第五章　非线性方法

5-1 近邻法

引言

贝叶斯分类需要知道样本的概率分布，但估计样本分布有时并不容易

直接对数据进行划分的方法：

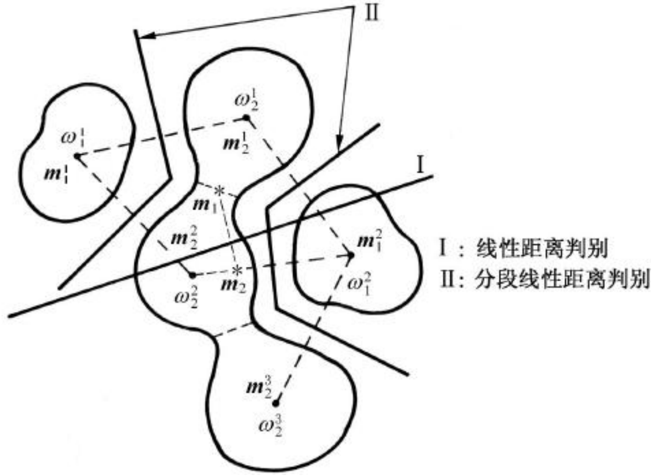
- 线性方法：简单、实用、经济，但数据不满足线性可分条件时错误可能大
- 非线性方法：解决线性不可分问题

分段线性判别函数 (piecewise linear discriminant functions)

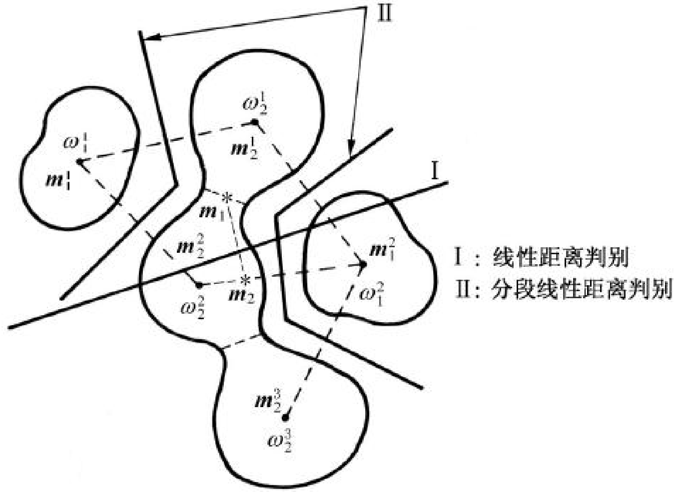
第二章提到的多类线性判别函数实际就是分段线性判别函数。

思路：

如果两类可以划分为线性可分的若干子类,则可以设计多个线性分类器，实现分段线性分类器。



最简单的分段线性分类器：把各类划分为若干子类，以子类中心作为类别代表点，考查新样本到各代表点的距离并将它分到最近的代表点所代表的类。



极端情况，将所有样本都作为代表点 ➔ 近邻法（Nearest-Neighbor method）

最近邻法

样本集 $S_N=\left\{(x_1,\omega_1),(x_2,\omega_2),\cdots,(x_N,\omega_N)\right\}$

x_i：样本，ω_1：类别标号，$\omega_i=\left\{1,2,\cdots,c\right\}$

样本 x_i 与 x_j 之间的距离 $\delta(x_i,x_j)$：比如欧氏距离 $\left\|x_i-x_j\right\|$

对未知样本 x，求 S_N 中与之距离最近的样本 x'，（类别为 ω'）

$$\delta(x,x')=\min_{j=1,\cdots,N}\delta(x,x_j)$$

则将 x 分到 ω' 类（或记作 $\hat{\omega}_1(x)$）

—— 最近邻决策（一近邻决策）

最近邻法的错误率（渐近分析）

近似表示为：$P \leq P \leq 2P^*$

其中：P^*：贝叶斯错误率

P：样本无穷多时最近邻法的错误率（渐近平均错误率）

前提：样本集独立同分布

证明详见参考文献： T. Cover & P. Hart, Nearest neighbor pattern classification，IEEE Transactions on Information Theory, 1967, 13(1):21-27

k-近邻法（kNN）

最近邻法（一近邻法）的推广：

找出 x 的 k 个近邻，看其中多数属于哪一类，则把 x 分到哪一类。

一般表示： c 类 ω_i， $i = 1, \cdots, c$， N 个样本。

k_i， $i = 1, \cdots, c$ 为 x 的 k 个近邻中属于 ω_i 的样本数

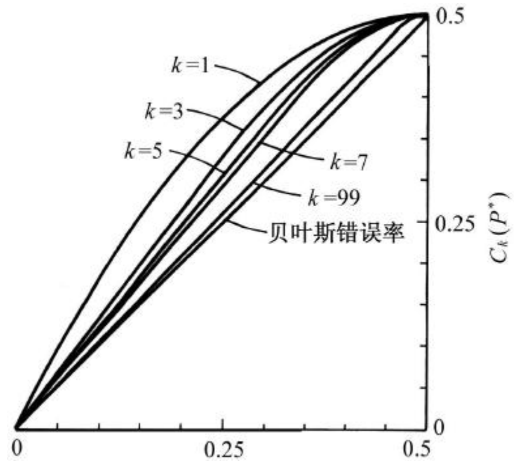
判别函数： $g_i(x) = k_i$， $i = 1, \cdots, c$

决策规则： if $g_j(x) = \max\limits_{i=1,\cdots,c} k_i$， then $x \in \omega_j$

KNN demo：http://vision.stanford.edu/teaching/cs231n-demos/knn/

渐近平均错误率的界：

N 无穷大时，k 越大，P_k 的上限越低（越靠近下限）。但 k 应始终是 N 中的一小部分，保证 k 个近邻均充分接近 x。否则这一关系不成立。



一般来说，总有 $P^* \le P_k \le 2P^*$

结论：原理简单，性能优异。

有没有问题？

问题

① 存储量和计算量

② 票数接近时风险较大，有噪声时风险加大

③ 样本无穷多时性能优异，有限样本下性能如何？

改进：

① 减少计算量和存储量

② 引入拒绝机制

③ 根据实际问题修正投票方式

如加权投票，否决票等

如距离加权，考虑样本比例及先验概率等

近邻法的快速算法

近邻法在计算上的问题：$\begin{cases}\text{需存储所有训练样本}\\ \text{新样本需与每个样本做比较}\end{cases}$

快速算法基本思想：

把样本集分级分成多个子集（树状结构）

每个子集（结点）可用较少几个量代表

通过将新样本与各结点比较排除大量候选样本

只有最后的结点（子集）中逐个样本比较，找出近邻

基本算法：分支定界算法（Branch-Bound Algorithm）

> □□□□□□□□□:
> □□:□□□□□□□□□□□□□□□□□□□□□;
> □□:□□□□□□□□□□□□□□□□□□□;
> □□·□□□□□□ □□□□□□□□□□□□□□□□□□□□□
> □□□□,□□□□□□□□□□□。

符号约定：

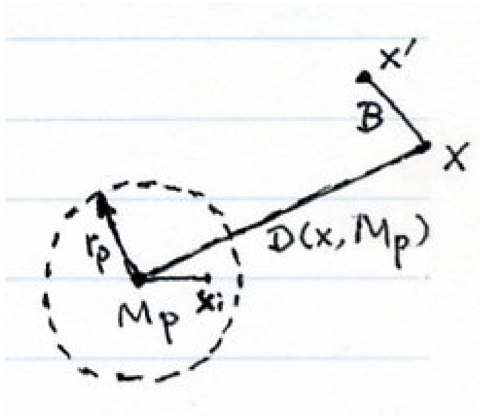
$\mathcal{X}_p$：结点 p 对应的样本子集

N_p：$\mathcal{X}_p$ 中的样本数

M_p：子集 $\mathcal{X}_p$ 中的样本均值（中心点）

$r_p = \max\limits_{x_i \in \mathcal{X}_p} D(x_i, M_p)$：$\mathcal{X}_p$ 中离中心点最远的距离

B：当前搜索到的最近邻距离



规则：1. 对新样本x，结点x_p

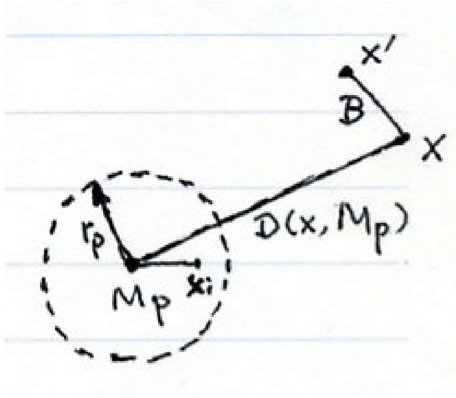
若$D(x,M_p) > B + r_p$

则x的近邻不可能在x_p中

2. 对新样本x，结点p中的样本$x_i \in \mathcal{X}_p$

若$D(x,M_p) > B + D(x_i,M_p)$

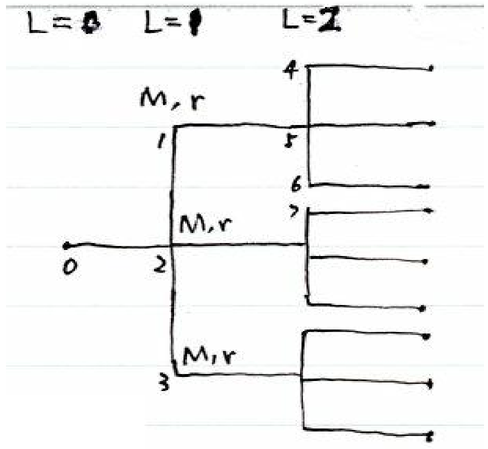
则x_i不是x的最近邻



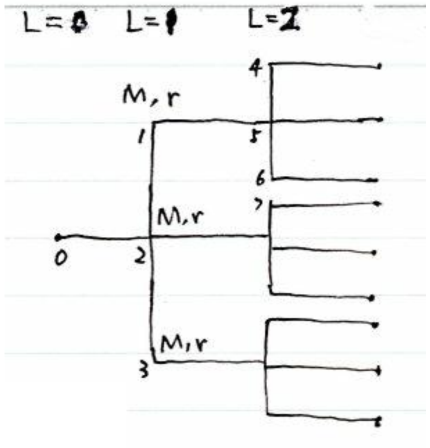
两大步：

1. 事先把样本子集划分好，计算并存储x_p的M_p，r_p及$D(x_i,M_p)$

2. 用分支定界算法搜索x的最近邻



搜索算法：（最近邻）



1° （初始化）

置 $B=\infty, L=0, p=0$（当前结点）。

2° （当前结点展开）

把当前结点的直接子结点放入（当前水平的）一个目录表（活动表）中，对它们计算并存储 $D(x,M_p)$。

（注意：活动表在每个水平上一个，下文均指当前水平的活动表）

3° （检验）

对活动表中每个结点，若 $D(x,M_p)>B+r_p$，则从表中去掉。

4° （回溯）

若活动表中已无结点，则回到上一级，置 $L = L - 1$

如 $L == 0$，则算法终止；

如 $L \neq 0$，则转 3°；

若活动表中有结点，则继续 5°。

5° （选择最近结点）

在目录表中选择最近结点（$D(x, M_p)$ 最小），记为 p'，以它为当前结点，若当前水平 L 为最终水平，则转 6°。

否则，置 $L = L + 1$，转 2°。

6° （检验）

对当前结点 p' 中的每个 x_i，

若 $D(x, M_p) > D(x_i, M_p) + B$，则非最近邻；（规则 2）

否则，计算 $D(x, x_i)$，

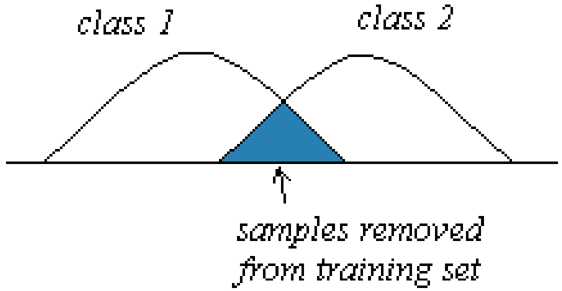
若 $D(x, x_i) < B$，则置 $NN = i$，$B = D(x, x_i)$

p' 中所有 x_i 被检验过之后，转 3°。

算法终止时，输出 x 的最近邻 x_{NN} 和 $D(x, x_{NN}) = B$

（K-近邻时只须修正上述算法的第 6°步）

剪辑近邻法



基本理解：

处在两类交界处或分布重合区的样本可能误导近邻法决策。

应将它们从样本集中去掉。

基本思路：

考查样本是否为可能的误导样本，

若是则从样本集中去掉——剪辑。

考查方法是通过试分类，认为错分样本为误导样本。

基本做法：

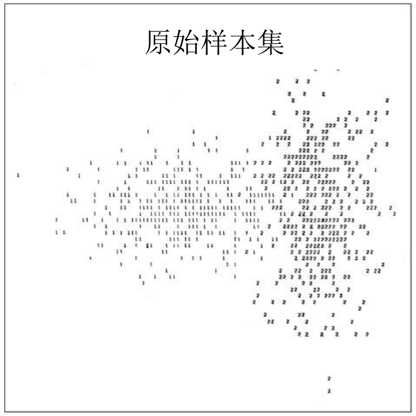
将样本集分为考试集 X^{NT} 和参考集 X^{NR}： $X^{N} = X^{NT} \cup X^{NR}$，$X^{NT} \cap X^{NR} = \phi$

剪辑：用 X^{NR} 中的样本对 X^{NT} 中的样本进行近邻法分类

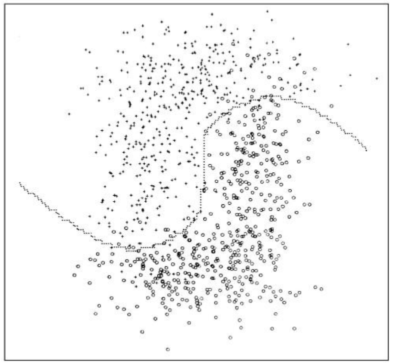
剪掉 X^{NT} 中被错分的样本，X^{NT} 中剩余样本构成剪辑样本集 X^{NTE}

分类：利用 X^{NTE} 和近邻法对未知样本 x 分类。

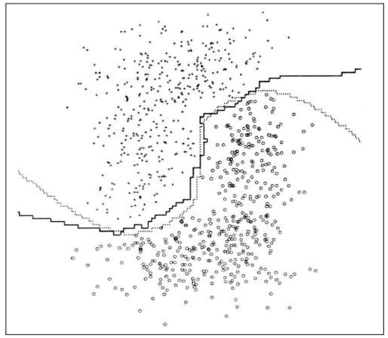
剪辑近邻法示例（教材图 6-6）：



原始数据贝叶斯分类面



剪辑近邻法分类面



压缩近邻法

主要用以减少存储量

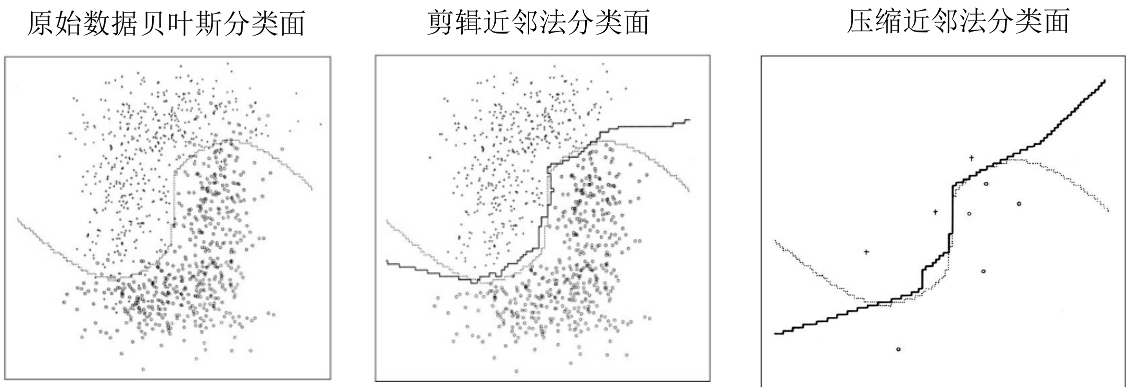
将 X^N 分为 X_s 和 X_G，开始时 X_s 中只有一个样本，X_G 中为其余样本。

考查 X_G 中每个样本，若用 X_s 可正确分类则保留，否则移入 X_s，……

最后用 X_s 作分类的样本集。

可与剪辑法配合使用。

例：



可做拒绝决策的近邻法

由于近邻法决策实际只取决于个别样本，因此有时风险较大，尤其是当两类近邻数接近时，为此，可考虑引入拒绝决策。

方法很简单：　　设某个 $k' > \frac{1}{2}(k+1)$，　$(k' < k)$

只有当 x 的 k 个近邻中有大于或等于 k' 个属于 ω_i 类时，

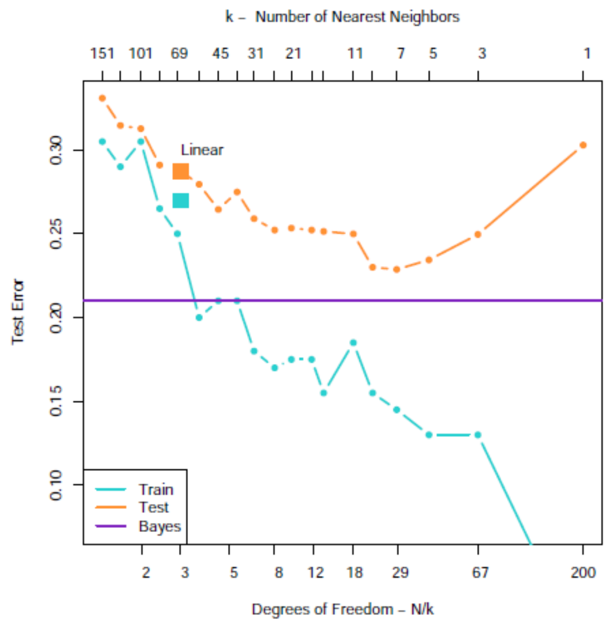
才决策 $x \in \omega_i$，否则拒绝

—— 简单多数 $\Rightarrow$ 绝对多数

拒绝决策同样可引入改进的近邻法中，比如剪辑近邻法

例：k 的选择的影响

Ref. Hastie, Tibshirani, Friedman, *The Elements of Statistical Learning*, Springer



Ref. Hastie, Tibshirani, Friedman, *The Elements of Statistical Learning*, Springer

高维空间的问题（维数灾难）

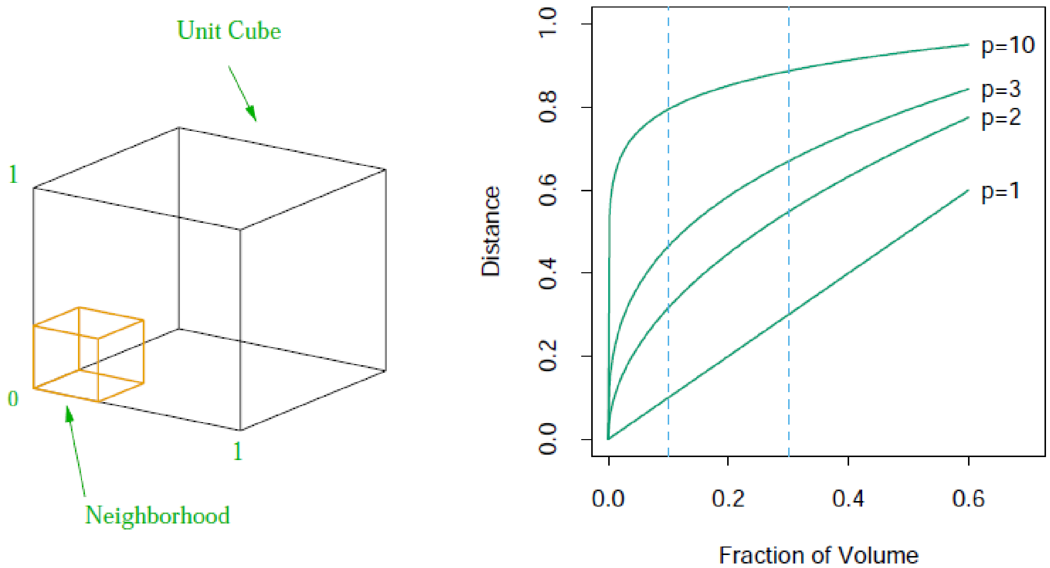


FIGURE 2.6. *The curse of dimensionality is well illustrated by a subcubical neighborhood for uniform data in a unit cube. The figure on the right shows the side-length of the subcube needed to capture a fraction r of the volume of the data, for different dimensions p. In ten dimensions we need to cover 80% of the range of each coordinate to capture 10% of the data.*

关于距离的计算

- 不同特征维间归一化的问题
- 不同的距离度量方式
  - 欧式距离
  - 曼哈顿距离
  - 闵可夫斯基距离 $dist_{mk}(\boldsymbol{x},\boldsymbol{y})=(\sum_{u=1}^{n}|x_u-y_u|^p)^{\frac{1}{p}}$
  - 编辑距离
  - ……