第八章

特征的提取与选择

特征的提取

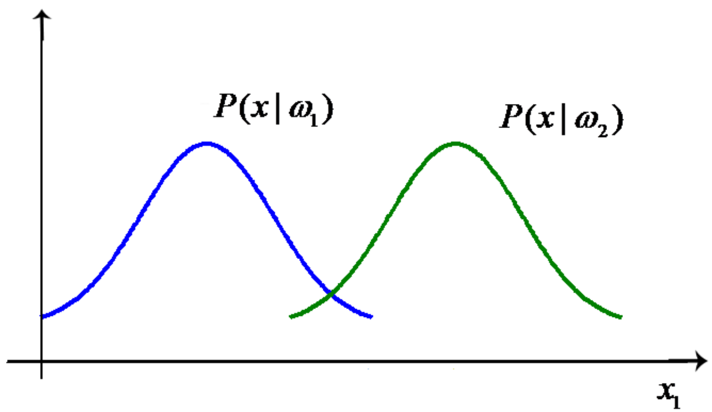
- 概念：特征提取，特征选择

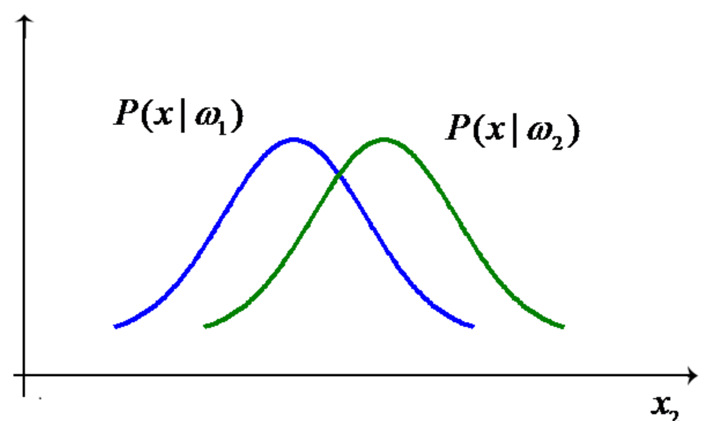
特征的提取

- 特征对分类器性能的影响

举例:

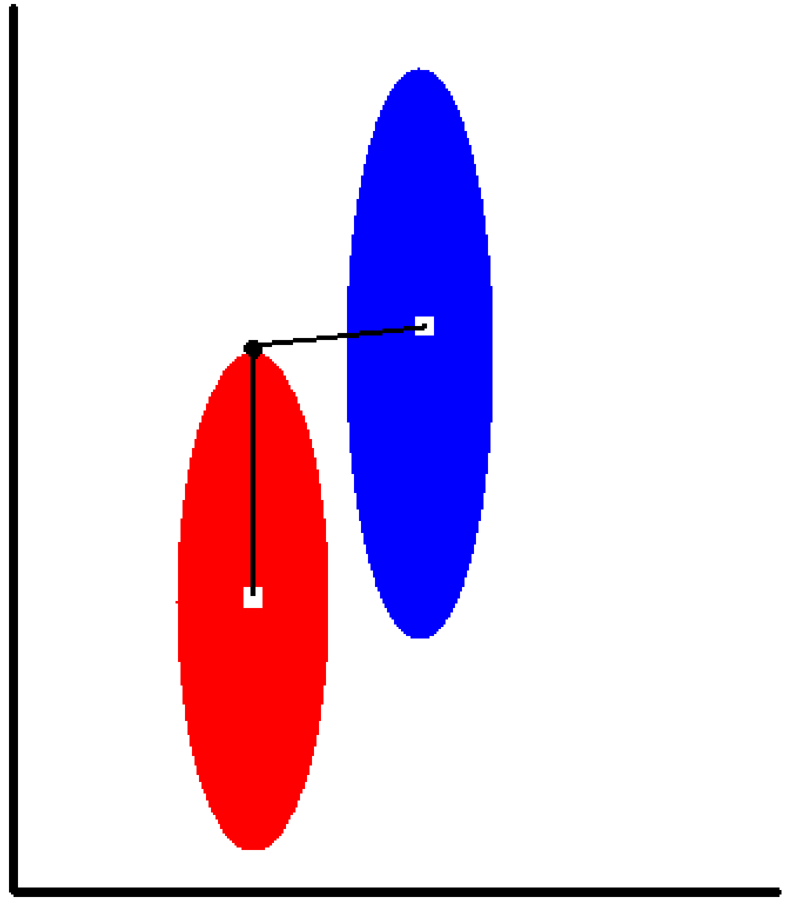
1. 玉米与杂草：高度、粗、化学物质
2. 长桌与方凳：长，宽，高，颜色，质地
3. 对人的识别





特征的提取

- 特征越多越好吗？



特征的提取

什么特征具有分类价值？

什么特征容易提取？

笔画的多少

像素的多少

赢 大

特征的提取

什么特征具有分类价值？
什么特征有好的稳定性？
人脸的几何信息稳定吗？
指纹的端点和分叉点？

特征的提取

什么特征具有分类价值？

获取什么特征代价比较小？

人脸？指纹？DNA?

特征的提取

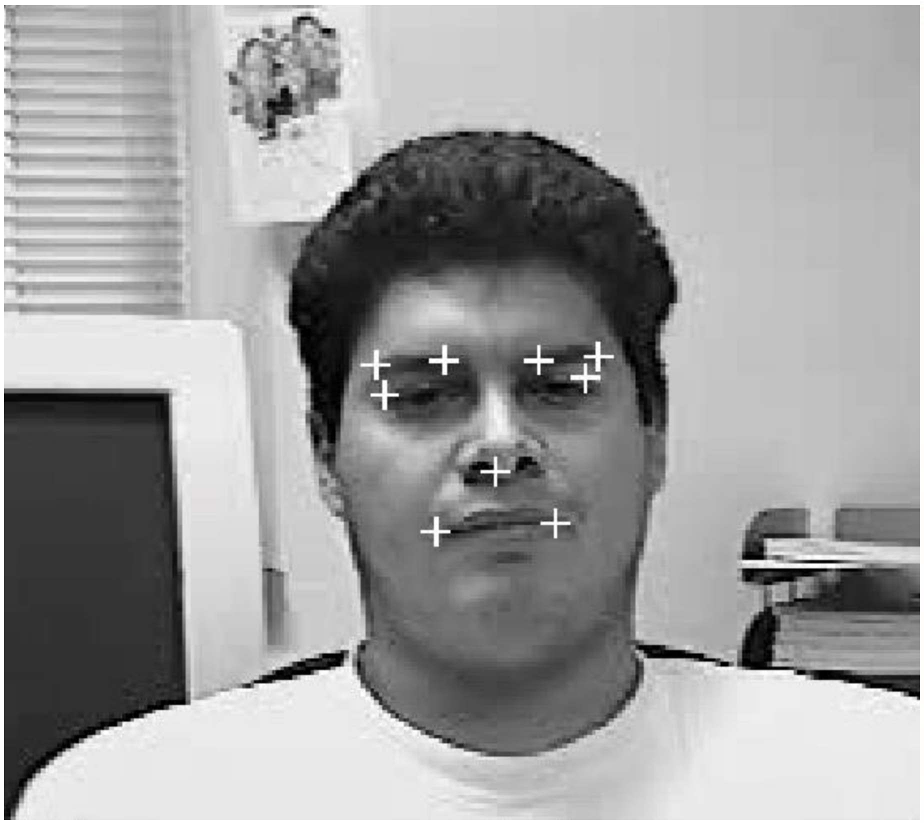
提取特征的方法

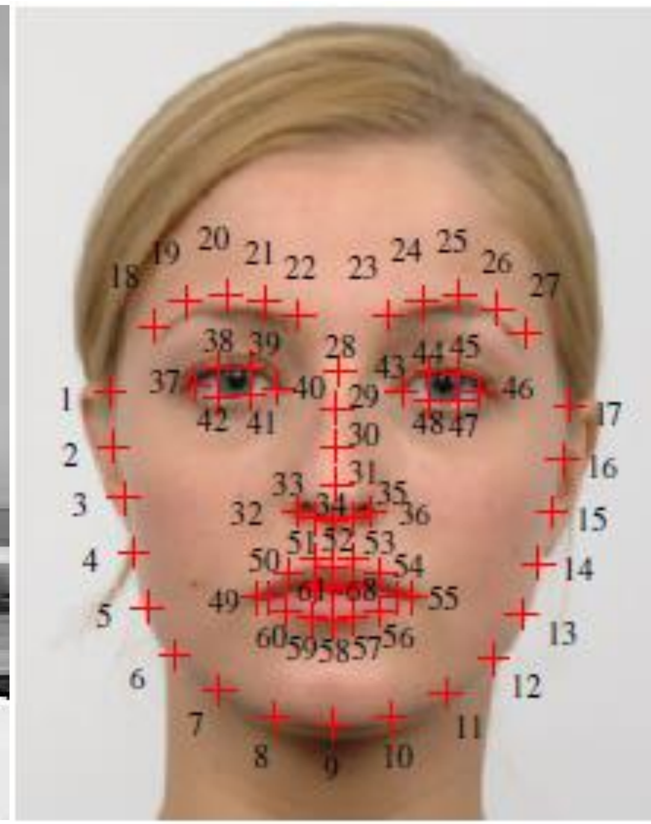
- 各种数据处理的理论和技术
- 信号处理，图象处理
- 生物医学信号处理，雷达信号处理，生物图象处理

特征的提取—图象处理

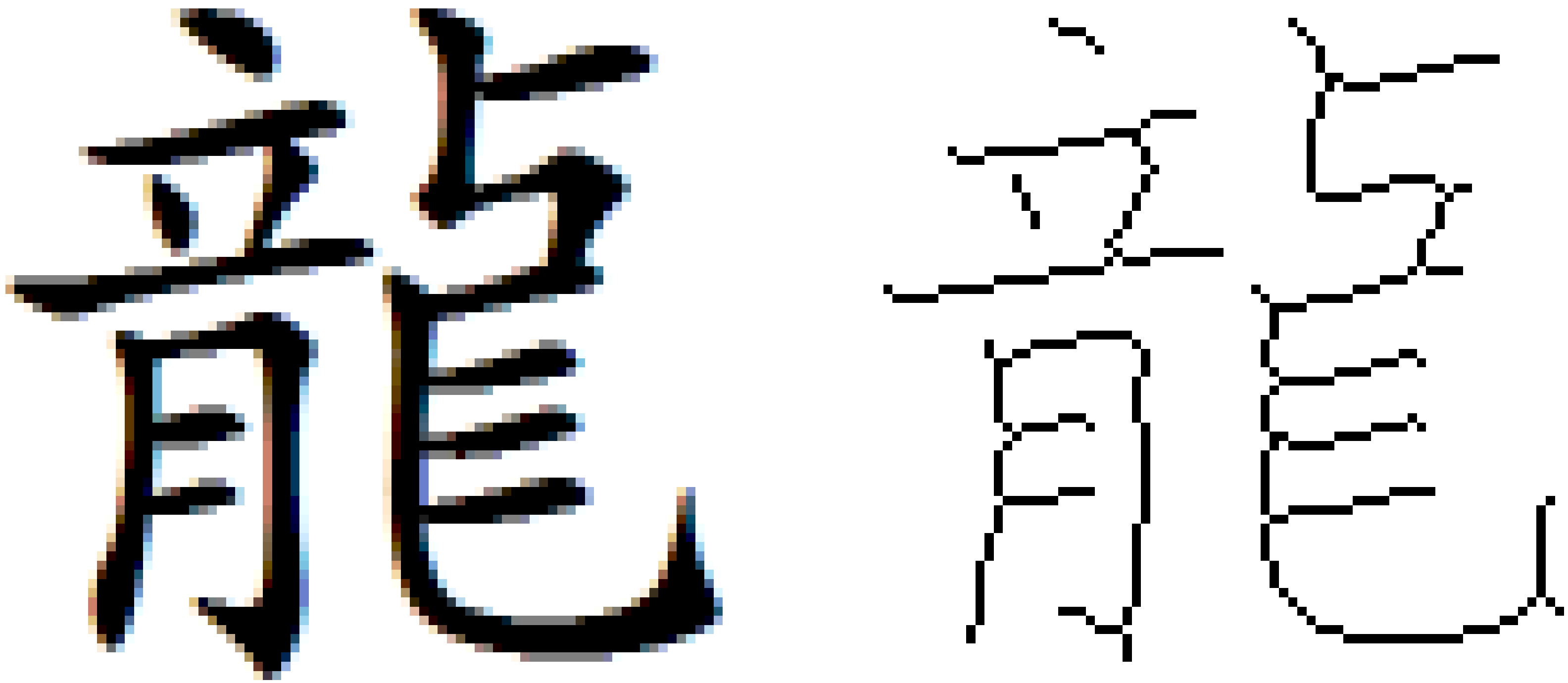
特征的提取—图象处理

人脸特征提取



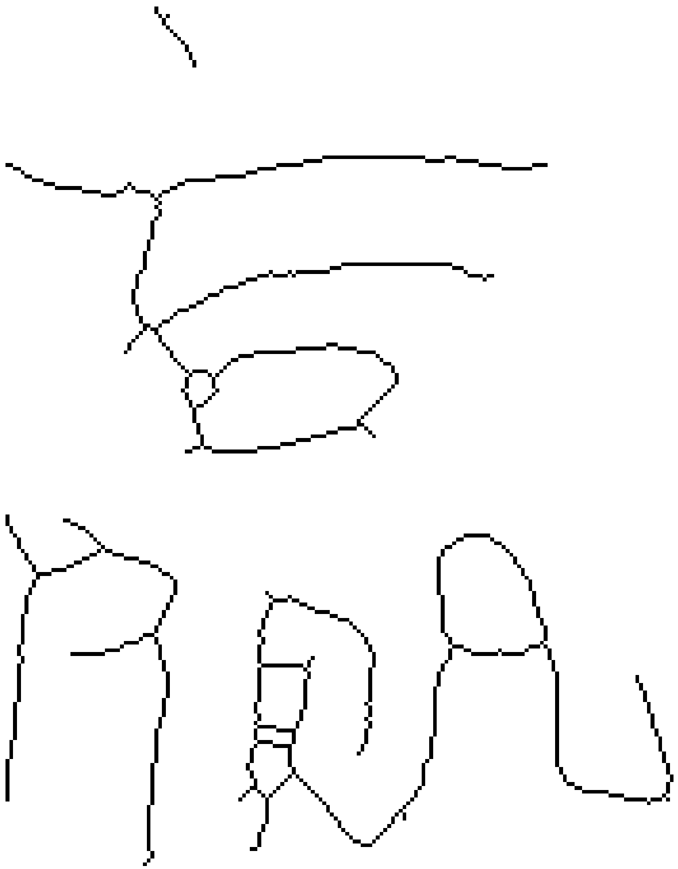


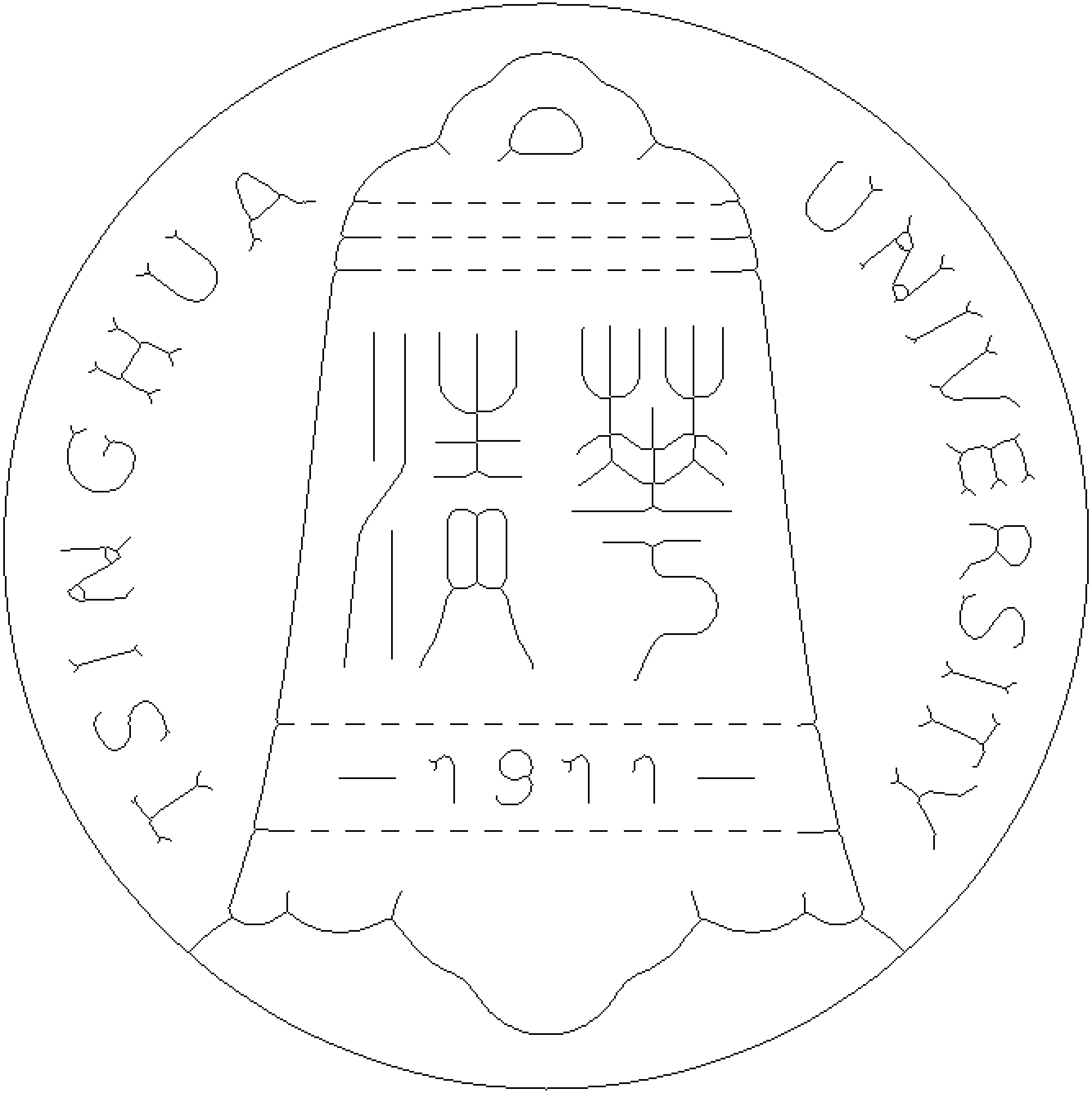
特征的提取—图象处理



特征的提取－图象处理

特征的提取－图象处理




TSINGHUA UNIVERSITY
1911

清華大學

自強不息 厚德載物

TSINGHUA UNIVERSITY

-1911-

清華大學

Tsinghua University

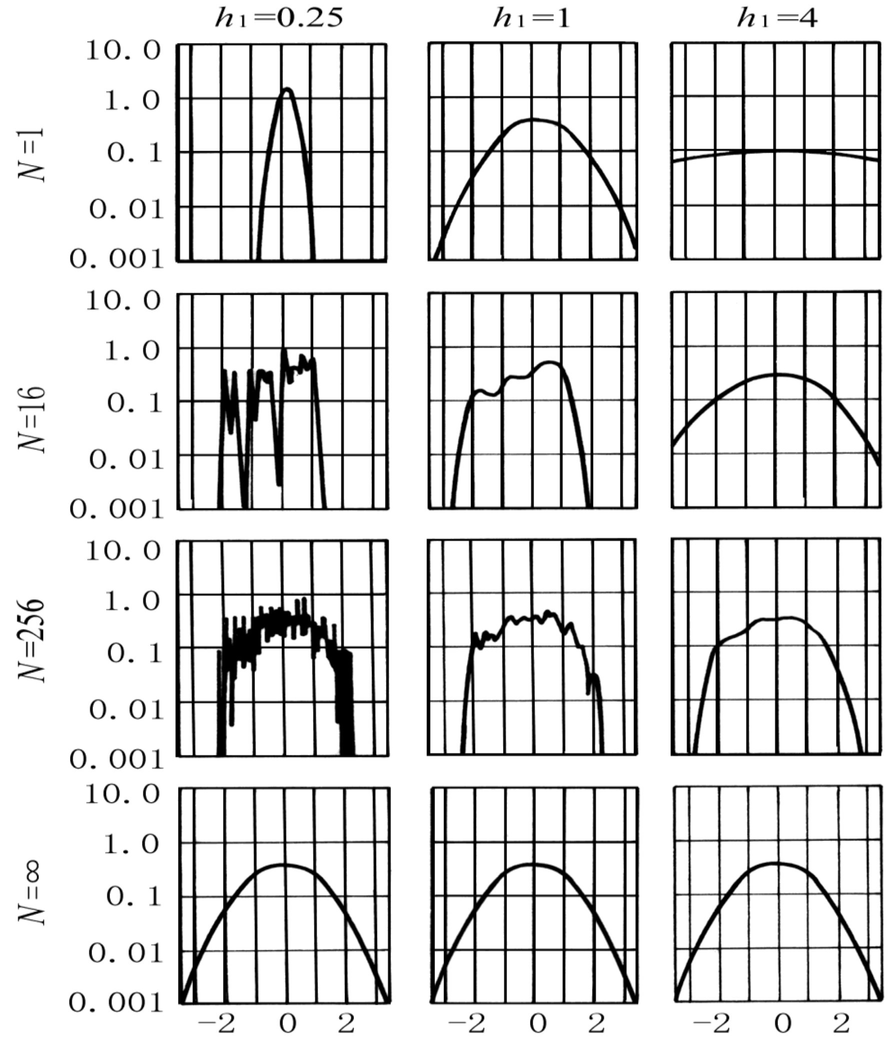

Tsinghua University

特征的提取

- 对差异性机理的研究
- 对专家的依赖性

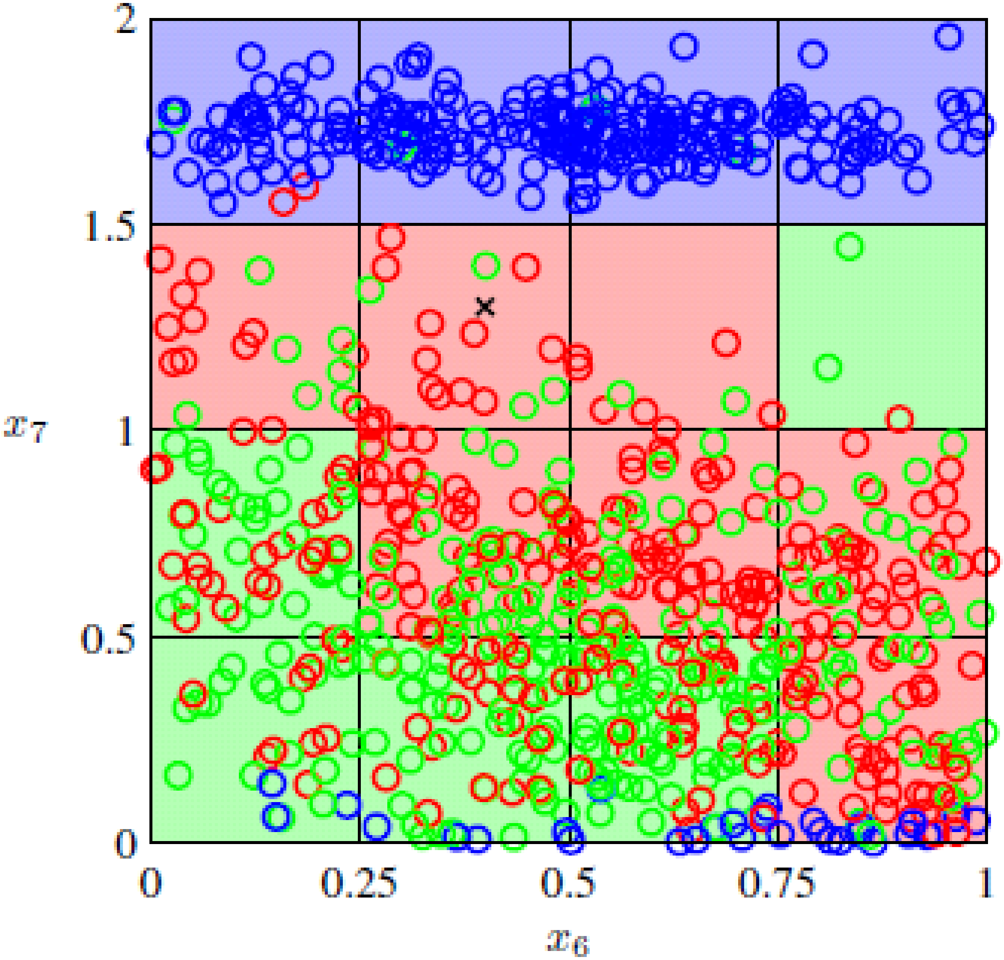
维数灾难（The Curse of Dimensionality）

概率密度函数估计



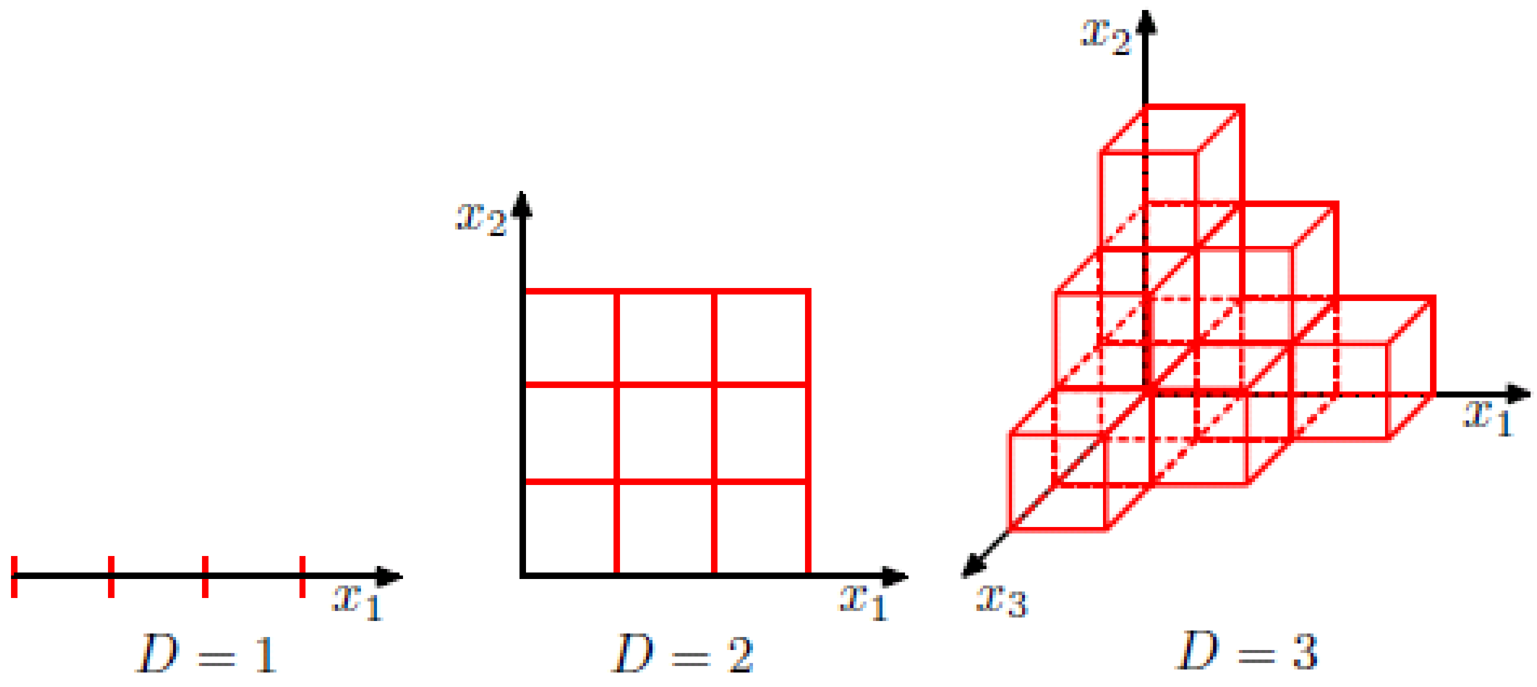
维数灾难

- 概率密度函数估计
- 二维方格



维数灾难

- 概率密度函数估计
- 方格数随维数的增长呈指数增长
- 大量格子中是空的

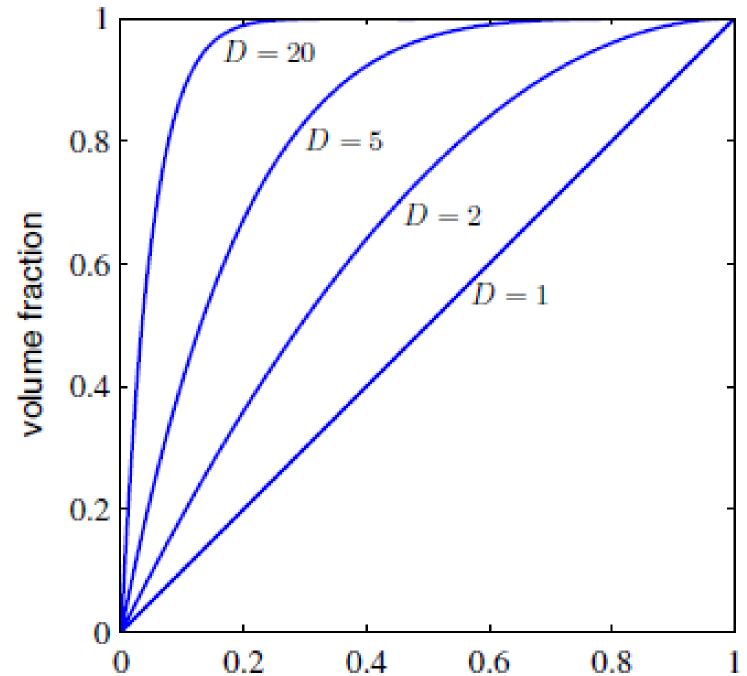


举例

- 三维几何直观使我们无法思考高维空间
- 一个D维空间半径r=1的球
- 该球处于半径 $r = 1 - \epsilon$ 和r=1之间的部分占整个体积的比例

$$V_D(r) = K_D r^D$$

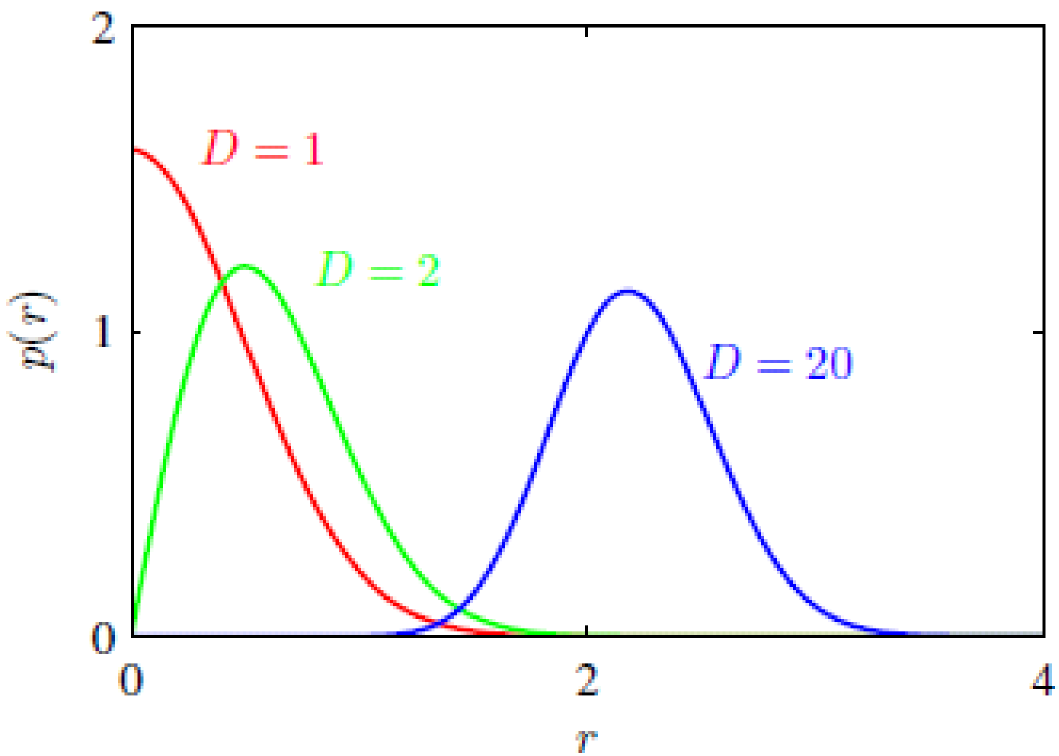
$$\frac{V_D(1) - V_D(1-\epsilon)}{V_D(1)} = 1 - (1-\epsilon)^D$$



高维数据

- 高斯分布：位于 r 厚度为 δr 的概率质量

$$p(r)\delta r$$

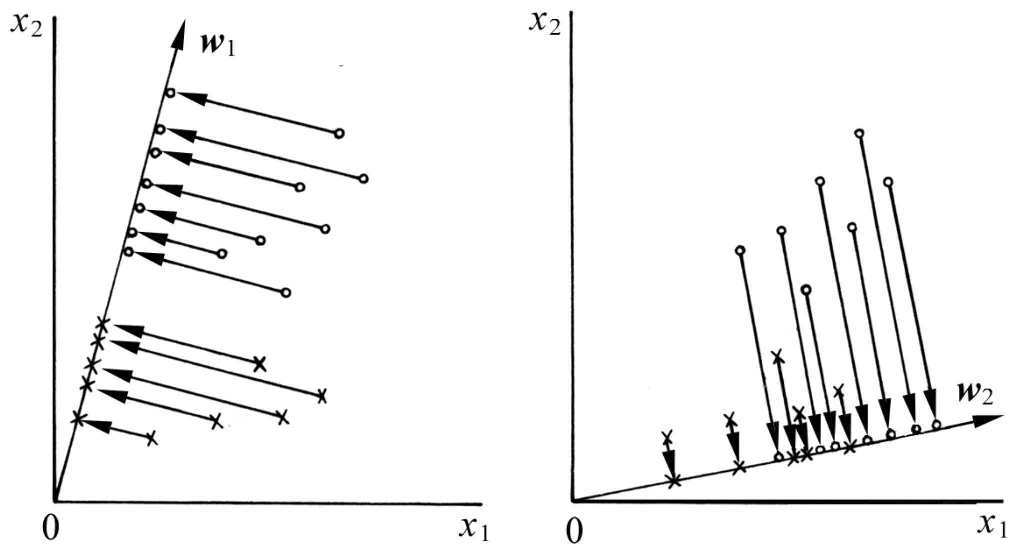


维数灾难

- 源自高维空间的困难被称作维数灾难
- 真实数据经常局限在空间中一个有着较低有效维数的区域中，目标变量的重要的变化方向可能是被局限的：流形
- 实际数据通常有一些平滑性质（至少是局部的），大部分情况下，输入变量的小的变化会产生目标变量的小的变化

Fisher准则

问题：把d维空间的样本投影到一条直线上，在这条直线上，样本能够最容易的分开



$N: x_1, \cdots, x_N$

$\chi_1: N_1$个样本构成的样本集,

$\chi_2: N_2$个样本构成的样本集

$N_1 + N_2 = N$

$$y_n = w^T x_n, n = 1,2,\cdots, N_i, i = 1,2$$

$$m_i = \frac{1}{N_i}\sum_{x\in\chi_i} x, \qquad i = 1,2$$

$$S_i = \sum_{x\in\chi_i}(x-m_i)(x-m_i)^T, \quad i = 1,2$$

S_i:类内离散度矩阵

$S_b = (m_1 - m_2)(m_1 - m_2)^T$

类间离散度矩阵

$S_w = S_1 + S_2$:总的类内离散度矩阵

$$\tilde{m}_i = \frac{1}{N_i}\sum_{y\in Y_i} y, \qquad i = 1,2$$

$$\tilde{S}_i^2 = \sum_{y\in Y_i} (y - \tilde{m}_i)^2, \qquad i = 1,2$$

$$\tilde{S}_w = \tilde{S}_1^2 + \tilde{S}_2^2$$

$$J_F(w) = \frac{(\tilde{m}_1 - \tilde{m}_2)^2}{\tilde{S}_1^2 + \tilde{S}_2^2}$$

$$\tilde{m}_i = \frac{1}{N_i}\sum_{y\in Y_i} y = \frac{1}{N_i}\sum_{x\in\chi_i} w^T x \; = w^T(\frac{1}{N_i}\sum_{x\in\chi_i} x) = w^T m_i$$

$$(\tilde{m}_1 - \tilde{m}_2)^2 = (w^T m_1 - w^T m_2)^2$$

$$= w^T(m_1 - m_2)(m_1 - m_2)^T w = w^T S_b w$$

$$\tilde{S}_i^{\,2} = \sum_{y\in Y_i}(y - \tilde{m}_i)^2 = \sum_{x\in\chi_i}(w^T x - w^T m_i)^2$$

$$= w^T[\sum_{x\in\chi_i}(x - m_i)(x - m_i)^T]w$$

$$= w^T S_i w$$

$$J_F(w) = \frac{w^T S_b w}{w^T S_w w}$$

令 $w^T S_w w = c \neq 0$ $\qquad L = w^T S_b w - \lambda(w^T S_w w - c)$

$$\frac{\partial L}{\partial w} = S_b w - \lambda S_w w = 0$$

$$S_b w = \lambda S_w w$$

$$S_w^{-1} S_b w^* = \lambda w^*$$

$$S_b w^* = (m_1 - m_2)(m_1 - m_2)^T w^*$$
$$= (m_1 - m_2)R$$

$$\lambda w^* = S_w^{-1}(S_b w^*) = S_w^{-1}(m_1 - m_2)R$$

$$w^* = S_w^{-1}(m_1 - m_2)$$

分类

采取下面的方法分类:

$$y_0^{(1)} = \frac{\tilde{m}_1 + \tilde{m}_2}{2} \qquad y_0^{(2)} = \frac{N_2\tilde{m}_1 + N_1\tilde{m}_2}{N_1 + N_2}$$

$$y_0^{(3)} = \frac{\tilde{m}_1 + \tilde{m}_2}{2} + \frac{\ln(P(w_1)/P(w_2))}{N_1 + N_2 - 2}$$

$$y > y_0 \rightarrow x \in w_1$$

$$y < y_0 \rightarrow x \in w_2$$

分类

还可以采取下面的方法分类:

- 在一维上估计概率密度函数,用Bayes决策方法.
- 考虑方差, 从中值向方差小的类别移动.

问题

- Fisher判别适合哪种数据的分布情况?
- 可以考虑多类吗?
- 有几个特征向量? 取哪一个向量?
- 散度矩阵前考虑先验加权.
- 可以投影到平面吗? 可以投影到一般的低维空间吗?
- 总的类内散度矩阵一定可逆吗? 不可逆怎么办?

Fisher准则的研究

- 非线性Fisher方法: Kernel Fisher
- 零子空间方法
- Fisher 与回归问题的等价性
- 局部Fisher方法
- Hastie T and Tibshirani R. Discriminant adaptive nearest neighbor classification. IEEE Trans. On PAMI, 1996, 18(6):409-415

线性判别分析

其它判据 $J_2 = \mathrm{tr}(S_w^{-1} S_b)$ $J_3 = \ln\left[\frac{|S_b|}{|S_w|}\right]$

$$J_4 = \frac{\mathrm{tr} S_b}{\mathrm{tr} S_w} \qquad J_5 = \frac{|S_w + S_b|}{|S_w|}$$

问题：求W使 $x = W^T y$ 的判据最大，$(J_2, \cdots, J_5)$

求 $S_w^{-1} S_b$ 的特征值 $\lambda_1, \lambda_2, \cdots, \lambda_D$ 有：

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_D$$

选前d个特征值对应的特征向量

$$W = [u_1, u_2, \cdots, u_d]$$

特征选择

D个特征，选d个

准则：

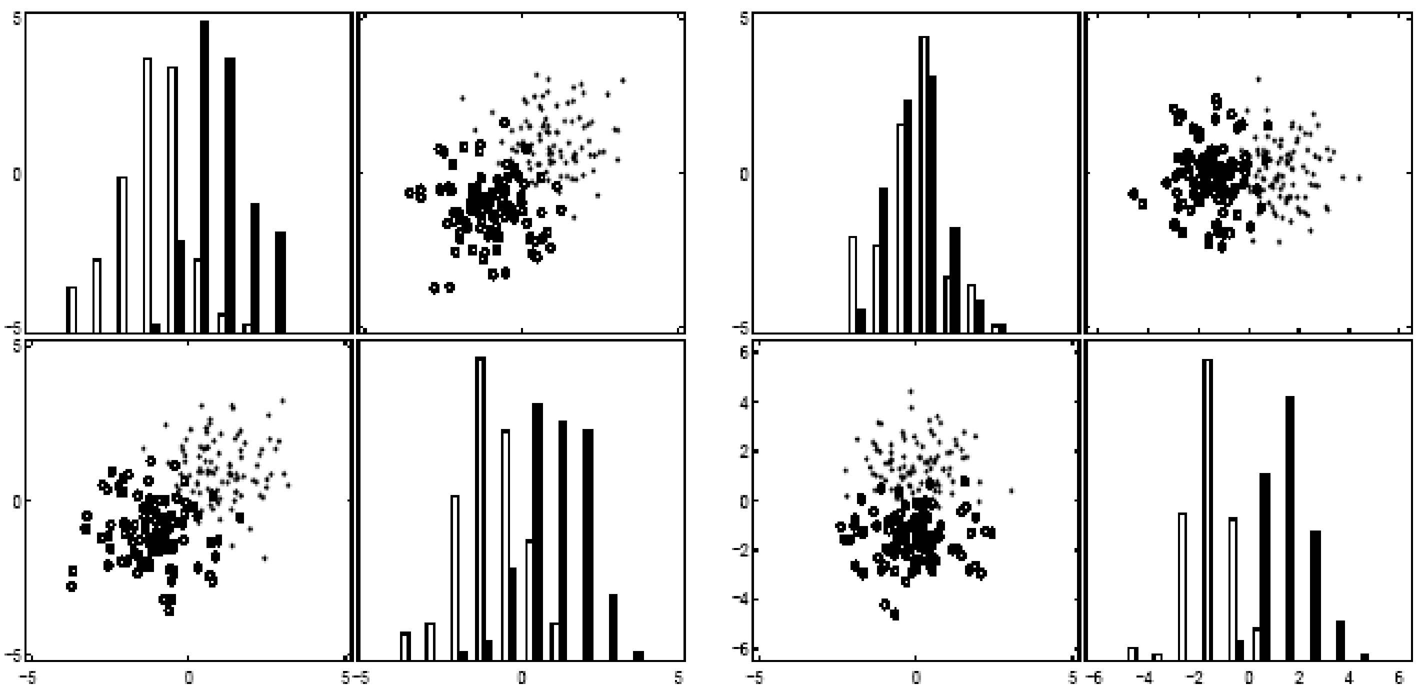
$$J(x_1) > J(x_2) > \cdots > J(x_d) > \cdots > J(x_D)$$

最好的d个特征组合在一起

问题？

为什么要特征选择？

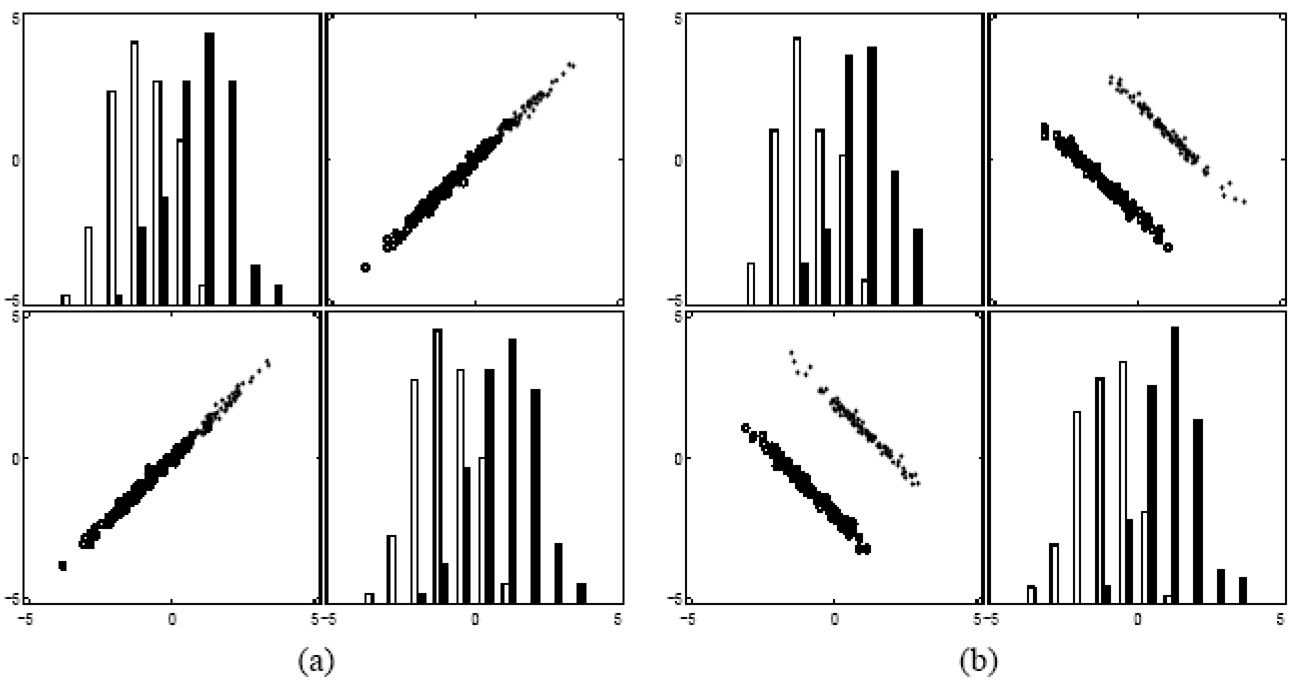
两个同样分布的变量，其包含的信息并不冗余。 样本(a)旋转45度后成(b)



(a) (b)

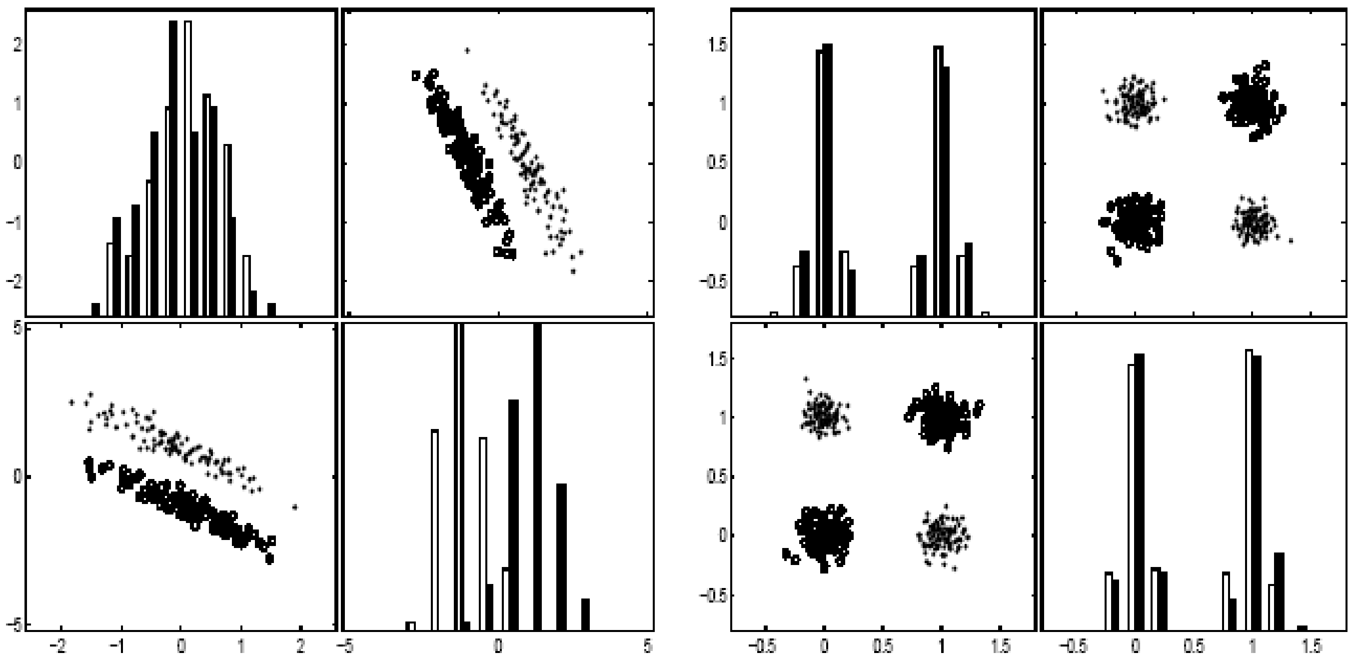
为什么要特征选择？

- **Perfectly correlated variables are truly redundant in the sense that no additional information is gained by adding them.**



为什么要特征选择？

- **Can a Variable that is Useless by Itself be Useful with Others?**



(a) (b)

特征选择-寻优算法

最优搜索法：分枝定界

次优搜索法：

a.单独最优特征组合

$$J(X)=\sum_{i=1}^{D}J(x_i) \qquad J(X)=\prod_{i=1}^{D}J(x_i)$$

b.顺序前进法

c.顺序后退法

次优搜索法：

d.增l减r法

e.模拟退火法

f. Tabu搜索法

g.遗传算法

Relief

输入： 训练集$X=\{\mathbf{x}_i \in \mathbb{R}^d\}_{i=1}^N$，随机选择的样本数$n$.

$S1$: 设定d维权重向量$\mathbf{w}=[w_1,\cdots,w_d]^T=\mathbf{0}$;

$S2$: for $i=1:n$

$S2a$: 从X中随机选择一个样本$\mathbf{x}$;

$S2b$: 计算X中$\mathbf{x}$最近的同类样本$\mathbf{h}$，不同类样本$\mathbf{m}$;

$S2c$: for $j=1:d$

$$w_j = w_j - \mathrm{diff}(j,\mathbf{x},\mathbf{h})/n + \mathrm{diff}(j,\mathbf{x},\mathbf{m})/n.$$

$S3$: 返回权重向量$\mathbf{w}$;

$S4$: 输出权重最大的前k个特征。

其中$\mathrm{diff}(j,\mathbf{x}_1,\mathbf{x}_2)$表示两个样本$\mathbf{x}_1,\mathbf{x}_2$在第$j$维上绝对值的差异。

Relief

对于离散变量： $$diff(j,x,h)=\begin{cases}0 & x_j=h_j\\1 & otherwise\end{cases}$$

对于连续变量： $$diff(j,x,h)=\frac{|x_j-h_j|}{x_{j\max}-x_{j\min}}$$

Extensions of RELIEF

- RELIEF series[3]
  - RELIEF-F: the widely used extension
    can handle noise data and multi-class problem
- RELIEF for regression[6]
  - RRELIEF, RRLIEF-F
- RELIEF as a decision tree splitting rule[4]
  - Myopic RELIEF

RELIEF-F (multi-class)

```
set all weights W[A] = 0.0
for i = 1 to m do
  begin
    randomly select an instance R
    find k nearest hits H_j
    for each class C ≠ class(R) do
      find k nearest misses M_j(C)
    for A = 1 to #attributes do
      W[A] = W[A] - Σ_{j=1}^{k} diff(A, R, H_j)/(m × k) +
      Σ_{C≠class(R)} [ P(C)/(1 - P(class(R))) Σ_{j=1}^{k} diff(A, R, M_j(C)) ]/(m × k)
  end
```

$$W[A] = W[A] - \sum_{j=1}^{k} \mathrm{diff}(A, R, H_j)/(m \times k) + \sum_{C \neq class(R)} \left[\frac{P(C)}{1 - P(class(R))} \sum_{j=1}^{k} \mathrm{diff}(A, R, M_j(C))\right]/(m \times k)$$

特征提取与选择中的过学习

- 样本太少
- 举例

References

[1] R. Kohavi and G. H. John. Wrappers for feature subset selection. Artificial Intelligence [J], 1997.

[2] Kira K and Rendell L. A practical approach to feature selection. In: ICML 1992

[3] I. Kononenko. Estimating attributes: Analysis and extensions of RELIEF. In ECML 94, pages 171–182, 1994.

[4] Kononenko, I., Simec, E., & Robnik- Sikonja, M. (1997). Overcoming the myopic of inductive learning algorithms with RELIEFF. Applied Intelligence [J], 1997.

[5] Molina L C, et al. Feature selection algorithms: a survey and experimental evaluation. ICDM 2002.

[6] Robnik-Sikonja M and Kononenko I. Theoretical and empirical analysis of RELIEF-F and RRELIEF-F. Machine Learning [J], 2003

[7] Ran Gilad-Bachrachy, et al. Margin Based Feature Selection - Theory and Algorithms, ICML 2004

[8] M. A. Hall and G. Holmes. Benchmarking attribute selection techniques for data mining, IEEE TKDE [J]. 2003.