第五章

近邻法

最近邻法

$$g_i(x) = \min_k \left\| x - x_i^k \right\|, k = 1, 2, \cdots, N_i$$

$$g_j(x) = \min_i g_i(x), i = 1, 2, \cdots, c$$

$$\Rightarrow x \in \omega_j$$

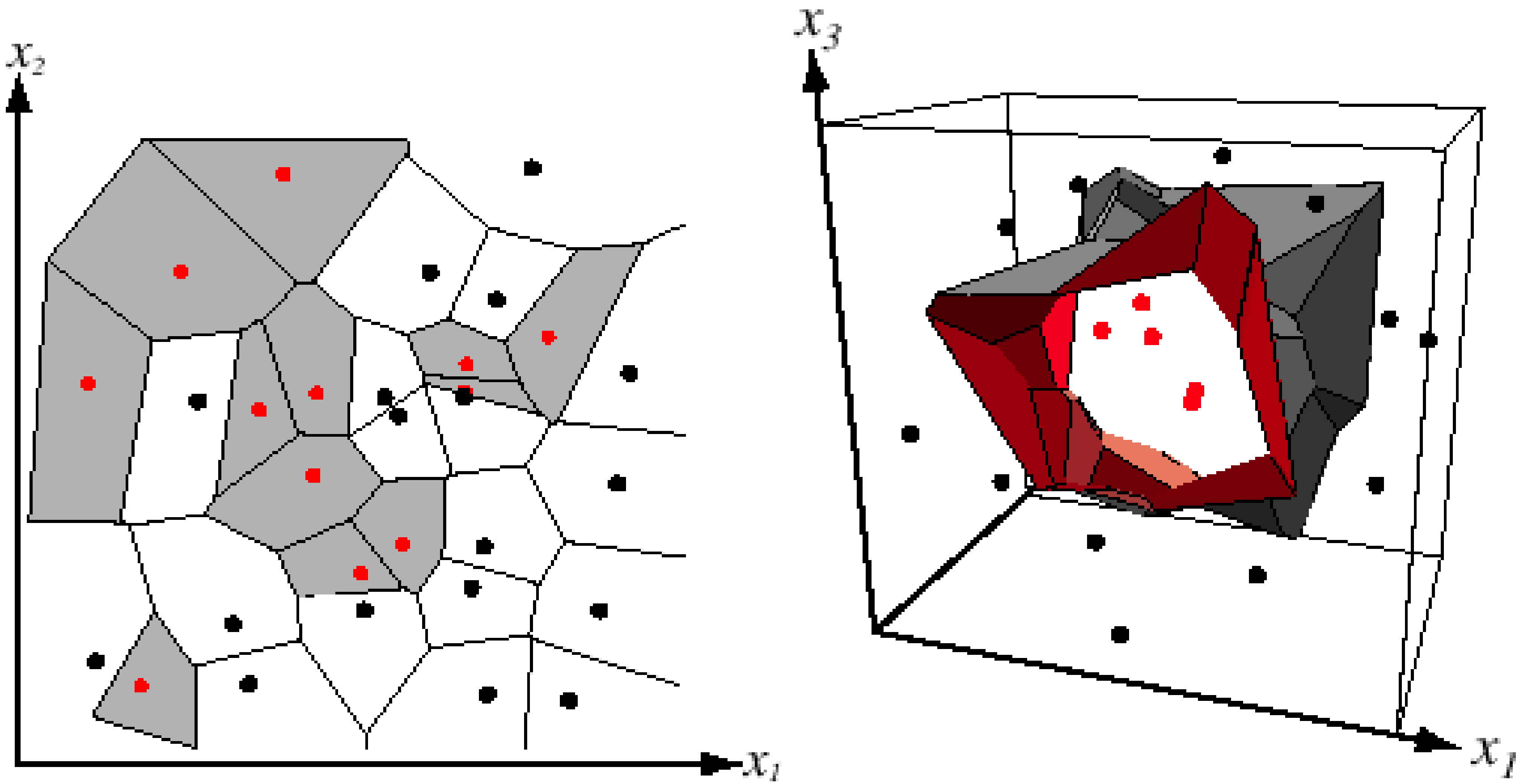


FIGURE 4.13. In two dimensions, the nearest-neighbor algorithm leads to a partitioning of the input space into Voronoi cells, each labeled by the category of the training point it contains. In three dimensions, the cells are three-dimensional, and the decision boundary resembles the surface of a crystal. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

错误率分析

决策：$P(\omega_m \mid x) = \max_i [P(\omega_i \mid x)]$

x 的错误率： $P^*(e \mid x) = 1 - P(\omega_m \mid x)$

总的错误率： $P^* = \int P^*(e \mid x) p(x) dx$

- N个样本时x的错误率（x'是x的最近邻）：

$$P_N(e \mid x) = \int P_N(e \mid x, x^{'}) p(x^{'} \mid x) dx^{'}$$

- N个样本时总的错误率：

$$P_N(e) = \int P_N(e \mid x) p(x) dx$$

$$P = \lim_{N \to \infty} P_N(e)$$

假定$p(x)>0$，连续。S：以x为中心的超球

$$P_s = \int_{x' \in s} p(x')dx'$$

$x_1, x_2, \cdots, x_N$, N 个样本落在超球S外的概率

$$P(x_1, x_2, \cdots, x_N) = (1 - P_s)^N$$

$$\lim_{N \to \infty} p(x' \mid x) = \delta(x' - x)$$

$$P_N(e|x,x') = 1 - \sum_{i=1}^{c} P(\theta = \omega_i, \theta' = \omega_i | x, x')$$

$$= 1 - \sum_{i=1}^{c} P(\omega_i | x) P(\omega_i | x')$$

抽取x ‘时与x的类别无关

$$\lim_{N \to \infty} P_N(e|x,x') = 1 - \sum_{i=1}^{c} P^2(\omega_i | x)$$

$$\lim_{N\to\infty} P_N(e \mid x) = \lim_{N\to\infty} \int P_N(e \mid x, x^{'})P(x^{'} \mid x)dx^{'}$$

$$= \int [1 - \sum_{i=1}^{c} P^2(\omega_i \mid x)]\delta(x^{'} - x)dx^{'}$$

$$= 1 - \sum_{i=1}^{c} P^2(\omega_i \mid x)$$

$$P = \lim_{N \to \infty} P_N(e) = \lim_{N \to \infty} \int P_N(e \mid x) p(x) dx$$

$$= \int \lim_{N \to \infty} P_N(e \mid x) p(x) dx$$

$$= \int [1 - \sum_{i=1}^{c} P^2(\omega_i \mid x)] p(x) dx$$

下界：$P^* \le P$

1、 $$P(\omega_m \mid x) = 1$$

$$P = \int [1-1] p(x) dx = 0$$

$$P^* = \int P^*(e \mid x) p(x) dx$$
$$= \int [1 - P(\omega_m \mid x)] p(x) dx = 0$$

2、 $P(\omega_i \mid x) = 1/c$

$$P = \int [1 - \sum_{i=1}^{c} (\frac{1}{c})^2] p(x) dx = \frac{c-1}{c}$$

$$P^* = \int (1 - \frac{1}{c}) p(x) dx = \frac{c-1}{c}$$

上界：

$$\sum_{i=1}^{c} P^2(\omega_i \mid x) = P^2(\omega_m \mid x) + \sum_{i \neq m} P^2(\omega_i \mid x)$$

当 $P^{'}(\omega_i \mid x) = A, i = 1, 2, \cdots, c; i \neq m$

$$\sum_{i=1}^{c} P^{'2}(\omega_i \mid x) = \min \sum_{i=1}^{c} P^2(\omega_i \mid x)$$

$$\sum_{i \neq m} P'(\omega_i \mid x) = (c-1)P'(\omega_i \mid x) = P^*(e \mid x)$$

$$P'(\omega_i \mid x) = \begin{cases} \dfrac{P^*(e \mid x)}{c-1}, i \neq m \\ 1 - P^*(e \mid x), i = m \end{cases}$$

$$\sum_{i=1}^{c} P^2(\omega_i \mid x) = P^2(\omega_m \mid x) + \sum_{i \neq m} P^2(\omega_i \mid x)$$

$$\geq [1 - P^*(e \mid x)]^2 + \sum_{i \neq m} \frac{P^{*2}(e \mid x)}{(c-1)^2}$$

$$= 1 - 2P^*(e \mid x) + \frac{c}{c-1} P^{*2}(e \mid x)$$

$$\Rightarrow 1 - \sum_{i=1}^{c} P^2(\omega_i \mid x) \leq 2P^*(e \mid x) - \frac{c}{c-1} P^{*2}(e \mid x)$$

$$E[P^*(e\,|\,x)] = P^*$$

$$Var[P^*(e\,|\,x) = \int [P^*(e\,|\,x) - P^*]^2 p(x)dx$$

$$= \int [P^{*2}(e\,|\,x)p(x) - 2P^*(e\,|\,x)P^* p(x) + P^{*2} p(x)]dx$$

$$= \int P^{*2}(e\,|\,x)p(x)dx - P^{*2} \geq 0$$

$$\Rightarrow P^{*2} \leq \int P^{*2}(e\,|\,x)p(x)dx$$

$$P = \int [1 - \sum_{i=1}^{c} P^2(\omega_i \mid x)] p(x) dx$$

$$\leq \int [2P^*(e \mid x) - \frac{c}{c-1} P^{*2}(e \mid x)] p(x) dx$$

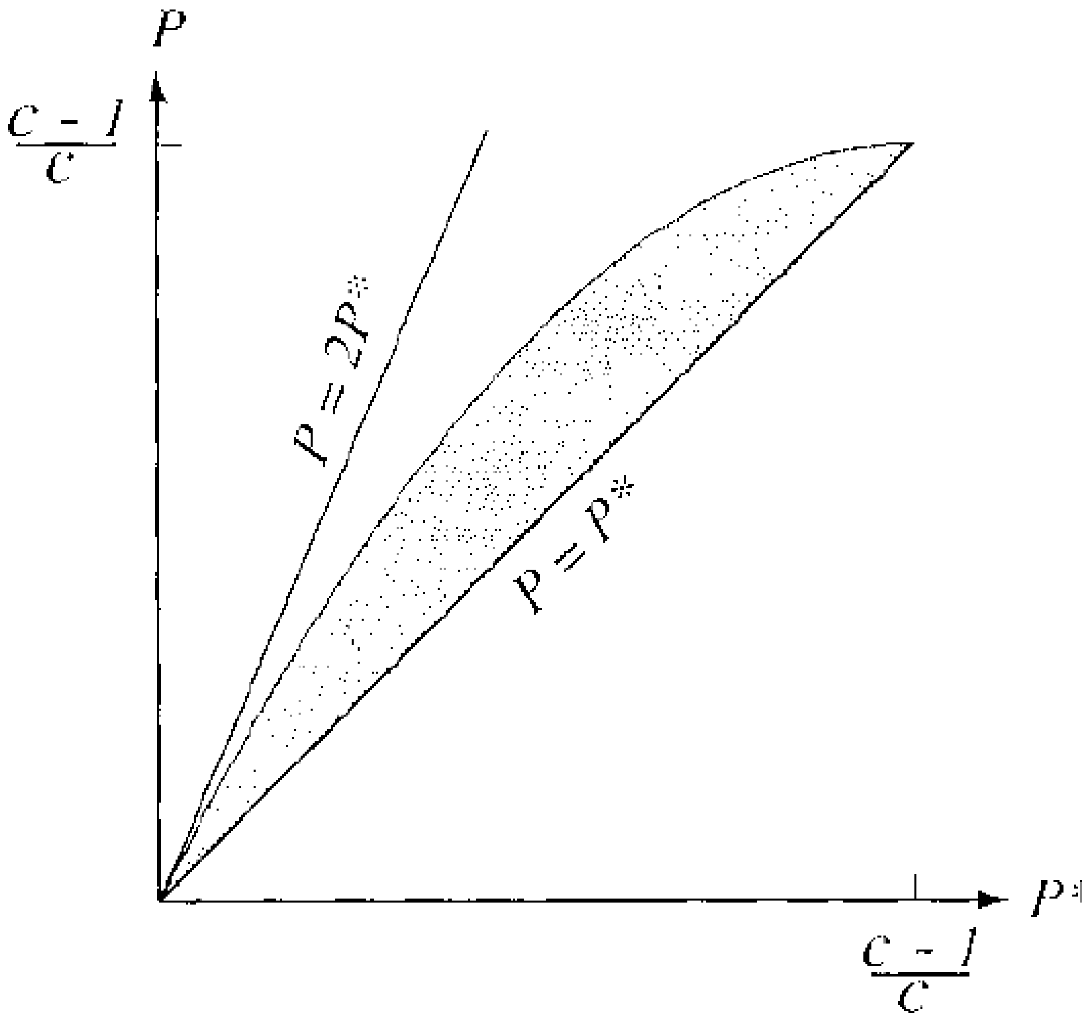
$$= 2\int P^*(e \mid x) p(x) dx - \frac{c}{c-1} \int P^{*2}(e \mid x) p(x) dx$$

$$\leq 2P^* - \frac{c}{c-1} P^{*2} \qquad = P^*(2 - \frac{c}{c-1} P^*)$$

$$P^* \le P \le P^*(2 - \frac{c}{c-1} P^*)$$

$$P^* \le P \le 2P^*$$

近邻法的错误率



K 近邻法

$$g_i(x) = k_i, i = 1, 2, \cdots, c$$

k_i： K个样本中属于第i类的样本数

$$g_j(x) = \max_i k_i \Rightarrow x \in \omega_j$$

k的选择与错误率的关系

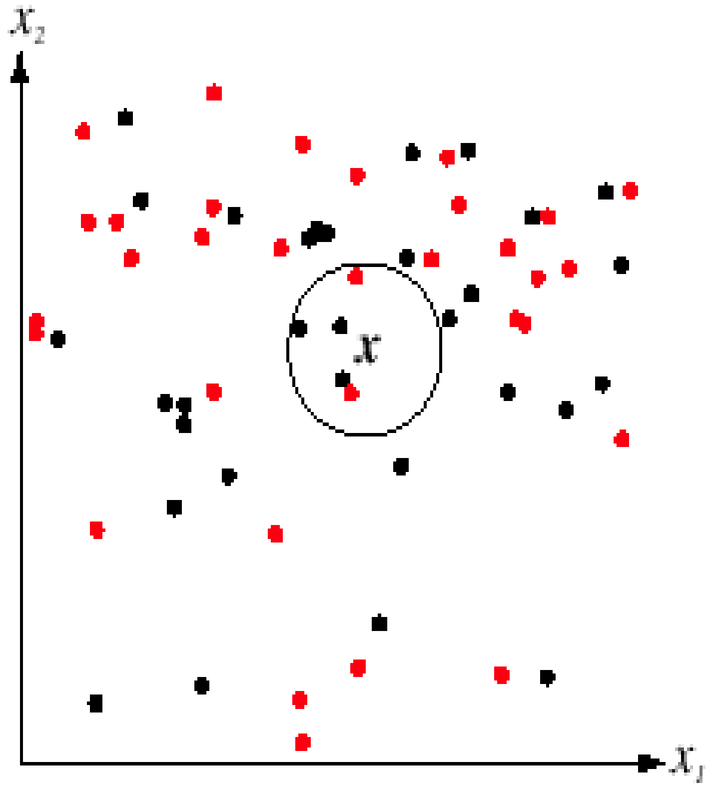
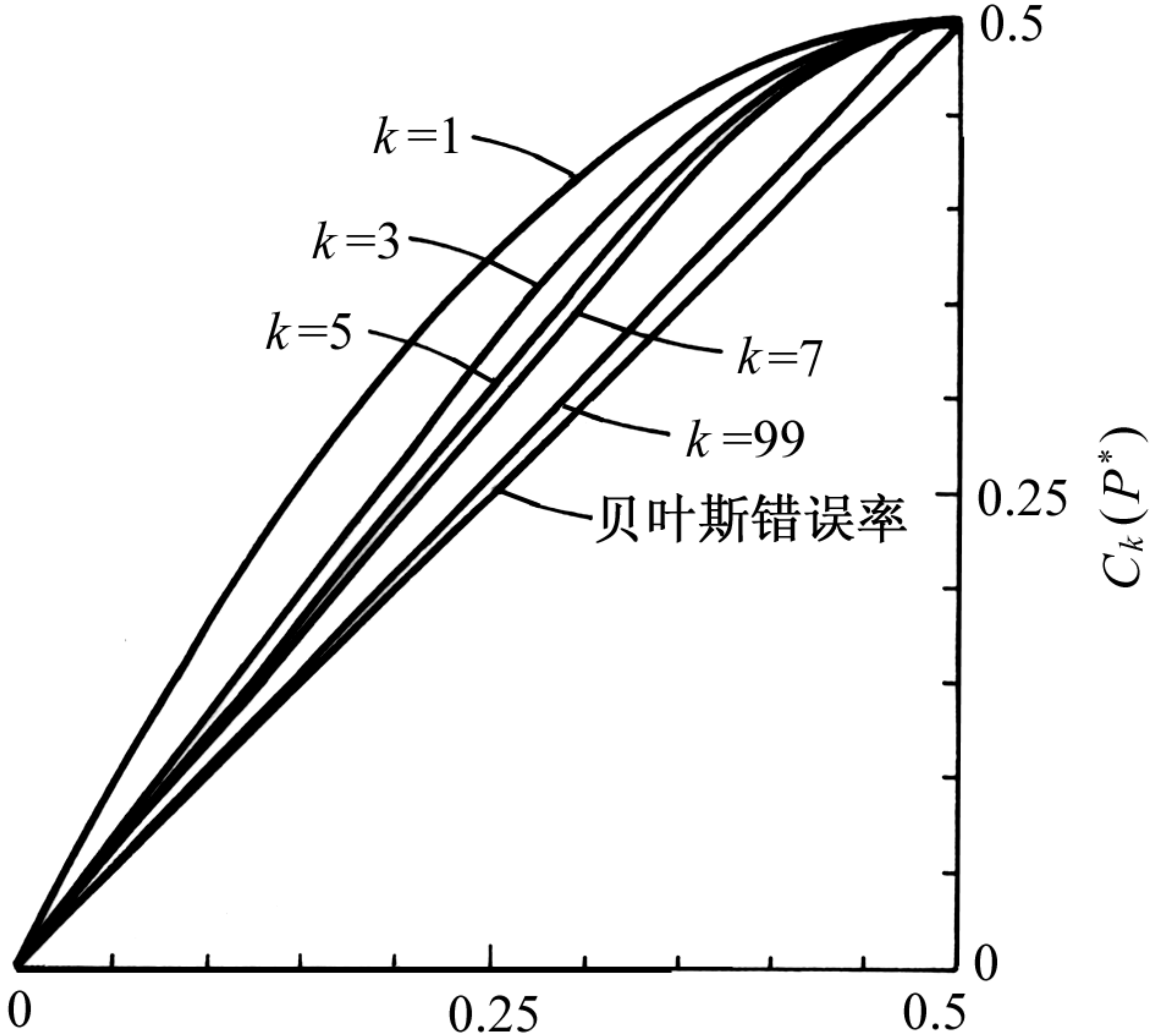


FIGURE 4.15. The k-nearest-neighbor query starts at the test point **x** and grows a spherical region until it encloses k training samples, and it labels the test point by a majority vote of these samples. In this $k = 5$ case, the test point **x** would be labeled the category of the black points. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

$$P^* \le P \le P^*(2 - \frac{c}{c-1} P^*)$$

K近邻法的错误率



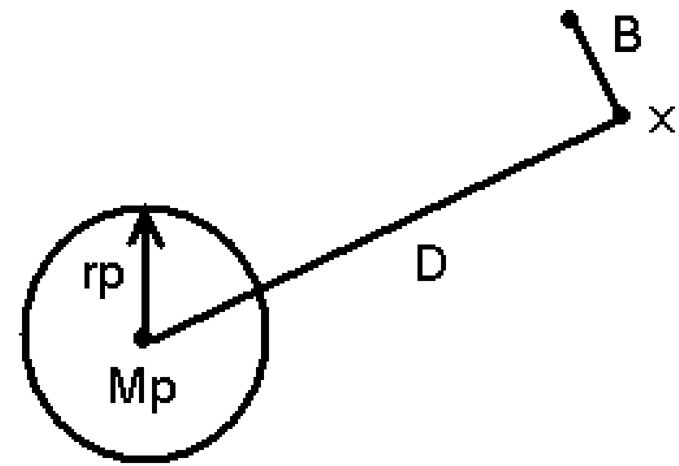
近邻法的快速方法

- 原来方法：对x计算N个距离
- 解决方法：判决时减少计算量

  判决前提前计算

1、 $B + r_p < D(x, M_p)$

2、 $B + D(x_i, M_p) < D(x, M_p)$



step1：分解样本集（聚类）

记录每一级中每一聚类样本的均值，以及每一样本与均值的距离，每一聚类中样本与均值的最大距离。

step2：近邻法实施

时刻保存x的最近邻 x' 及 $D(x,x')$ 分级权与每级中聚类数目 。

K近邻：保存x的K个近邻及距离值 。

压缩近邻法

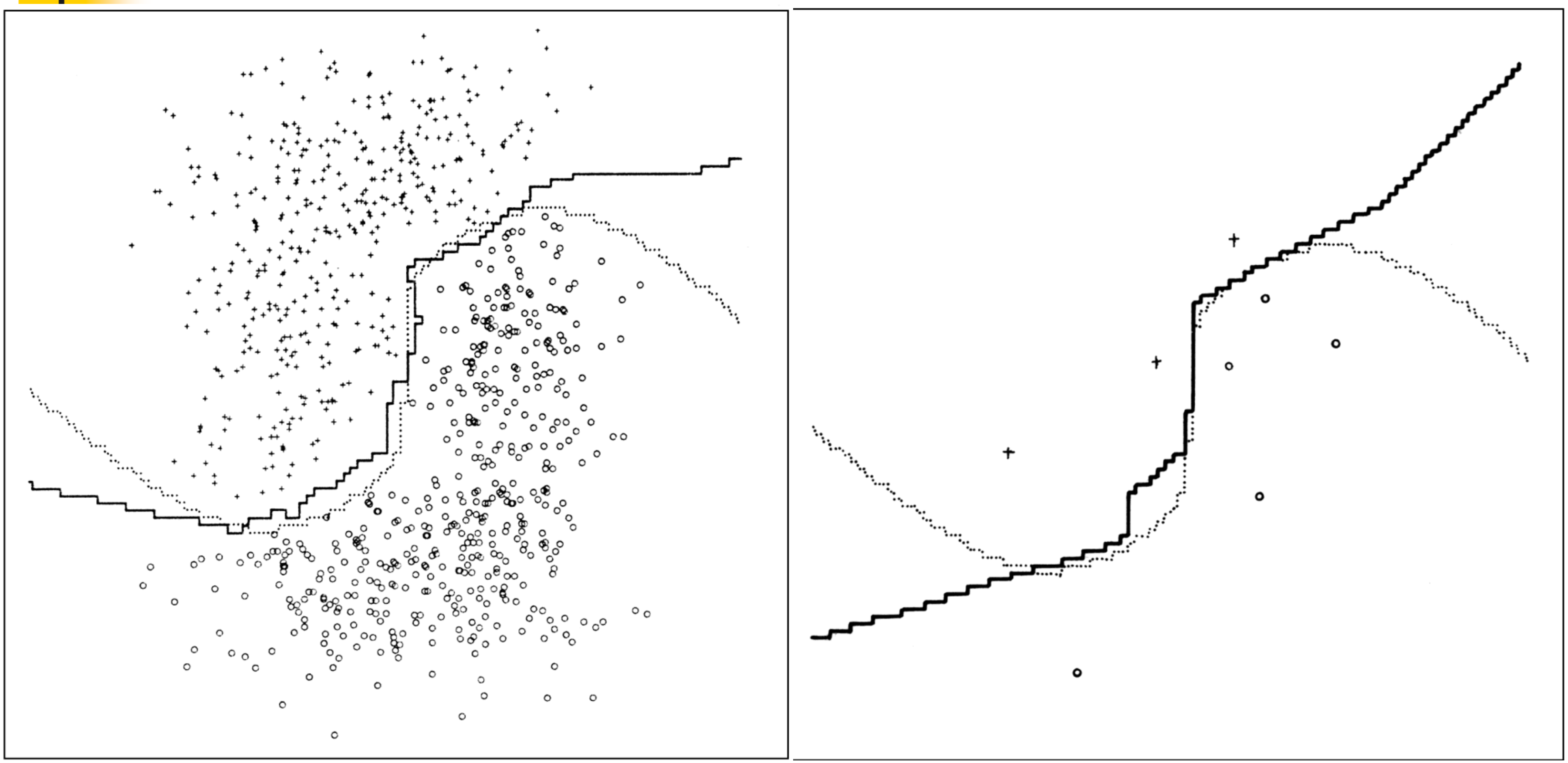
step1：从Grabbag中选择一个样本放入store中；

Step2：用store中样本以近邻法测试Grabbag中样本。如果分错，则将该样本放入Store。

Step3：重复上面方法直到Grabbag中没有样本再转到Store中，或Grabbag为空则停止。

Step4：用Store中样本作为近邻法设计集。

压缩近邻法结果



各种距离度量

1、s 阶 Minkowski 度量

$$D_M(\mathbf{x},\mathbf{y})=[\sum_{j=1}^{d}|\mathbf{x}_j-\mathbf{y}_j|^s]^{1/s}$$

s=1时，$D_c(\mathbf{x},\mathbf{y})=\sum_{j=1}^{d}|\mathbf{x}_j-\mathbf{y}_j|$

2、欧氏距离度量（ s=2时）

$$D_E(\mathbf{x},\mathbf{y})=[\sum_{j=1}^{d}(\mathbf{x}-\mathbf{y})^2]^{1/2}=[(\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y})]^{1/2}$$

各种距离度量

3、Chebychev距离

$$\delta_T(x_k, x_l) = \max_j \left| x_{kj} - x_{lj} \right|$$

4、平方距离：Q正定标尺矩阵

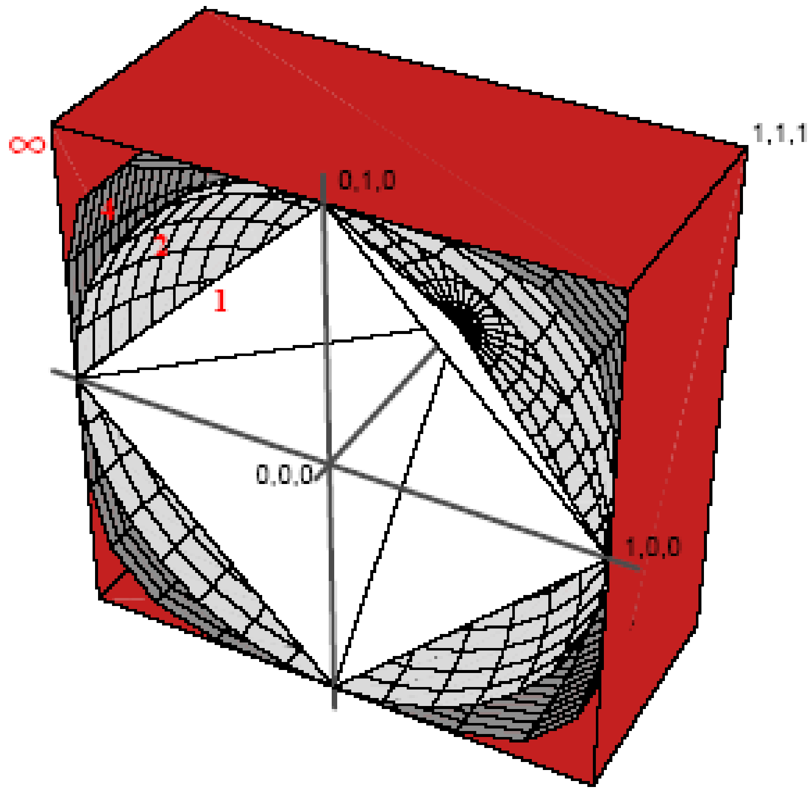
$$\delta_Q(x_k, x_l) = (x_k - x_l)^T Q(x_k - x_l)$$

问题

- 几种距离度量之间的关系?

问题

- 几种距离度量之间的关系?



各种距离度量

- 其他距离度量?
- 相似性与不相似性

相似性度量举例

- 文本分类
- 问题：长文本和短文本
- 问题：亮图像和暗图像
- 特征的归一化（normalization）

两个概率分布的相似性

$$J_B = -\ln \int [p(x|\omega_1)p(x|\omega_2)]^{1/2} dx$$

$$J_C = -\ln \int p^s(x|\omega_1)p^{1-s}(x|\omega_2)dx$$

KL散度：

$$J_D = \int_X [p(x|\omega_i) - p(x|\omega_j)] \ln \frac{p(x|\omega_i)}{p(x|\omega_j)} dx$$

正态分布且 $\Sigma_i = \Sigma_j = \Sigma$时，

$$8J_B = J_D = (\mu_i - \mu_j)^T \sum^{-1} (\mu_i - \mu_j) = J_M$$

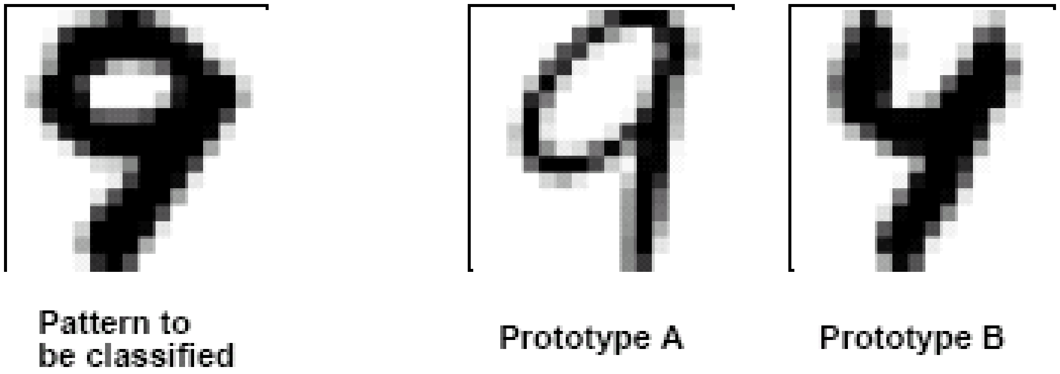
$$D(f_1, f_2) = \int f_1(x) \log \frac{f_1(x)}{f_2(x)} dx$$

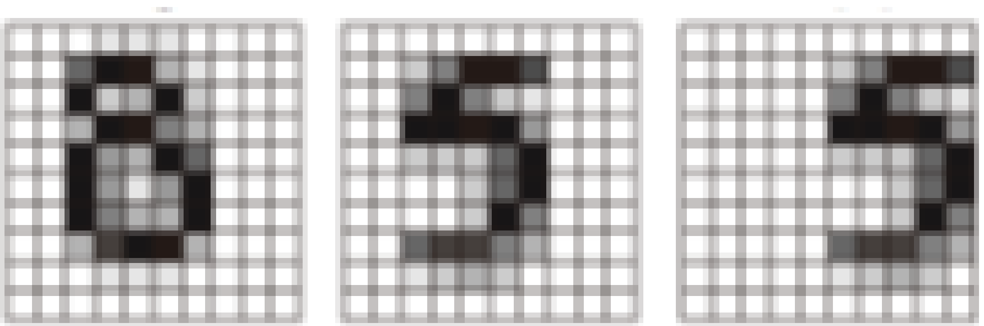
$$J_D = D(f_1, f_2) + D(f_2, f_1)$$

Tangent distance in visual patterns

手写数字识别

平移，旋转，尺度，线条粗细

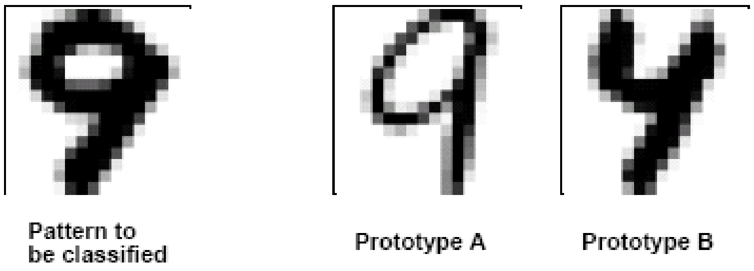


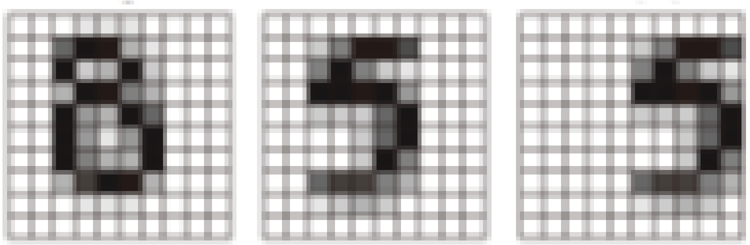


Tangent distance in visual patterns

解决方法：

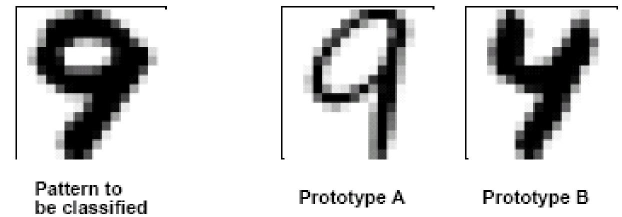
- 通过技术手段减少同类样本之间的差异：图像重心归一化消除平移的影响
- 生成大量新样本
- 距离度量的选择

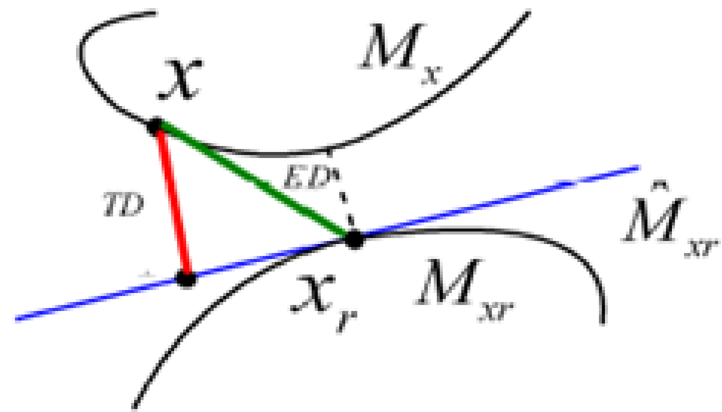


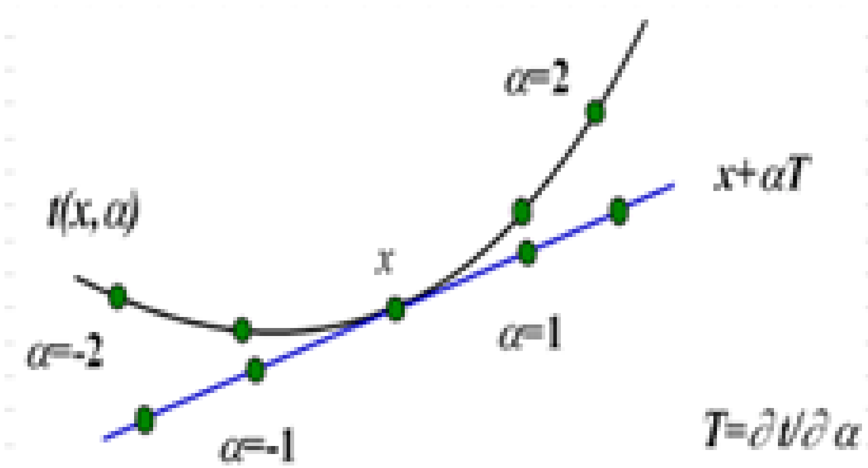


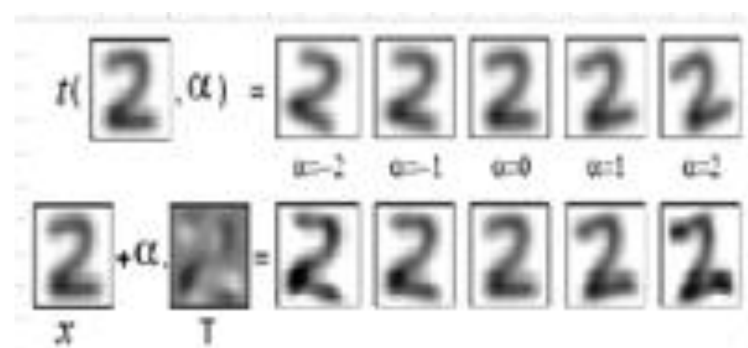
Tangent distance in visual patterns

- 样本微小变化构成流形
- 样本到流形的距离
- 流形的局部线性近似
- 计算切距离







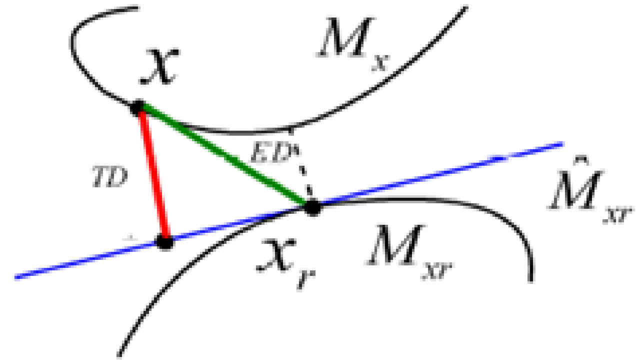


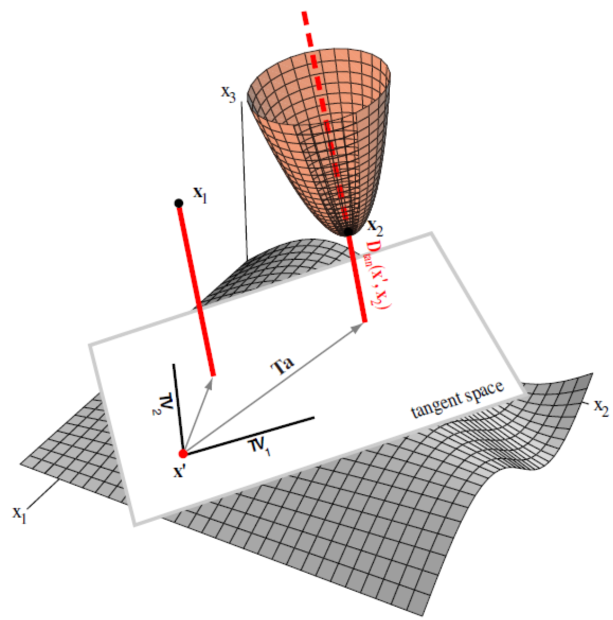
Tangent distance in visual patterns

切距离的计算

- 原型样本点 $\mathbf{x}^t$都进行每一种变换操作 $F_i(\mathbf{x}^t)$，代表图像经过第i种变换（如旋转固定角度）得到的新图像。构造切向量

- 对每一个原型样本点 $\mathbf{x}^t$ ，可以构造 $r \times d$ 的矩阵 $\mathbf{T}$，$\mathbf{T}$ 由 $\mathbf{x}^t$ 处的切向量组成



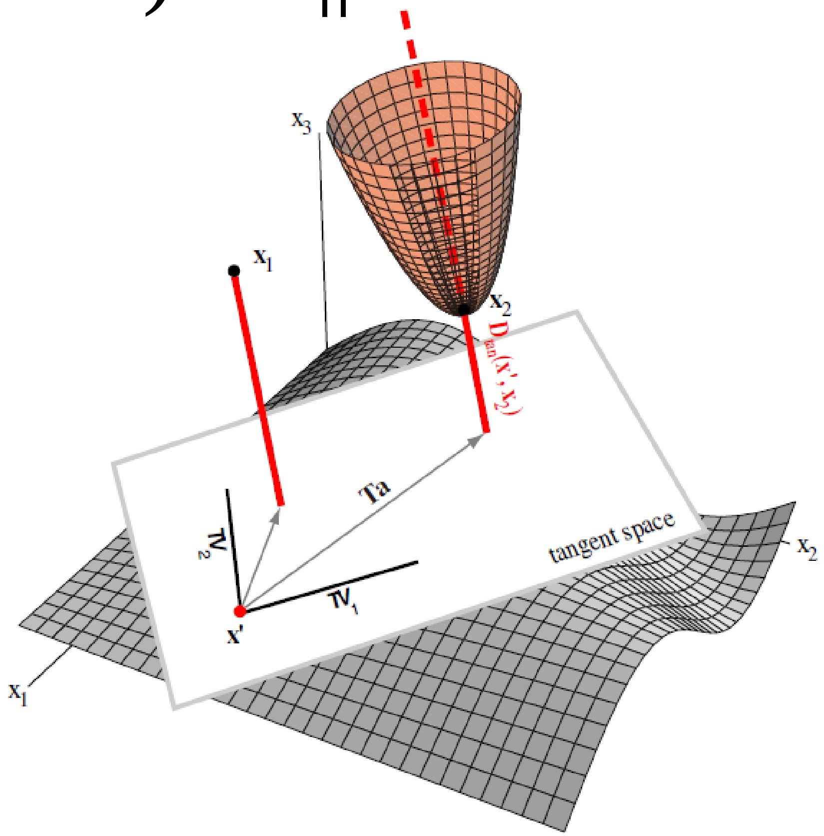


Tangent distance in visual patterns

切距离的计算

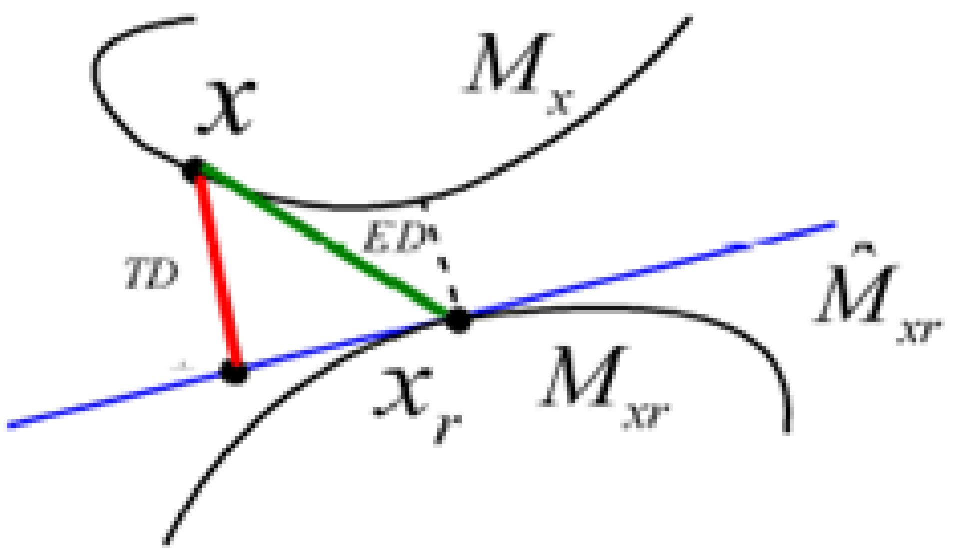
$$D(\mathbf{x},\mathbf{x}') = \min_a \| (\mathbf{x}' + a\mathbf{T}) - \mathbf{x} \|$$

- 目标函数是二次型
- 简单的搜索算法，比如：迭代梯度下降法



Tangent distance in visual patterns

- 单边(one-side)切距离
- 双边（two-side）切距离



Tangent distance in visual patterns

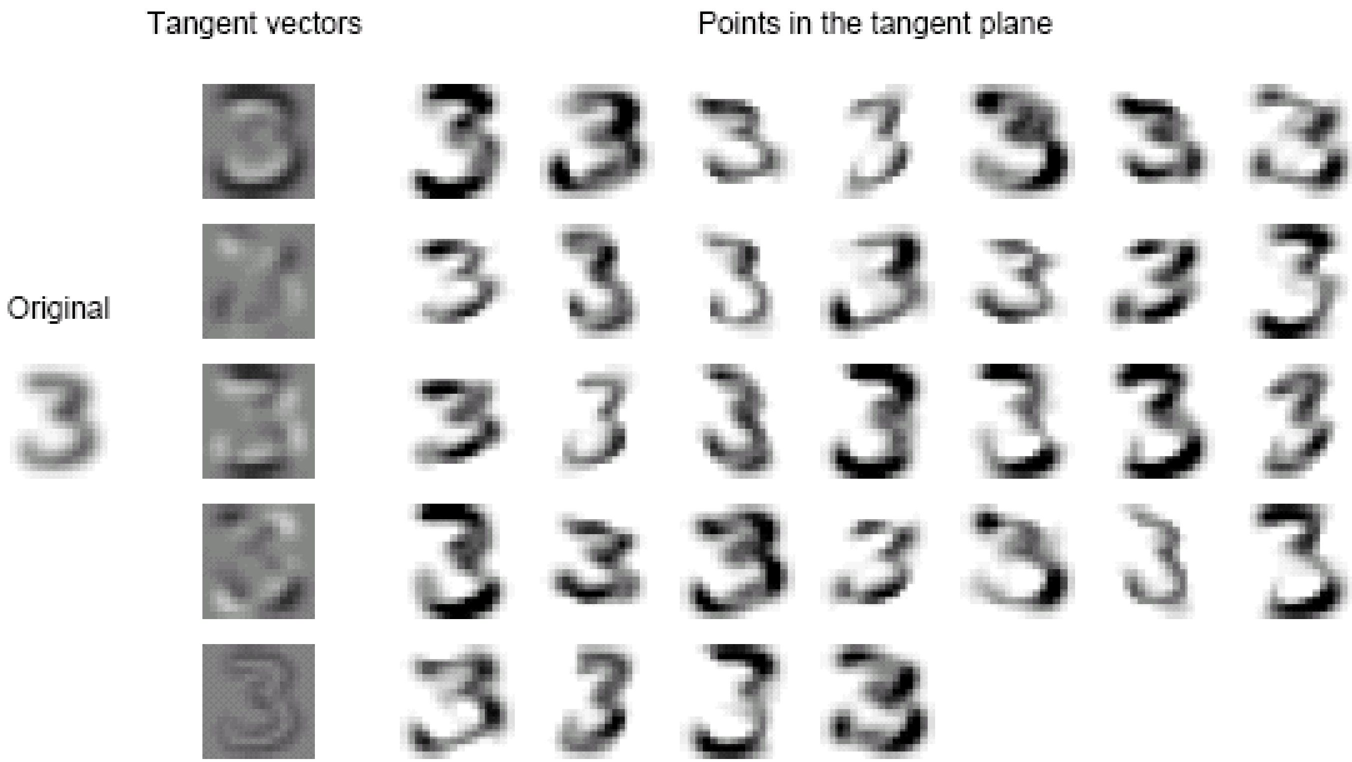


Fig. 6. Left: Original image. Middle: 5 tangent vectors corresponding respectively to the 5 transformations: scaling, rotation, expansion of the X axis while compressing the Y axis, expansion of the first diagonal while compressing the second diagonal and thickening. Right: 32 points in the tangent space generated by adding or subtracting each of the 5 tangent vectors.

近邻法的其它工作

- Kernel Nearest Neighbor
- 近邻线方法
- …

文献

- Patrice Y. Simard, Yann A. Le Cun, John S. Denker, Bernard Victorri, “*Transformation Invariance in Pattern Recognition -- Tangent Distance and Tangent Propagation*”. In Neural Networks: Tricks of the Trade, G. B. Orr and K-R Muller (Eds), Chapter 12, Springer, 1998.

Exercise

$$\Sigma = CBC^T$$
$$= CB_1 B_1^T C^T$$
$$= CB_1 (CB_1)^T$$
$$= A \cdot A^T$$
$$\Sigma^{-1} = (A^T)^{-1} \cdot A^{-1}$$

$$B = \begin{pmatrix} \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_d \end{pmatrix}$$

$$B_1 = \begin{pmatrix} \sqrt{\lambda_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sqrt{\lambda_d} \end{pmatrix}$$

Exercise

$$r^2(a,b) = (a-b)^T (A^T)^{-1} \cdot A^{-1}(a-b)$$

$$= [A^{-1}(a-b)]^T [A^{-1}(a-b)]$$

$$= [A^{-1}a - A^{-1}b]^T [A^{-1}a - A^{-1}b]$$

$$= (a'-b')^T (a'-b')$$

$$= \|a'-b'\|^2 \leftarrow \text{欧氏距离}$$

Exercise

$$a \to A^{-1}a = a'$$

$$r(a,b) = \|a' - b'\|$$

$$r(a,b) = \|a' - b'\| = \|b' - a'\| = r(b,a)$$

当且仅当$a' = b'$时$\|a' - b'\| = 0 = r(a,b)$

$$r(a,c) = \|a' - c'\|$$

$$\le \|a' - b'\| + \|b' - c'\| = r(a,b) + r(b,c)$$

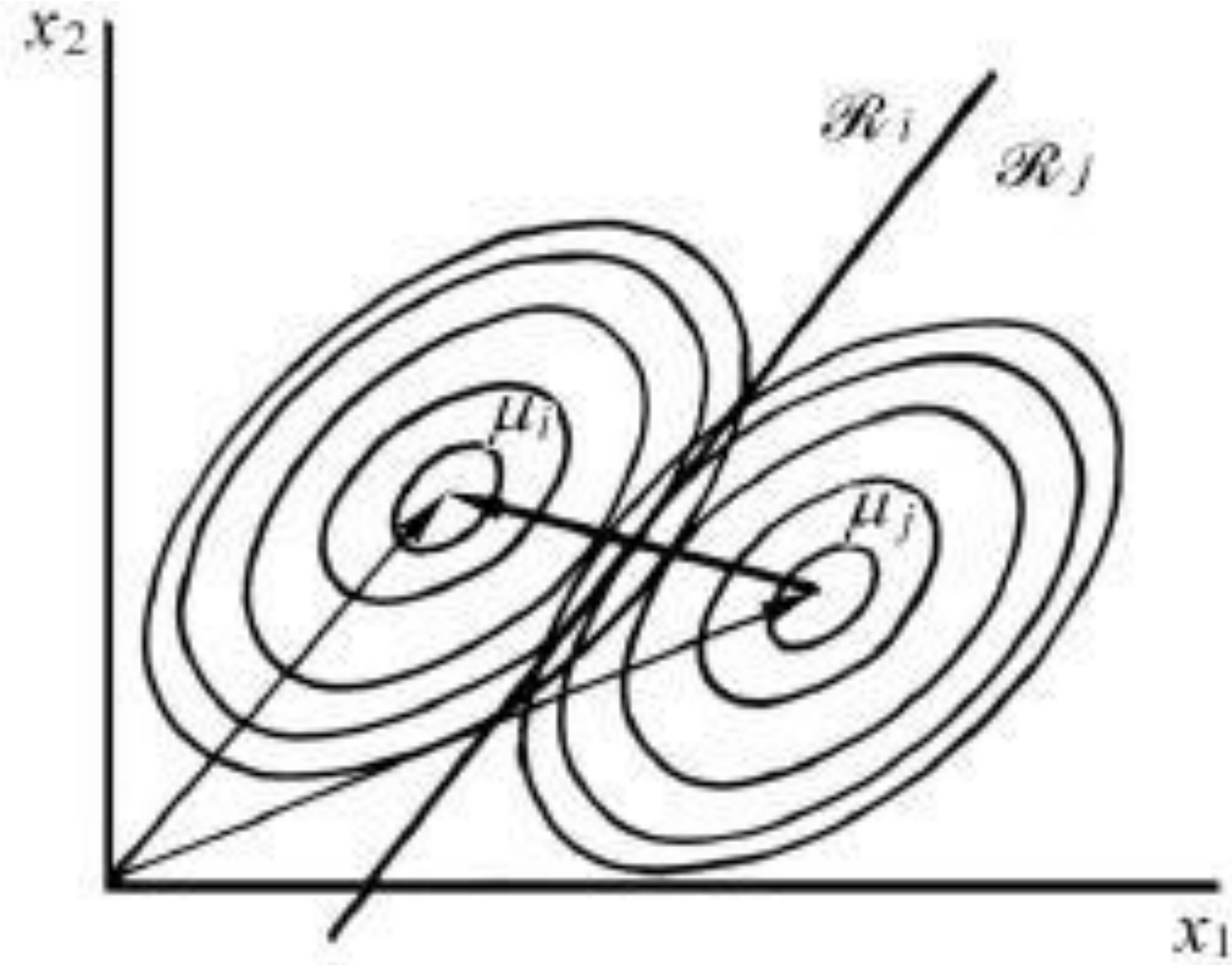
Exercise

$$\sum^{-1}: \sum > 0, 正定?$$

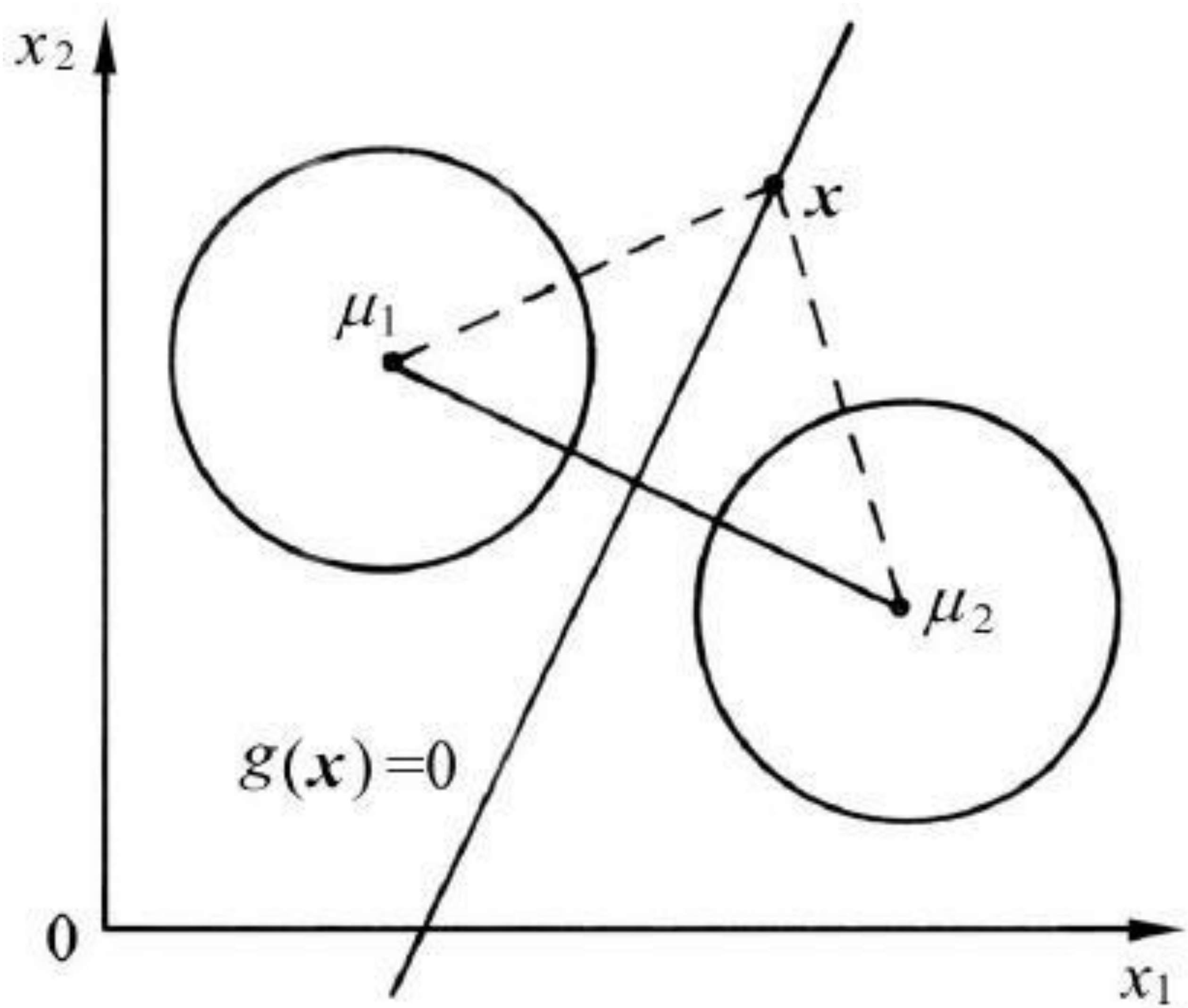
几何意义

- 协方差矩阵是对角阵
- 对角元素的大小
- 马氏距离与欧式距离的关系
- 样本距离对样本分布的依赖性

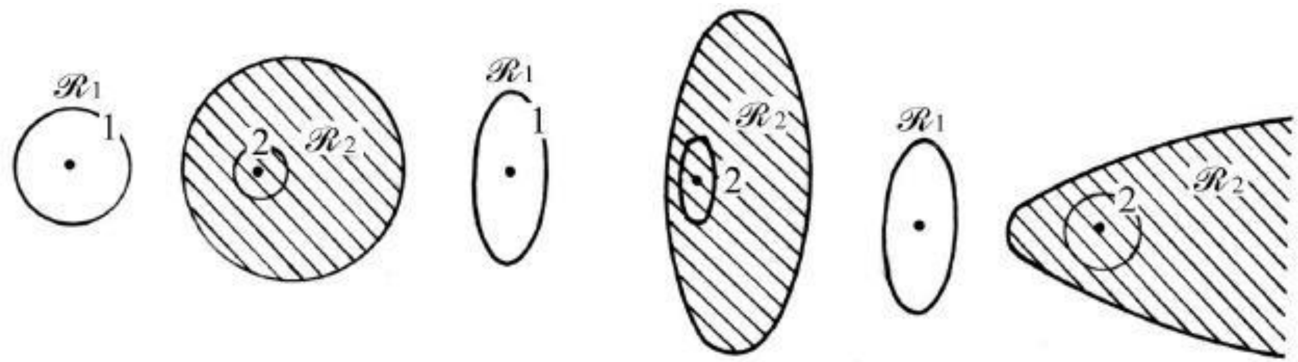
几何意义



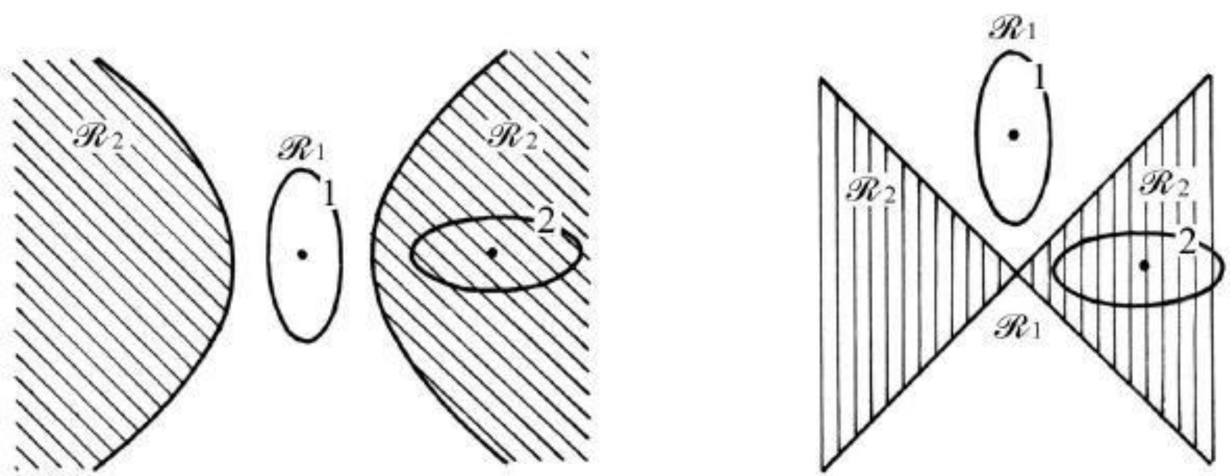
几何意义



几何意义



(a) 圆　　(b) 椭圆　　(c) 抛物线



(d) 双曲线　　(e) 直线