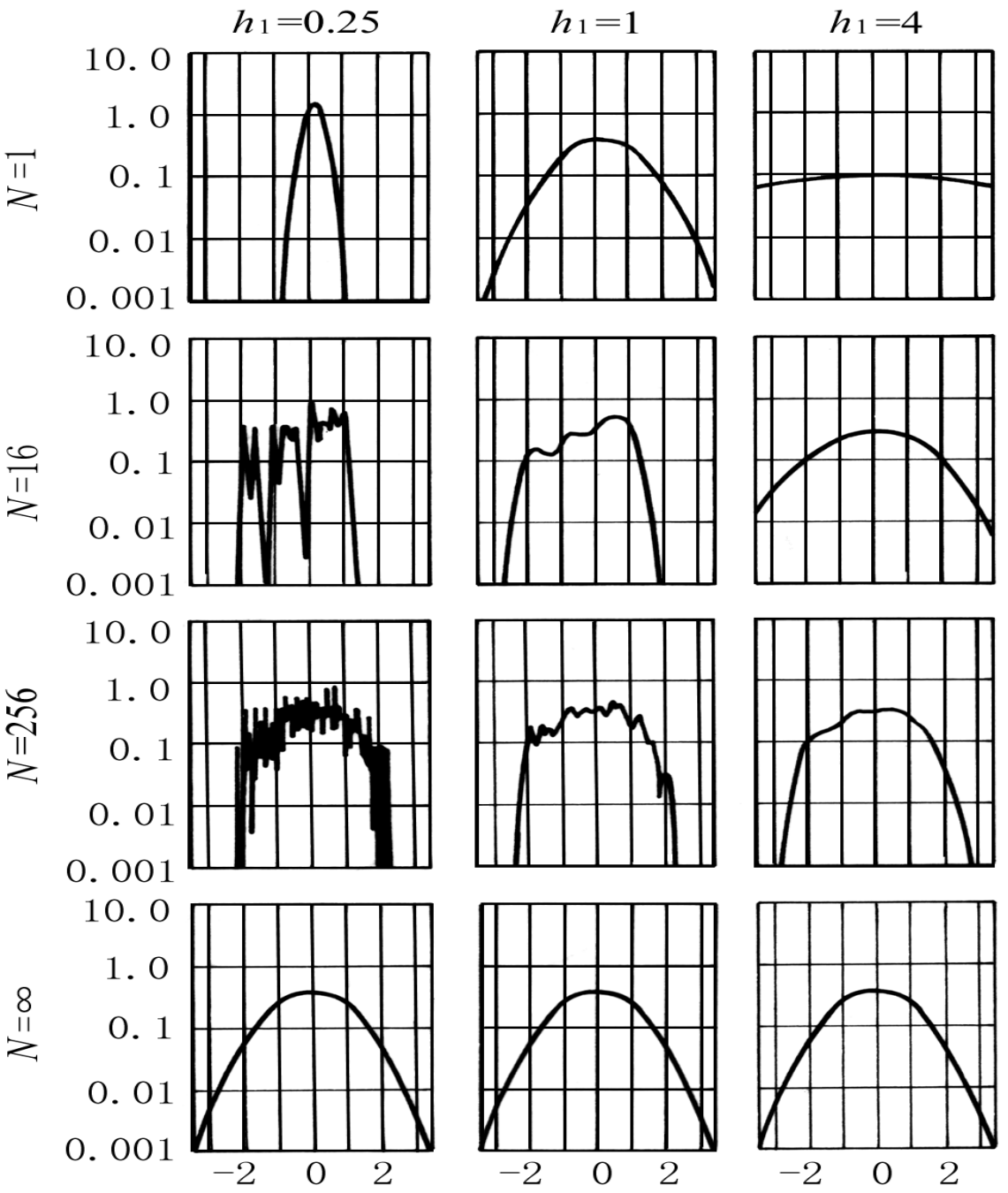
$$\sup_u \varphi(u) < \infty \qquad \lim_{\|u\|\to\infty} \varphi(u)\prod_{i=1}^{d} u_i = 0$$

（3）窗宽满足：$\lim_{N\to\infty} V_N = 0$ $\lim_{N\to\infty} NV_N = \infty$

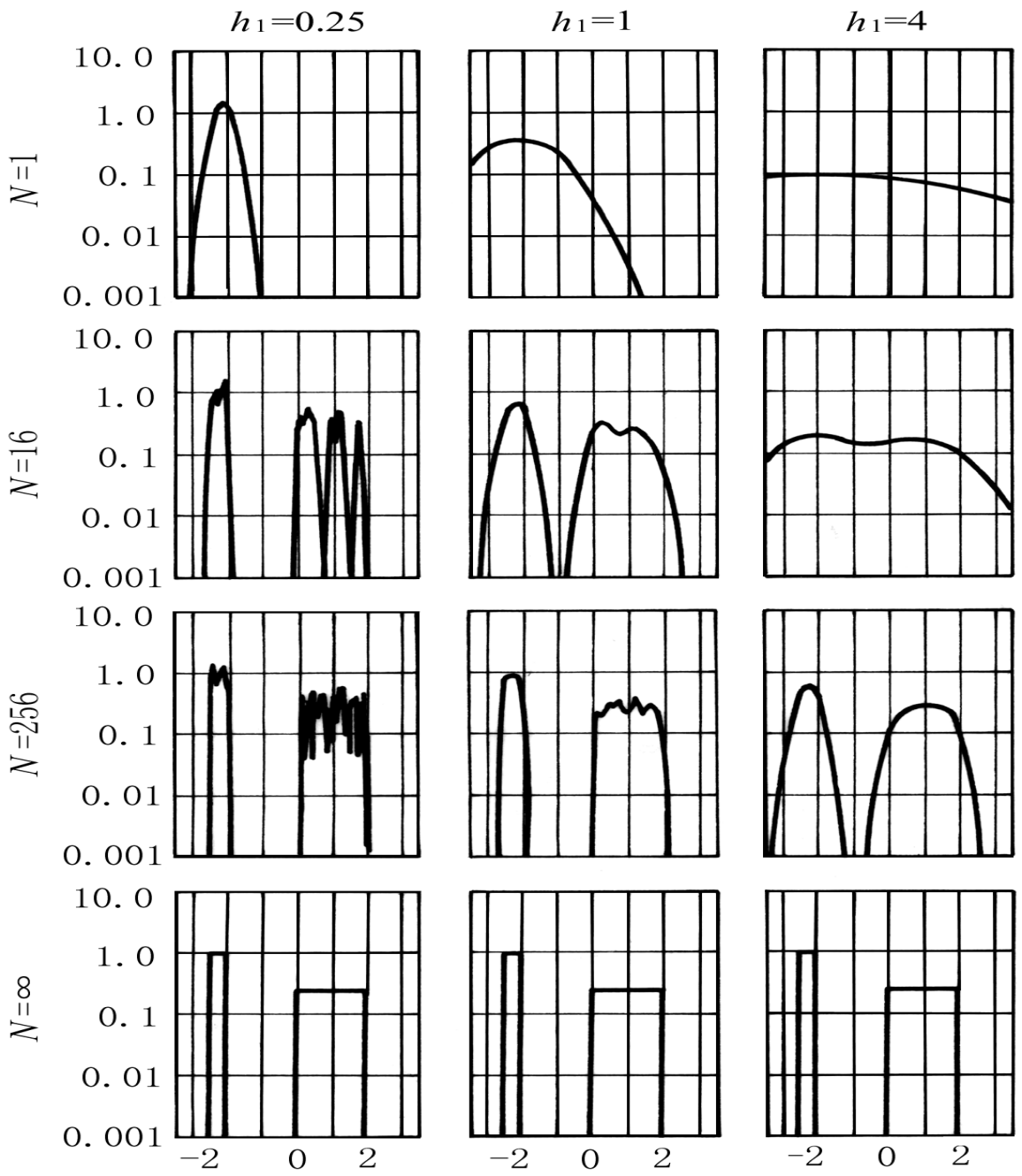
窗宽的影响

- 窗宽很大
- 窗宽很小

应用示例



应用示例



k_N 最近邻估计

- k_N 最近邻

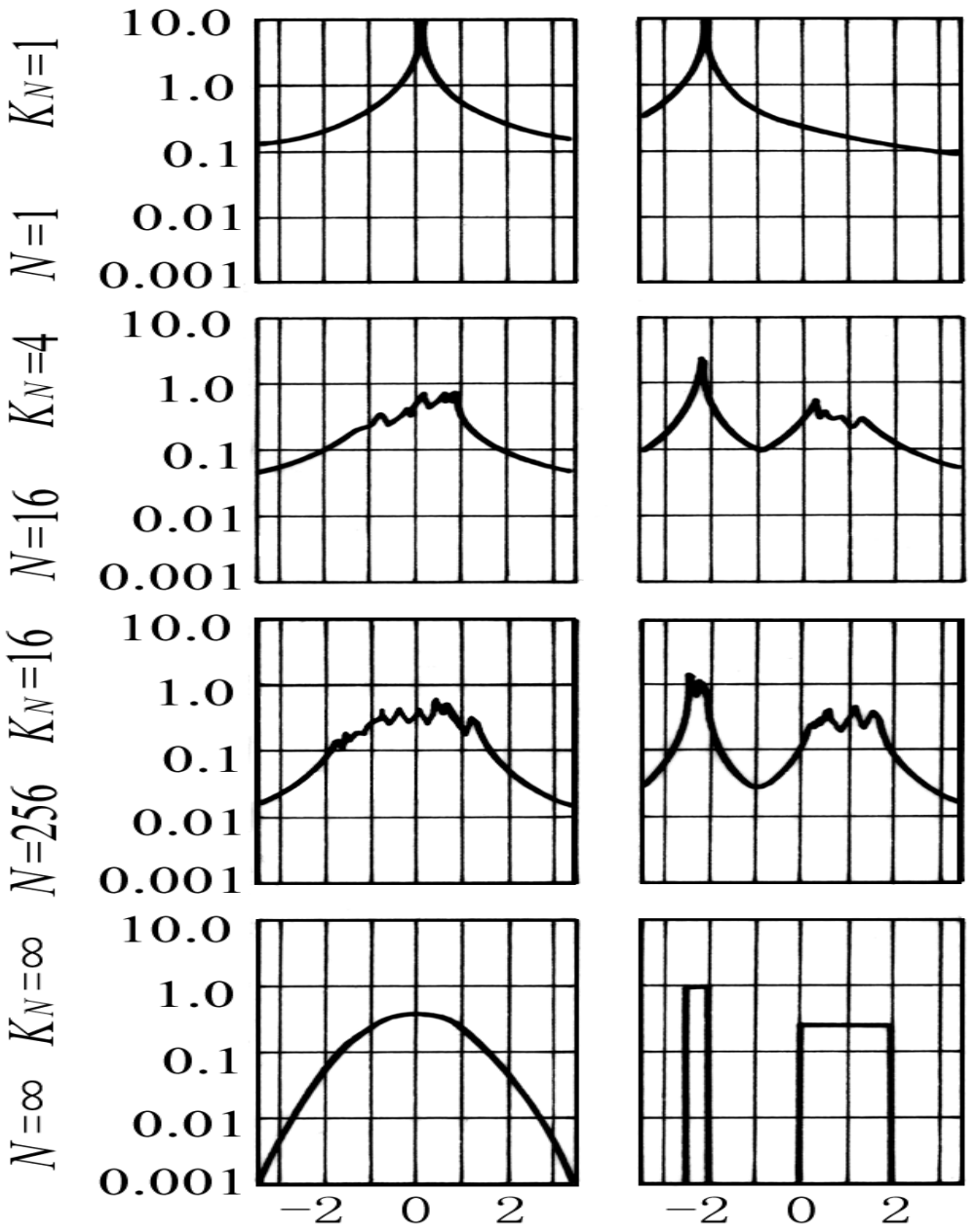
$$\hat{p}_N(x) = \frac{k_N / N}{V_N}$$

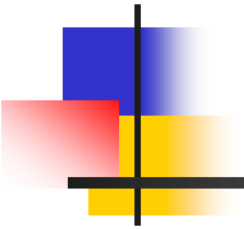
1) $\lim_{N\to\infty} V_N = 0$

2) $\lim_{N\to\infty} k_N = \infty$

3) $\lim_{N\to\infty} \frac{k_N}{N} = 0$

应用示例





- 最近邻估计

  构造 k_N 序列

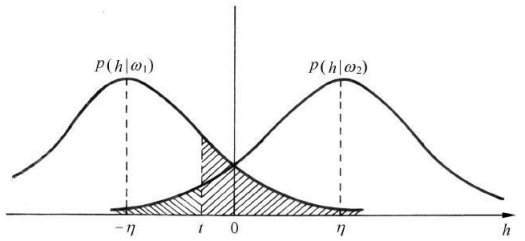
- Parzen窗

  构造 V_N 序列

一些问题：概率密度函数估计的准确性与分类器性能的关系

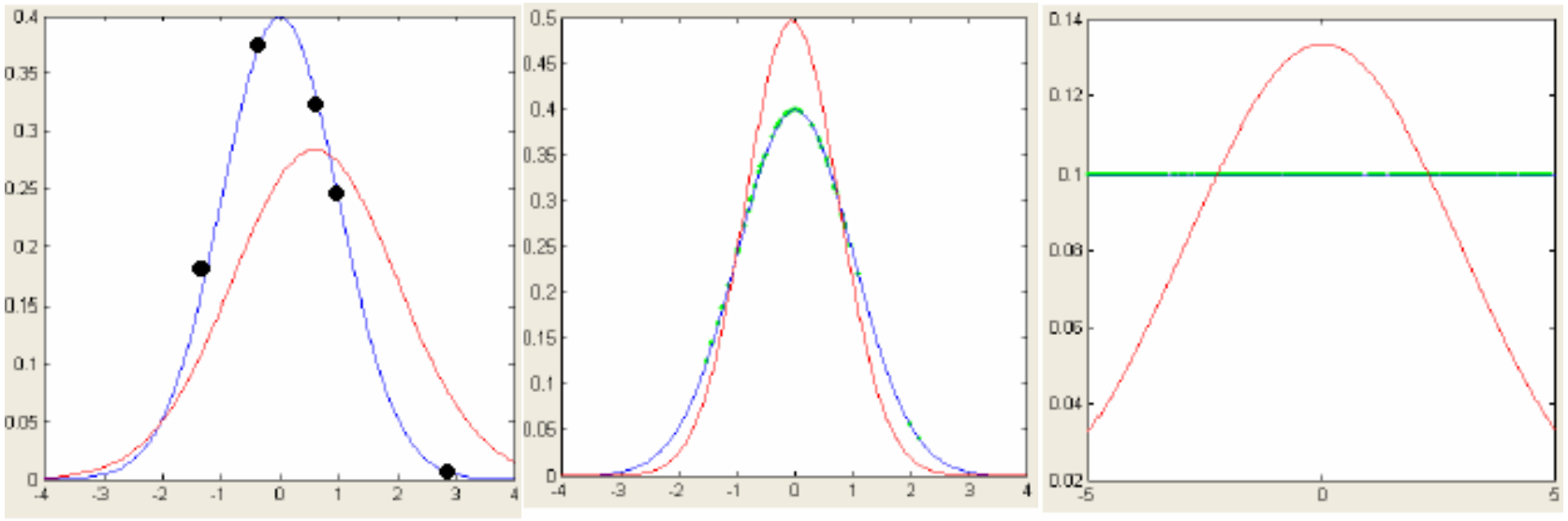
导致分类器产生误差的因素：

- **贝叶斯误差**
- 这种误差是由于不同的类条件概率分布函数之间的相互重叠引起的
- 是问题本身所有固有的，在分类器设计阶段是无法消除的



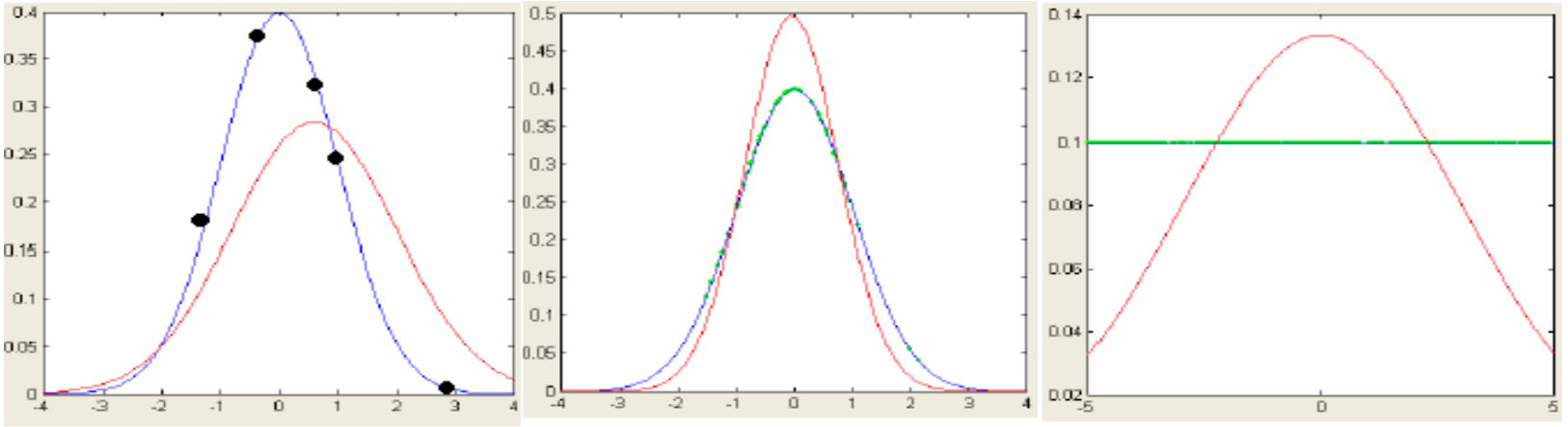
一些问题：概率密度函数估计的准确性与分类器性能的关系

- 模型误差
- 选择了不正确的模型所导致的分类误差
- 模型选择问题



一些问题：概率密度函数估计的准确性与分类器性能的关系

- 估计误差
- 由于采用有限样本进行估计所带来的误差
- 增加样本量
- 采用更好的估计方法



一些问题：维数问题

- **在估计一维概率密度函数时用数百个样本一般可以得到较好的结果**
- 一维概率密度函数：100　个样本
- 二维概率密度函数：100^2个样本
- D维：　　100^d个样本

- 维数灾难

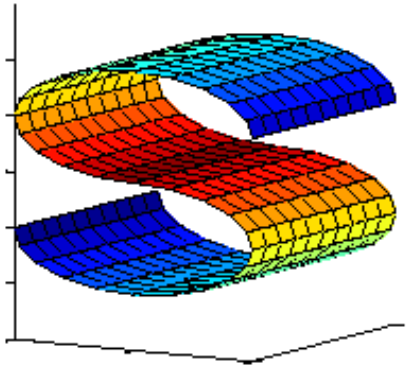
一些问题：维数问题

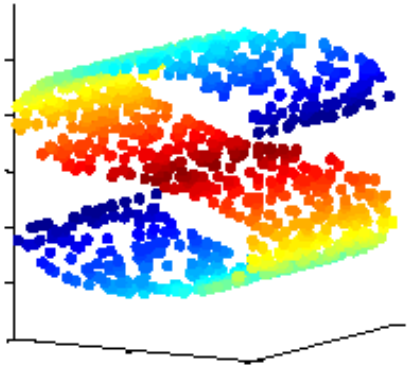
减小维数灾难的思路

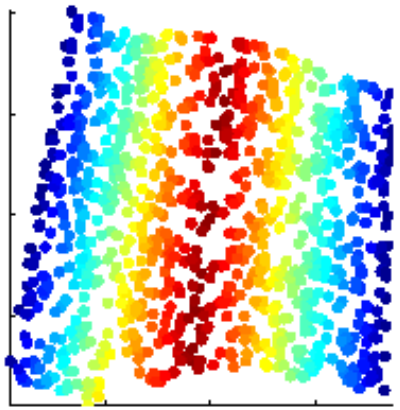
- 特征之间的独立性

$$p(x, y) = p(x)p(y)$$

- 大量的高维数据实际上嵌入在一个低维的流形上

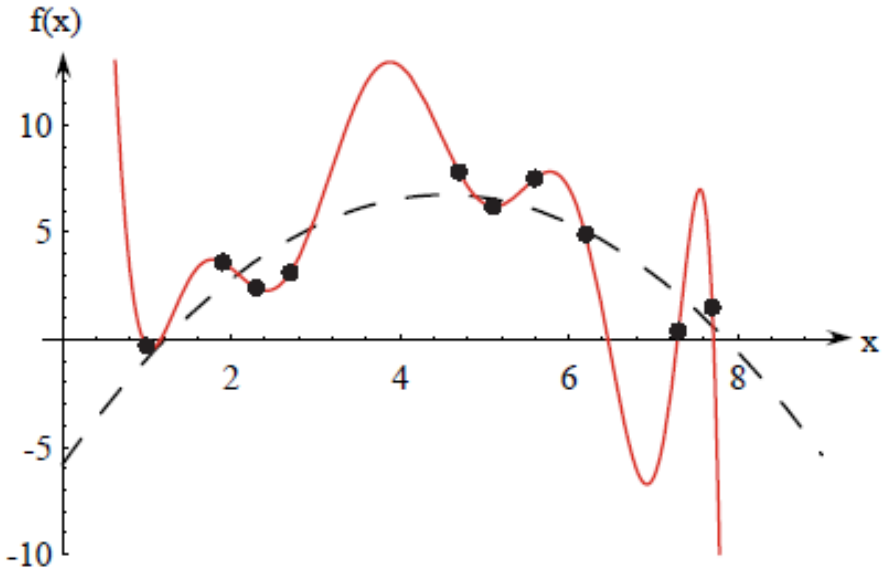






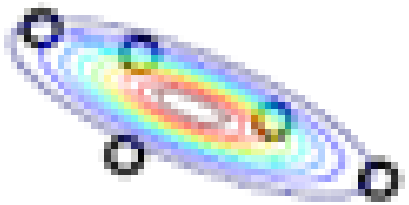
一些问题：过拟合

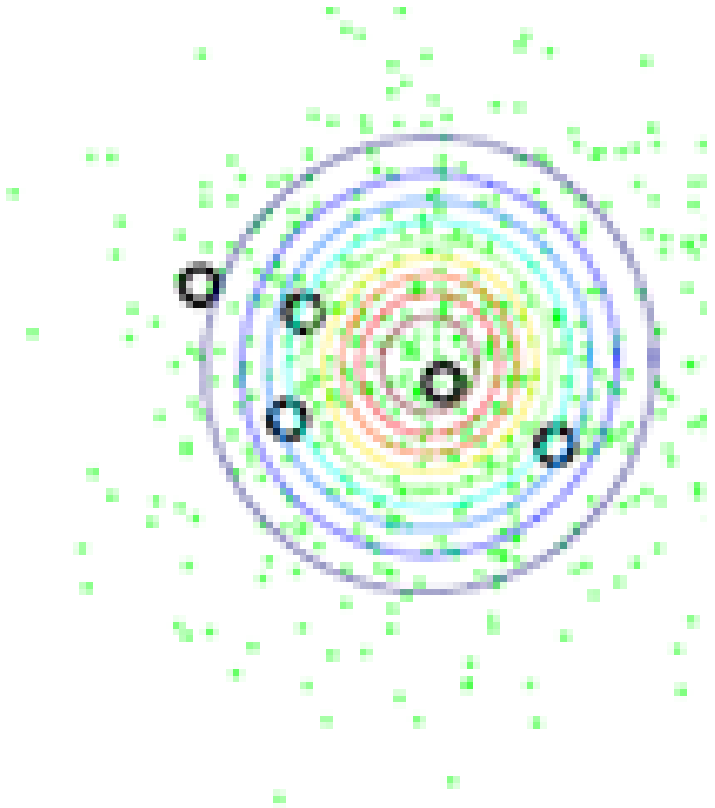
- **数据由抛物线函数**加上一个噪声项生成
- 用一条10阶的多项式曲线拟合
- 对于训练样本以外的其他样本的预测能力被称作做泛化能力，或推广能力



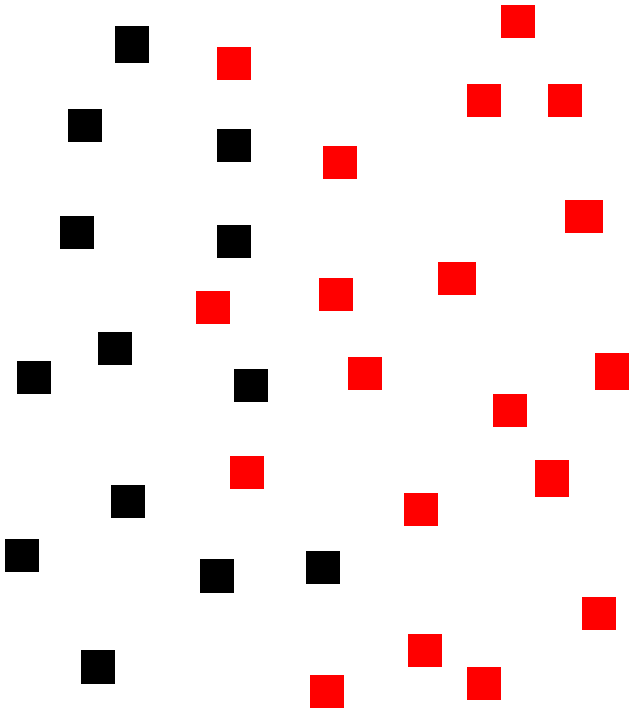
一些问题：过拟合

- 利用5个点估计一个正态分布

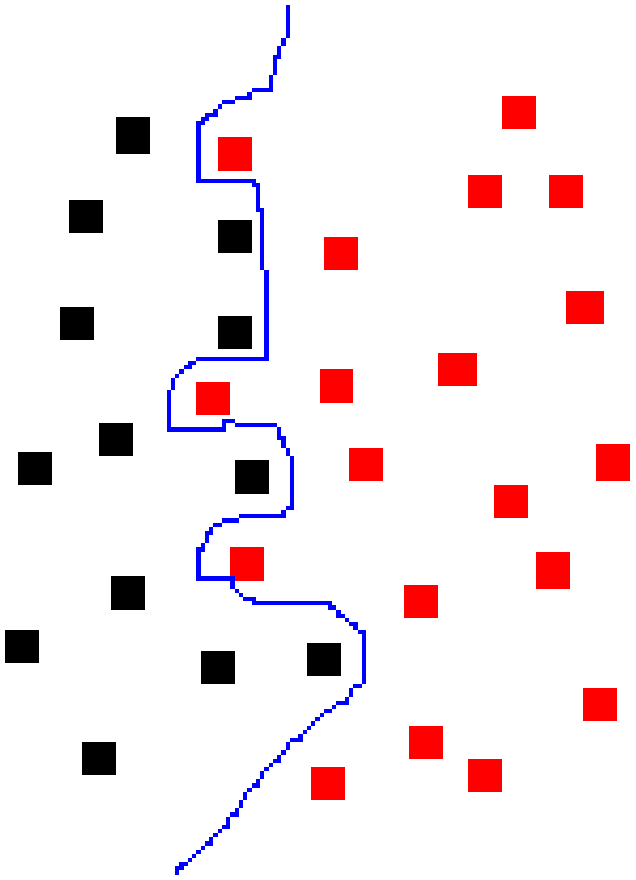




一些问题：过拟合



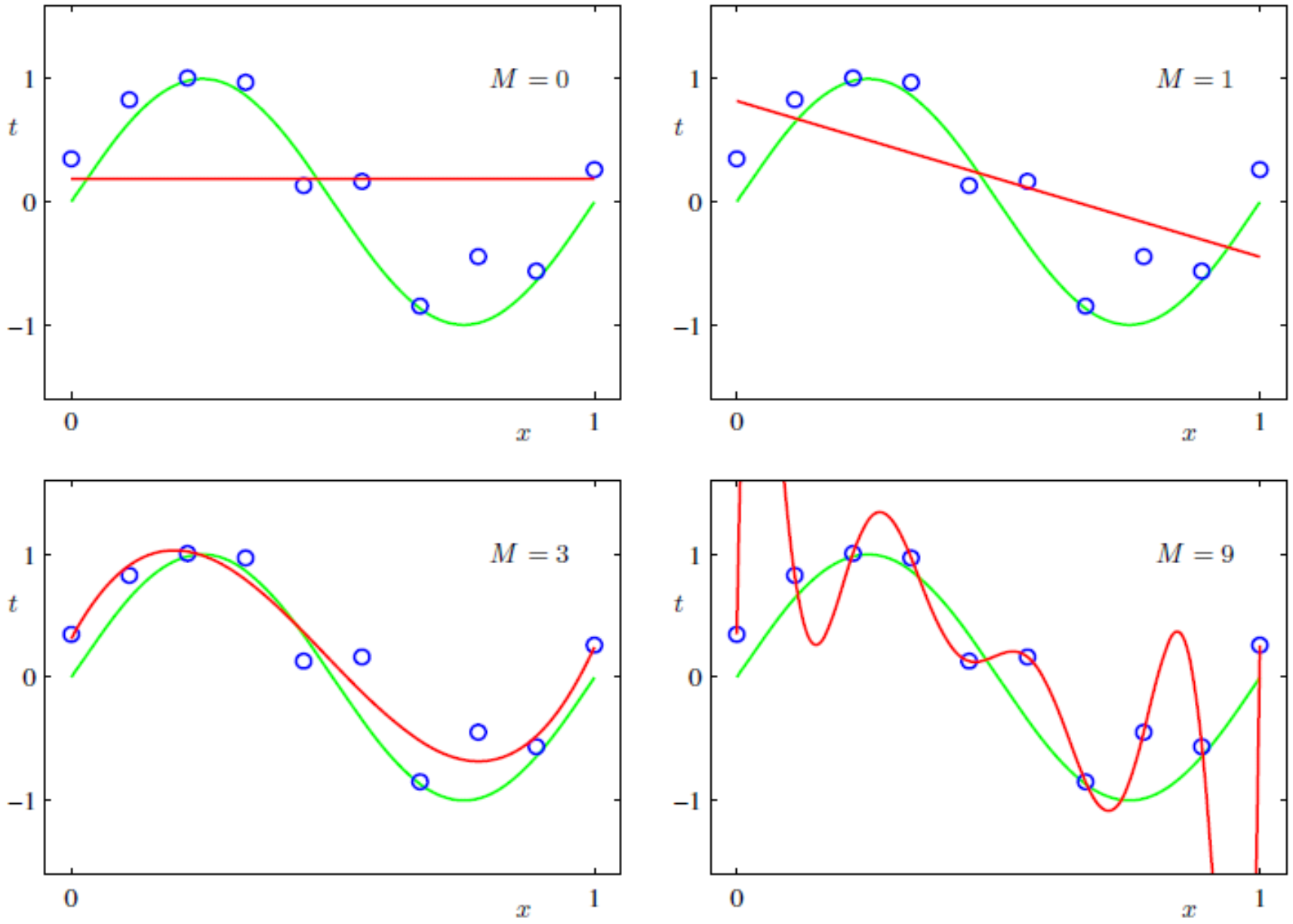
一些问题：过拟合



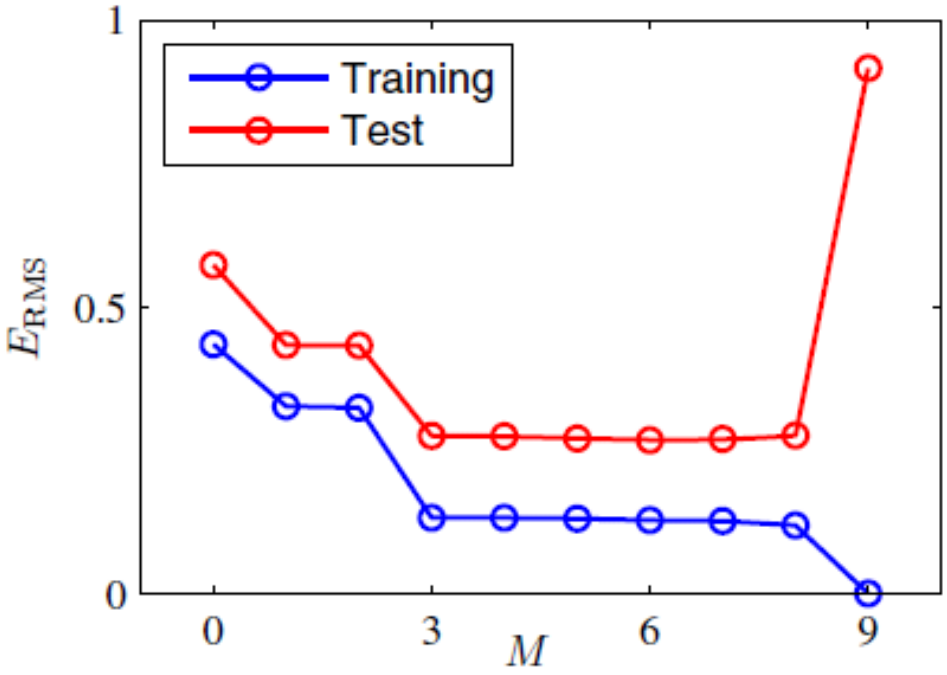
数据的多项式拟合

$$E(w)=\frac{1}{2}\sum_{n=1}^{N}\{y(x_n,w)-t_n\}^2$$

$y(x,w)$ 是多项式函数


M = 0
M = 1
M = 3
M = 9
t
x
0
1
−1

拟合误差



模型选择

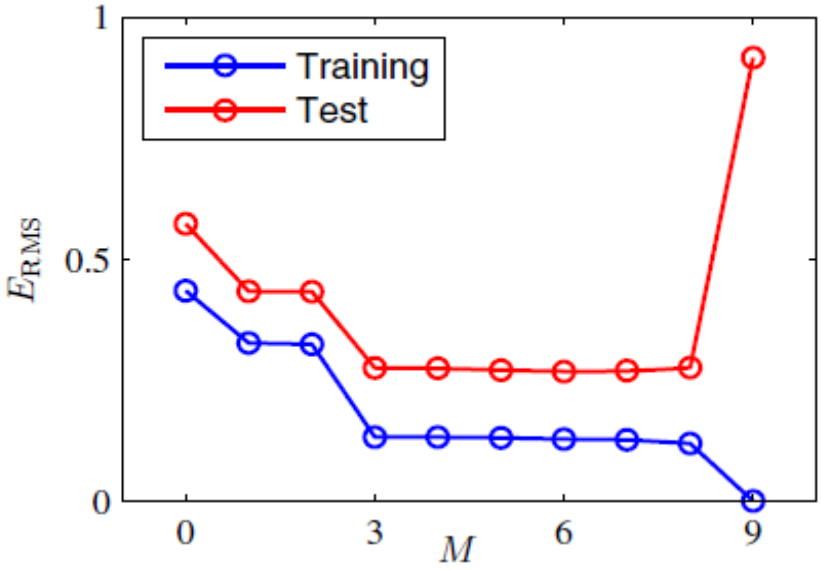
多项式的阶次M=?

M小，模型简单，不够灵活

M大，模型复杂，灵活

交叉验证：

训练集，验证集

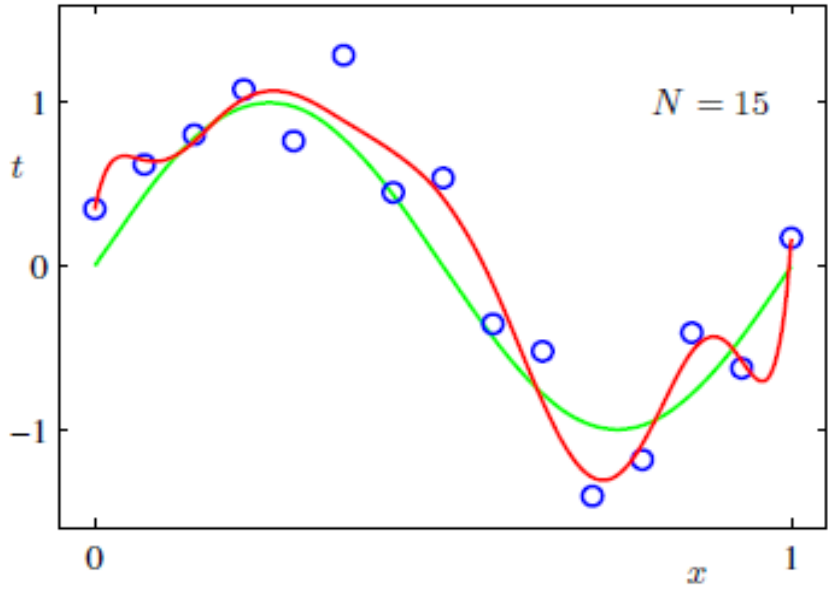


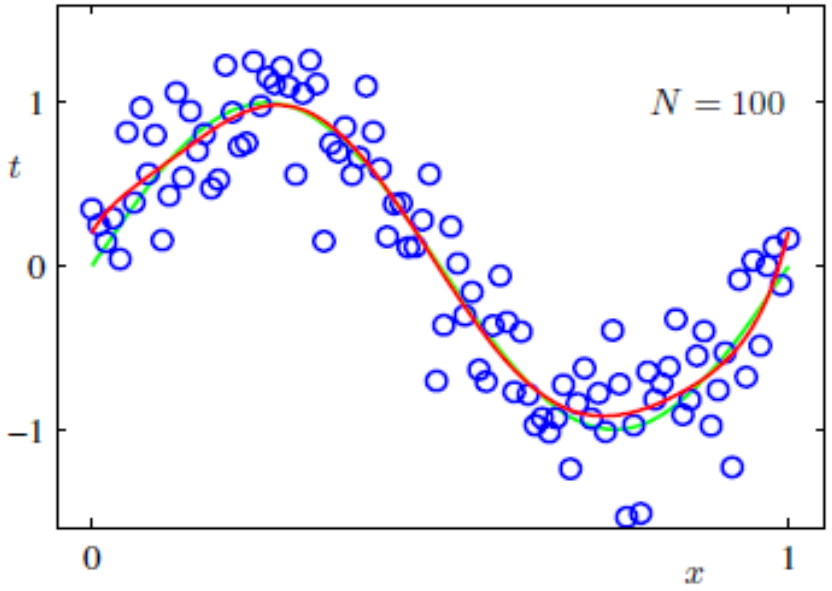
过拟合

- 用最小二乘方法来寻找模型的参数是最大似然的一个特例
- 过拟合问题是最大似然方法的一个普遍特性
- 采用贝叶斯方法可以避免过拟合问题。

如何避免过拟合？

- 样本数增加
- M=9





如何避免过拟合？

- 正则化方法

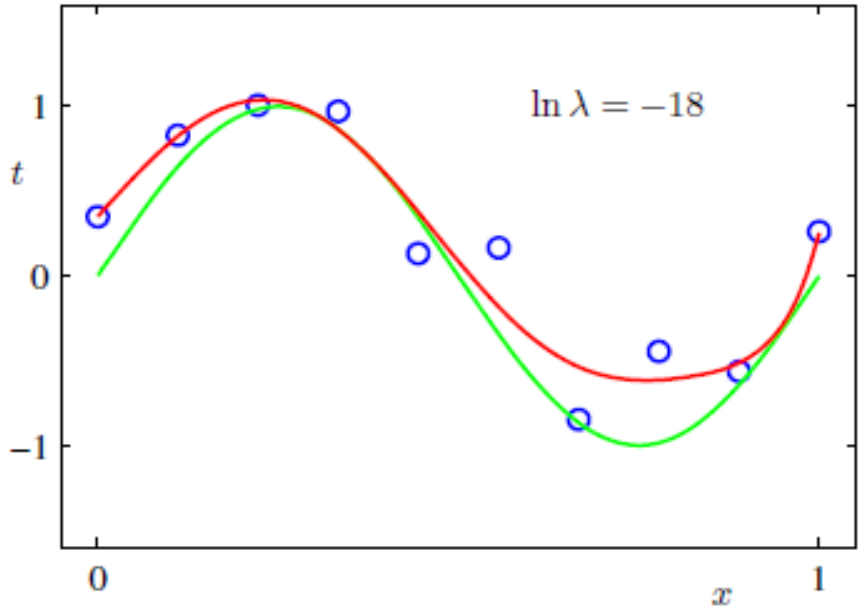
$$\tilde{E}(w) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, w) - t_n\}^2 + \frac{\lambda}{2}\|w\|^2$$

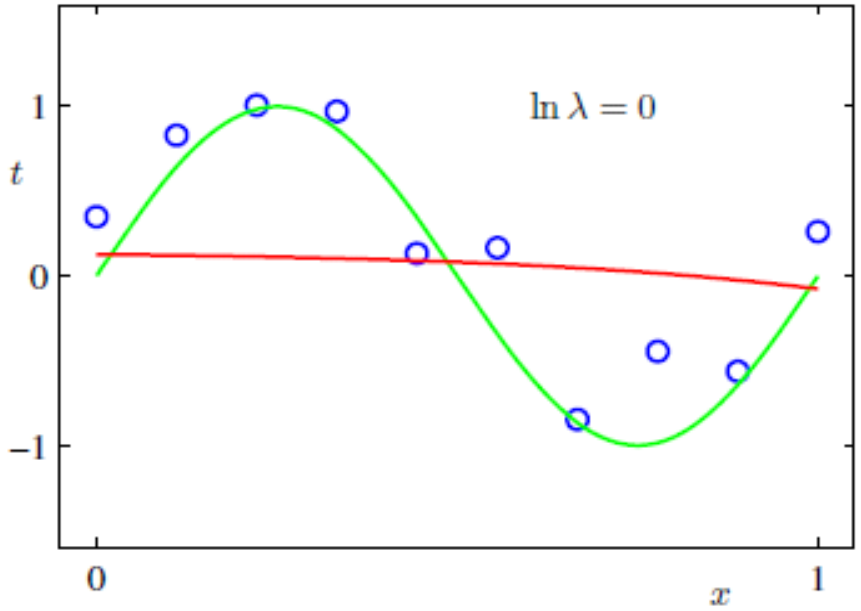
$$\|w\|^2 \equiv w^T w = \omega_0^2 + \omega_1^2 + \cdots + \omega_M^2$$

如何避免

- 正则化方法

	$M=0$	$M=1$	$M=6$	$M=9$
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43





- AIC： Akaike信息准则
- BIC： Bayesian信息准则， AIC的一个变形

$$\ln p(\mathcal{D}|w_{ML}) - M$$

减少参数的方法

- 使用参数化模型
- 使用共享参数
- 采用简单模型
- 低阶的多项式
- 假设协方差阵为对角阵
- 具有更少参数的模型

- 过拟合=过学习

估计错误率

- 理论上，精确讨论错误率存在困难
- 更多依赖于实验来估计错误率

两种情况

- 已设计好的分类器：使用检验集
- 未设计好的分类器：训练集,检验集

- 1. $P(\omega_1), P(\omega_2)$ 未知

抽出N 个样本，k 个样本分类错误的概率：

$$P(k) = C_N^k \varepsilon^k (1-\varepsilon)^{N-k}$$ ε：实际错误率

$$\frac{\partial \ln P(k)}{\partial \varepsilon} = \frac{\partial(\ln C_N^k + \ln \varepsilon^k + \ln(1-\varepsilon)^{N-k})}{\partial \varepsilon}$$

$$= \frac{k}{\varepsilon} - \frac{N-k}{1-\varepsilon} = 0 \qquad \hat{\varepsilon} = \frac{k}{N}$$

$$\hat{\varepsilon} = \frac{k}{N}$$

$$E(k) = N\varepsilon$$

$$Var(k) = N\varepsilon(1-\varepsilon)$$

$$E(\hat{\varepsilon}) = E[\frac{k}{N}] = \frac{E[k]}{N} = \frac{N\varepsilon}{N} = \varepsilon$$

$$Var[\hat{\varepsilon}] = {Var[k]}/{N^2} = {\varepsilon(1-\varepsilon)}/{N} \qquad (1)$$

2. $P(\omega_1), P(\omega_2)$ 已知：

$\omega_1 : N_1$个$\mathrm{P}(\omega_1)$

$\omega_2 : N_2$个$\mathrm{P}(\omega_2)$ $\quad N = N_1 + N_2$

- ω_1 类中有 k_1 个样本分错
- ω_2 类中有 k_2个样本分错

$$P(k_1, k_2) = P(k_1)P(k_2) = \prod_{i=1}^{2} C_{N_i}^{k_i} \varepsilon_i^{k_i} (1-\varepsilon_i)^{N_i - k_i}$$

- $\varepsilon_1, \varepsilon_2$ 是 ω_1, ω_2 类的实际错误率

$$\hat{\varepsilon}_i = \frac{k_i}{N_i} \qquad i = 1,2$$

总错误率：$\hat{\varepsilon}' = P(\omega_1)\hat{\varepsilon}_1 + P(\omega_2)\hat{\varepsilon}_2 = \sum_{i=1}^{2} P(\omega_i)\hat{\varepsilon}_i$

$$E[\hat{\varepsilon}'] = P(\omega_1)E[\hat{\varepsilon}_1] + P(\omega_2)E[\hat{\varepsilon}_2]$$
$$= P(\omega_1)\varepsilon_1 + P(\omega_2)\varepsilon_2 = \varepsilon$$

$$Var[\hat{\varepsilon}'] = \frac{1}{N}\sum_{i=1}^{2} P(\omega_i)\varepsilon_i(1-\varepsilon_i) \qquad (2)$$

比较（1）和（2）的方差：

$$
\begin{aligned}
&(1)-(2)\\
&=[\varepsilon(1-\varepsilon)-P(\omega_1)\varepsilon_1(1-\varepsilon_1)-P(\omega_2)\varepsilon_2(1-\varepsilon_2)]/N\\
&=[P(\omega_1)P(\omega_2)(\varepsilon_1-\varepsilon_2)^2]/N\geq 0
\end{aligned}
$$

容易理解

（1）极大似然估计

（2）无偏

3. 训练和测试总共有N个样本。

- 错误率与训练集和测试集相关
- 如何划分数据集？
- 留一法 1，2，……，N ：取出一个样本

K: 错分样本数

$$\hat{\varepsilon} = \frac{K}{N}$$

当前研究进展

- 概率图模型[1]中的参数估计
- 鲁棒估计

  鲁棒K均值法[3], L1范数参数估计[5]
- 变分估计法[4]
- 非参数贝叶斯方法[6]

参考文献

- [1] Probabilistic graphical models: principles and techniques. Koller, D. & Friedman, N. The MIT Press, 2009
- [2]: Loopy belief propagation for approximate inference: An empirical study.Murphy, K.; Weiss, Y. & Jordan, M. Proceedings of Uncertainty in AI, 1999, 9, 467-475

参考文献

- [3] A review of robust clustering methods .Garcia-Escudero, L.; Gordaliza, A.; Matrán, C. & Mayo-Iscar, A. Advances in Data Analysis and Classification, Springer Berlin / Heidelberg, 2010, 4, 89-109
- [4] Graphical models, exponential families, and variational inference. Wainwright, M. & Jordan, M. Foundations and Trendstextregistered in Machine Learning, Now Publishers Inc., 2008, 1, 1-305

参考文献

- [5] L1-penalized robust estimation for a class of inverse problems arising in multiview geometry.Dalalyan, A.; Keriven, R. & Marne-la-Vallée, F. Annual Conference on Neural Information Processing Systems, Vancouver, Canada, 2009
- [6] A Tutorial on Bayesian Nonparametric Models. Samuel J. Gershmana, David M. Blei.