近邻法
Nearest Neighbour

杜逆索

近邻法

5.1 最近邻法

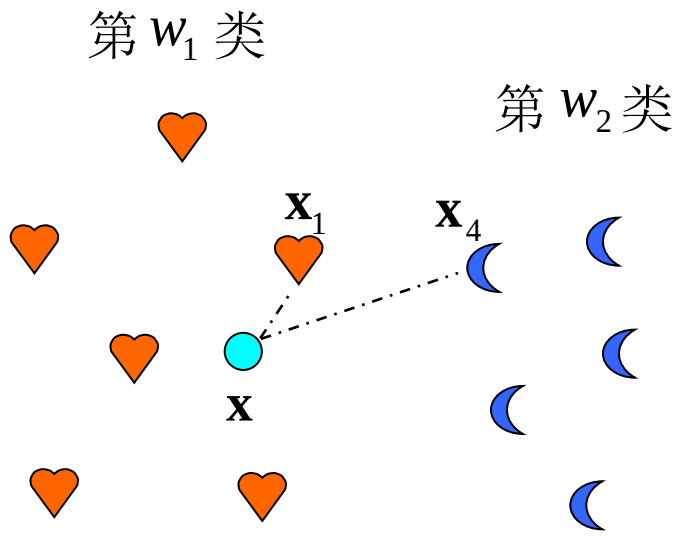
5.2 K近邻法

5.3 近邻法的错误率

5.4 剪辑近邻法

关于近邻法

- 最经典的模式识别方法之一
- 方法简单，便于理论分析
- 是其它模式识别方法的标尺
- “距离”的度量方式有很多种



近邻法原理示意图

近邻法

- **近邻法属于有监督学习，聚类属于无监督学习。**
- 它是在已知模式类别的训练样本的条件下，绕开概率的估计，按最近距离原则对待识别模式直接进行分类。

近邻法

- 最近邻分类器(nearest neighborhood classifier, nnc)：最小距离分类器的一种极端的情况，以全部训练样本作为代表点，计算测试样本与所有样本的距离，并以最近邻者的类别作为决策类。

- 最初的近邻法是由Cover和Hart于1968年提出的，随后得到理论上深入的分析与研究，是非参数法中最重要的方法之一。

近邻法

- 我们常说，物以类聚，人以群分，判别一个人是一个什么样品质特征的人，常常可以从他/她身边的朋友入手，所谓观其友，而识其人。

最近邻决策规则

对于有c个类别$(w_1, w_2, \ldots, w_c)$的模式识别问题，每类有 $N_i (i = 1, 2, \ldots, c)$个样本，则第$i$ 类 w_i 的判别函数为

$$\varphi_i(x) = \min_k \left\| \vec{x} - \vec{x}_i^{\,k} \right\|, k = 1, 2, \ldots, N_i$$

其中，$\vec{x}_i^{\,k}$ 的角标i 表示w_i类，k表示 w_i类 N_i个样本的第个样本；$\|\bullet\|$ 表示距离 ，这只是一个象征性的表示，可以采用任何一种相似性度量。

如果

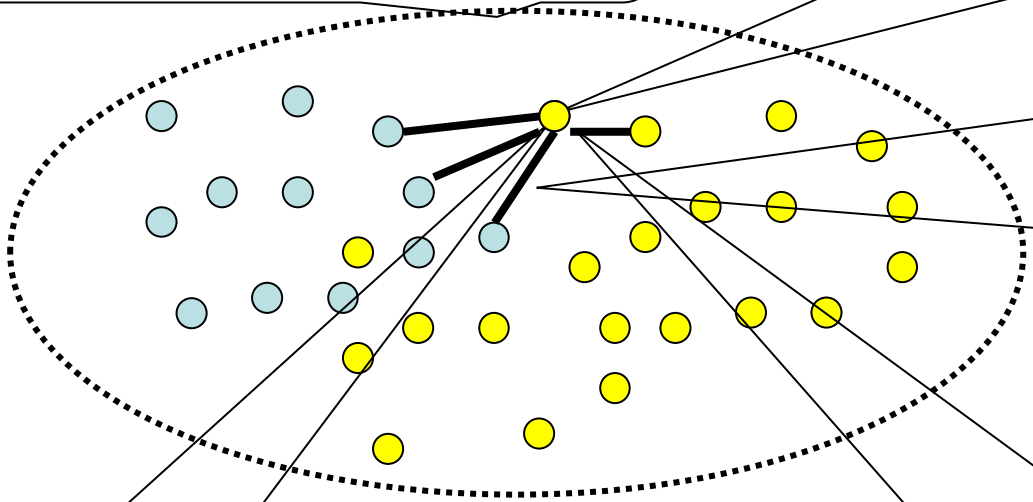
$$\varphi_i(\vec{x}) = \min_i \varphi_i(\vec{x}), i = 1, 2, \ldots c$$

那么决策 $\vec{x} \in w_j$，称这一决策方法为最近邻法

最近邻方法

(1) N个已知类别样本X

(2) 输入未知类别样本x



(3) 计算x到$x_i \in X$, (i=1, 2, ..., N)的距离$d_i(x)$

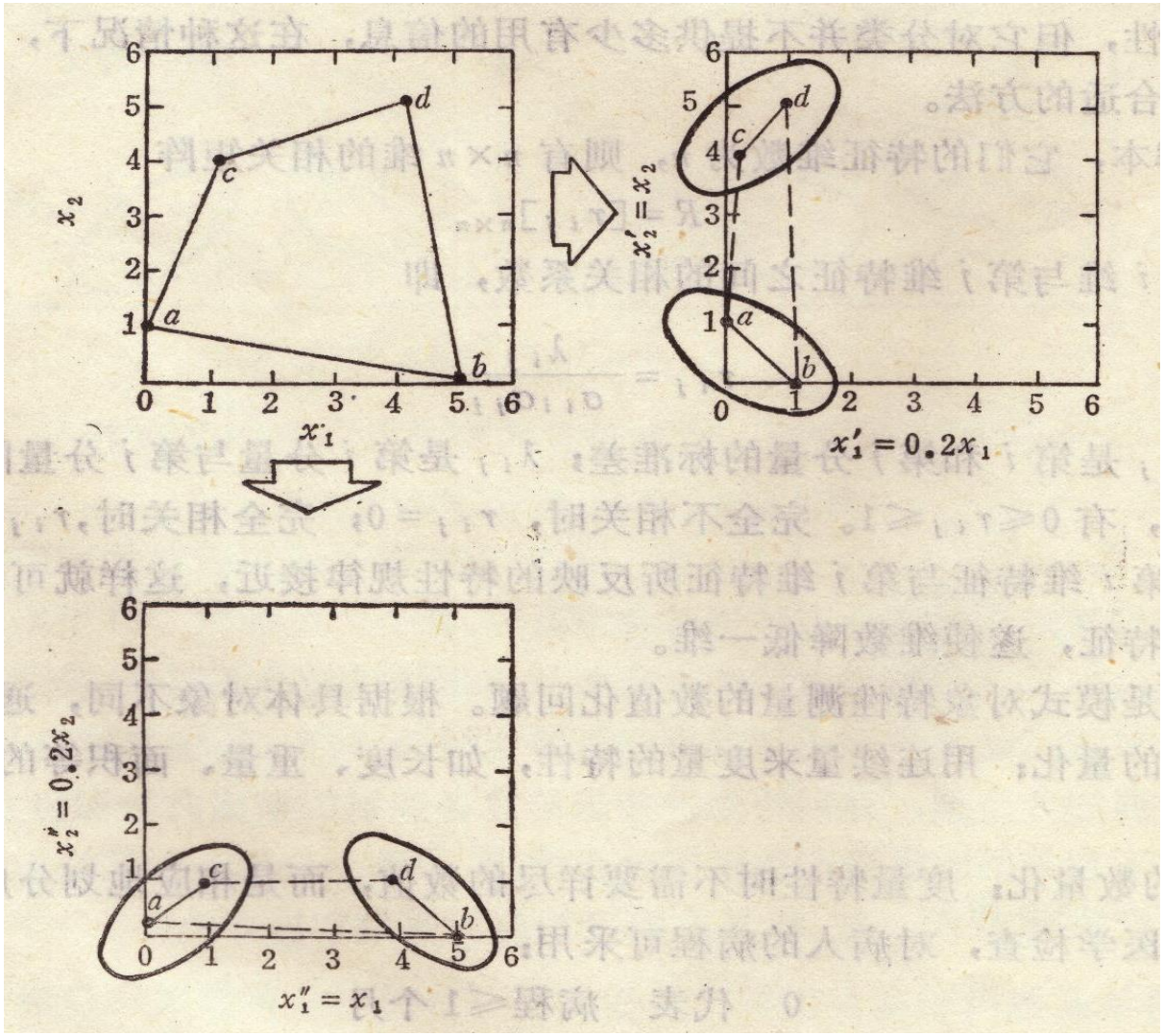
(4) 找出最小距离 $d_m(x)=\min\{d_i(x)\}$

(5) 判$x \in \omega_m$

距离的测度

- 欧氏距离
  - 量纲对分类的影响（下页图例）
- 马氏距离
  - 特点：排除了模式样本之间的相关性
  - 问题：协方差矩阵在实际应用中难以计算
- 一般化的明氏距离
- 角度相似性函数
  - 特点：反映了几何上相似形的特征，对于坐标系的旋转、放大和缩小等变化是不变的。
  - 当特征的取值仅为(0,1)两个值时的特例

量纲的影响（图例）



欧氏距离

- 设x和z为两个模式样本，其欧氏距离定义为：$D = \| x - z \|$
- 例：$x = (x1, x2)$，$z = (z1, z2)$，则
- 显然，模式x和z之间的距离越小，它们越相似。欧氏距离的概念和习惯上距离的概念是一致的。

马氏距离

- 设x是模式向量，m是均值向量，C为模式总体的协方差矩阵，则马氏距离的表达式：

$$D^2 = (x - m)^T C^{-1} (x - m)$$

一般化的明氏距离

- 模式样本向量x_i和x_j之间的明氏距离表示为：

$$D_m(x_i, x_j) = \left[\sum_k (x_{ik} - x_{jk})^m \right]^{1/m}$$

- 其中x_{ik}和x_{jk}分别表示x_i和x_j的第k各分量。
- 显然，当m=2时，明氏距离即为欧氏距离
- 特例：当m=1时，亦称为街坊距离。

角度相似性函数

- 表达式：$S(x,z)=\frac{x^{T}z}{\|x\|\cdot\|z\|}$
- 它表示模式向量x和z之间夹角的余弦，也称为x的单位向量与z的单位向量之间的点积。
- 特例：当特征的取值仅为(0, 1)两个值时，夹角余弦度量具有特别的含义，即当模式的第i个分量为1时，认为该模式具有第i个特征；当模式的第i个分量为0时，认为该模式无此特征。这时，$x^{T}z$的值就等于x和z这两个向量共同具有的特征数目。

近邻法

问题描述:

特征向量　　类别

- X=(0.1,0.1)　　?

特征向量	类别
(0.1,0.2)	**W1**
(0.2,0.1)	**W1**
(0.4,0.5)	**W2**
(0.5,0.4)	**W2**

近邻法

– 设有6个五维模式样本为3个类如下，按最小距离准则进行聚类：

x_1: 0, 3, 1, 2, 0

x_2: 1, 3, 0, 1, 0

x_3: 3, 3, 0, 0, 1

x_4: 1, 1, 0, 2, 0

x_5: 3, 2, 1, 2, 1

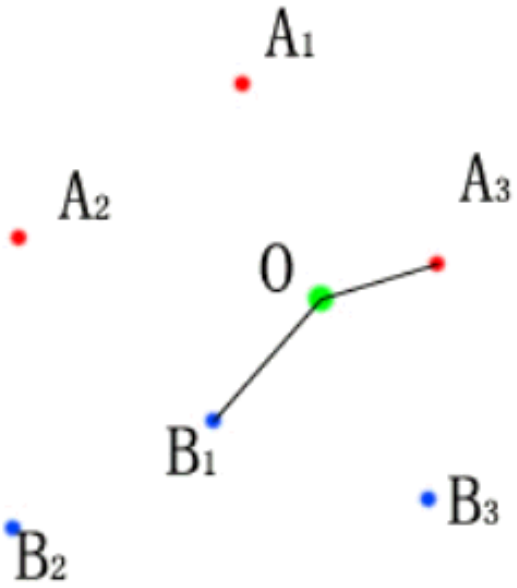
x_6: 4, 1, 1, 1, 0

问x_7: 1, 2, 1, 2, 1　按近邻法分类属于哪一类

最近邻法的错误率

- 最近邻法的错误率是比较难计算的，这是因为训练样本集的数量总是有限的，有时多一个少一个训练样本对测试样本分类的结果影响很大。

- 红点表示A类训练样本，蓝点表示B类训练样本，而绿点O表示待测样本。
- 假设以欧氏距离来衡量，O的最近邻是A_3，其次是B_1，因此O应该属于A类；
- 但若A_3被拿开，O就会被判为B类。

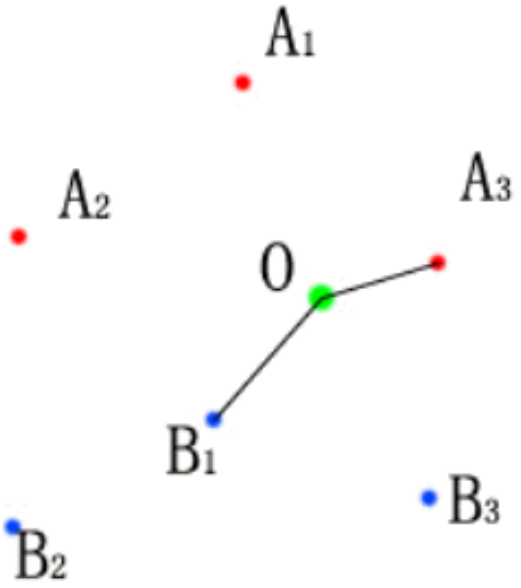


最近邻法的错误率

- 这说明计算最近邻法的错误率会有偶然性，也就是指与具体的训练样本集有关。

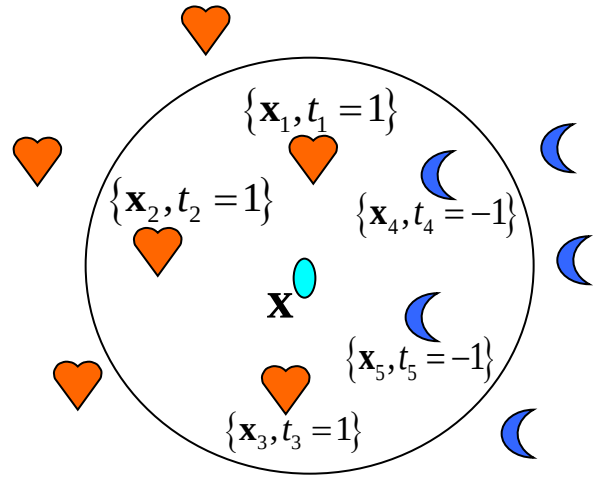
- 同时还可看到，计算错误率的偶然性会因训练样本数量的增大而减小。

- 因此我们就利用训练样本数量增至极大，来对其性能进行评价。这要使用渐近概念，书本以及后面都是在渐近概念下来分析错误率的。



k近邻法

- 是近邻法的一种推广；
- 原理：先找出 $\mathbf{x}$ 的k个近邻，这k个近邻中，哪一类的样本数量占优势，就将 $\mathbf{x}$ 归为哪一类。
- 选择合适的k值很重要；



k−最近邻方法原理示意图

k=5

K-近邻法

K近邻法是最近邻法的推广。K近邻法的具体描述如下，在N个已知样本中，找出x的K个近邻。设在这N个样本中，来自 w_1 类的样本有N_1个，来自 w_2 类的有N个，……，来自 w_i 类的有N_i个，若 $k_1, k_2, \ldots, k_c$ 分别是K个近邻中属于 $w_1, w_2, \ldots, w_c$ 类的样本数，则我们可以定义判断函数为：

$$\varphi_i(x) = k_i, i = 1, 2, \ldots, c$$

的决策规则，若

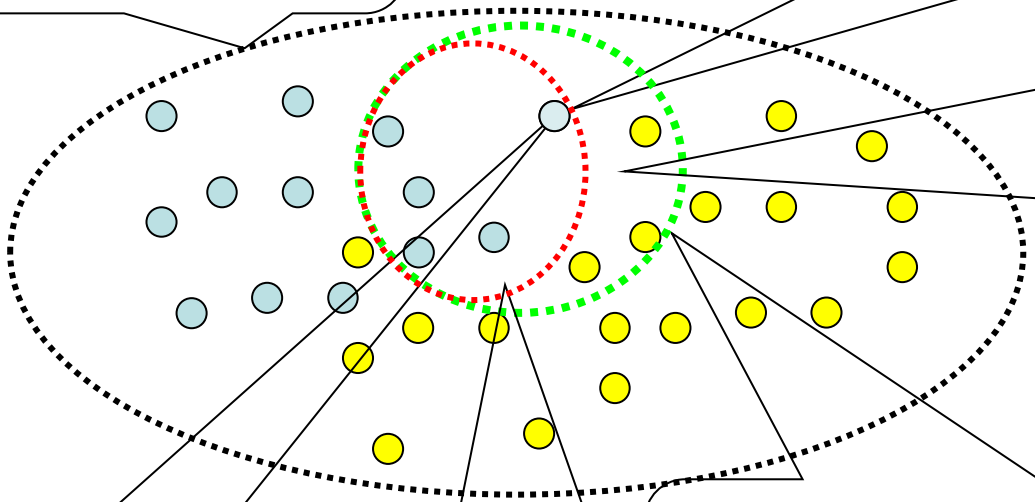
$$\varphi_i(x) = \max(\varphi_i(x)), i = 1, 2, \ldots, c$$

则决策 $x \in w_i$ ，这种方法通常称为K近邻法，即K-NN法。

K近邻法

(1) 已知N个已知类别样本X

(2) 输入未知类别样本x



(3) 计算x到$x_i \in X$, (i=1, 2, …, N) 的距离$d_i(x)$

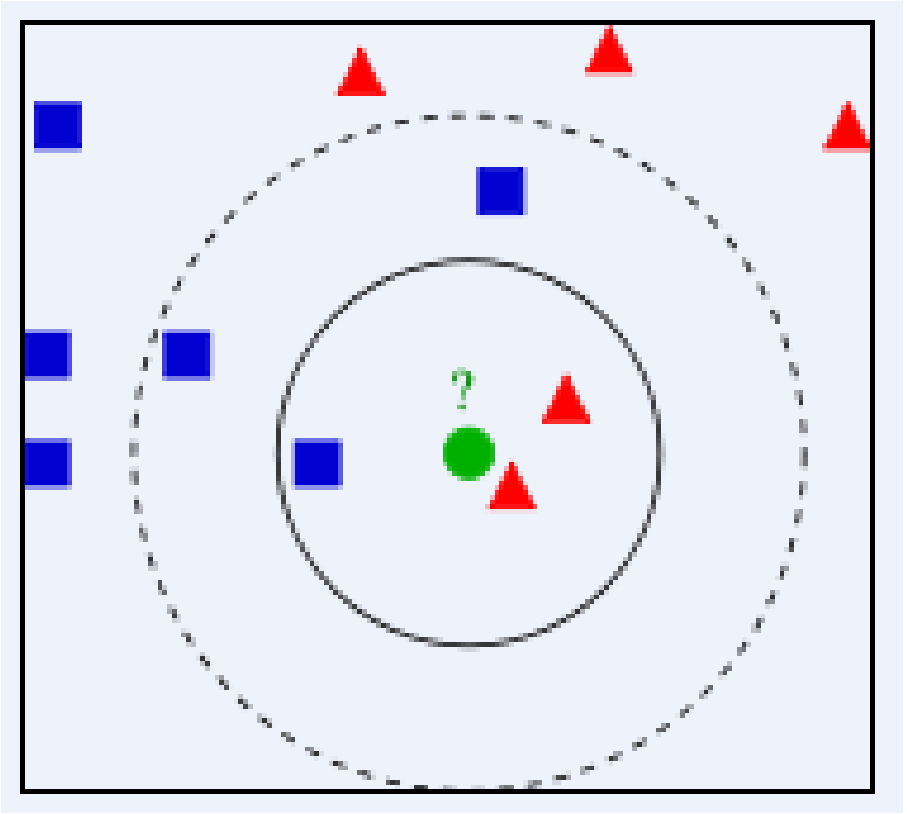
(4) 找出x的k个最近邻元$X_k=\{x_i, i=1, 2, \ldots, k\}$

(5) 看X_k中属于哪一类的样本最多$k_1=3<k_2=4$

(6) 判$x \in \omega_2$

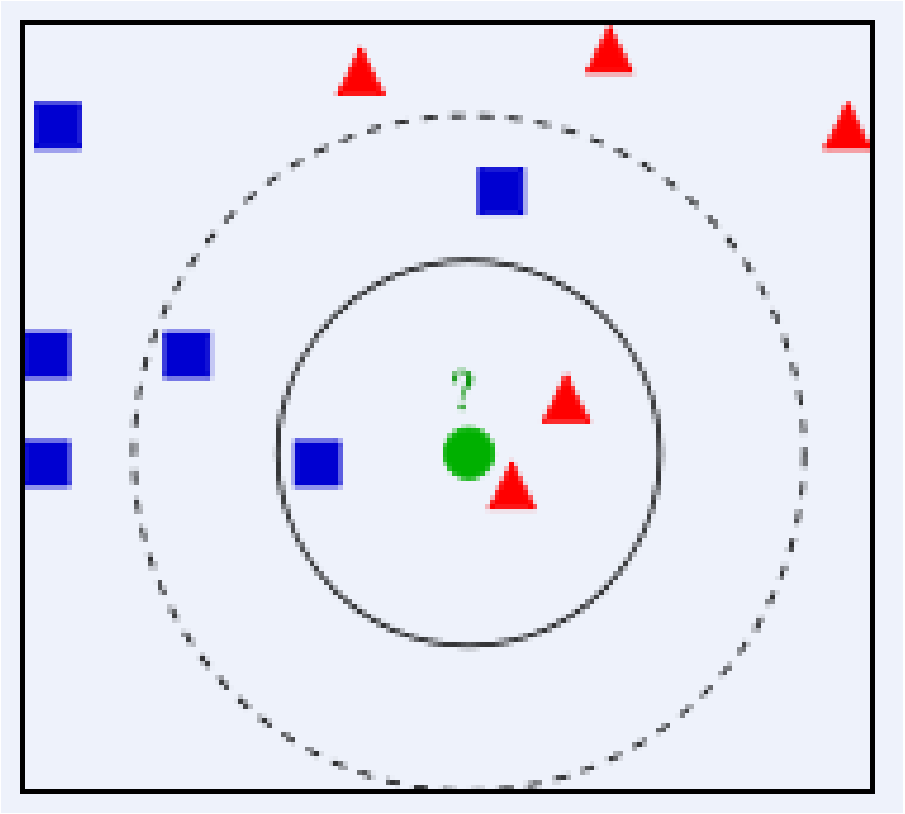
K的重要性

- 如图所示，有两类不同的样本数据，分别用蓝色的小正方形和红色的小三角形表示，而图正中间的那个绿色的圆所标示的数据则是待分类的数据。

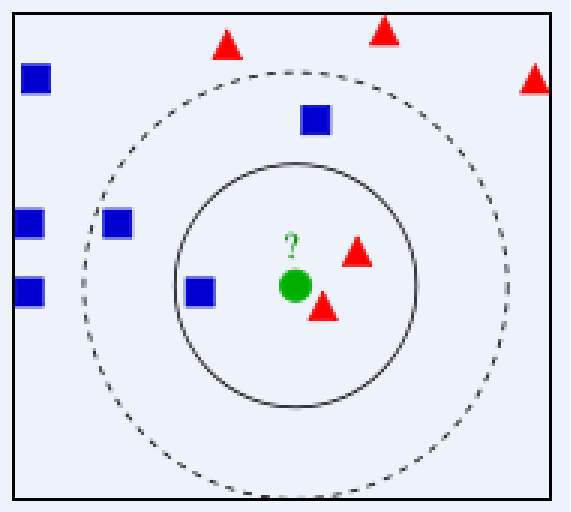


K的重要性

- 我们不是要判别上图中那个绿色的圆是属于哪一类数据么，好说，从它的邻居下手。但一次性看多少个邻居呢？

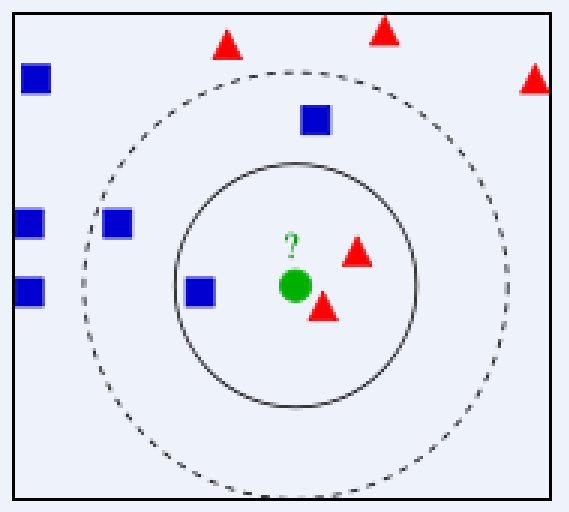


- 如果K=1，绿色圆点的最近的1个红色小三角形和1个蓝色小正方形，判定绿色的这个待分类点属于红色的三角形一类。
- 如果K=3，绿色圆点的最近的3个邻居是2个红色小三角形和1个蓝色小正方形，少数从属于多数，基于统计的方法，判定绿色的这个待分类点属于红色的三角形一类。



K的重要性

- 如果K=5，绿色圆点的最近的5个邻居是2个红色三角形和3个蓝色的正方形，还是少数从属于多数，基于统计的方法，判定绿色的这个待分类点属于蓝色的正方形一类。



KNN算法特征

- 于此我们看到，当无法判定当前待分类点是从属于已知分类中的哪一类时，我们可以依据统计学的理论看它所处的位置特征，衡量它周围邻居的权重，而把它归为(或分配)到权重更大的那一类。这就是K近邻算法的核心思想。
- KNN算法中，所选择的邻居都是已经正确分类的对象。该方法在定类决策上只依据最邻近的一个或者几个样本的类别来决定待分样本所属的类别。

KNN算法特征

- KNN 算法本身简单有效，它是一种 lazy-learning 算法，分类器不需要使用训练集进行训练，训练时间复杂度为0。KNN 分类的计算复杂度和训练集中的文档数目成正比，也就是说，如果训练集中文档总数为 n，那么 KNN 的分类时间复杂度为O(n)。

KNN算法特征

- KNN方法虽然从原理上也依赖于极限定理，但在类别决策时，只与极少量的相邻样本有关。由于KNN方法主要靠周围有限的邻近的样本，而不是靠判别类域的方法来确定所属类别的，因此对于类域的交叉或重叠较多的待分样本集来说，KNN方法较其他方法更为适合。

KNN模型

- K 近邻算法使用的模型实际上对应于对特征空间的划分。
- K 值的选择
- 距离度量
- 分类决策规则
- 是该算法的三个基本要素。

K 值的选择

- 1. K值的选择会对结果产生重大影响。K值较小意味着只有与输入实例较近的训练实例才会对预测结果起作用，但容易发生过拟合；如果K值较大，优点是可以减少学习的估计误差，但缺点是学习的近似误差增大，这时与输入实例较远的训练实例也会对预测起作用，是预测发生错误。在实际应用中，K值一般选择一个较小的数值，通常采用交叉验证的方法来选择最优的K值。K一般取奇数。

距离度量

- 2.距离度量一般采用欧氏距离，在度量之前，应该将每个属性的值规范化，这样有助于防止具有较大初始值域的属性比具有较小初始值域的属性的权重过大。
-

距离度量

- 对事物进行分类本身是依据同类样本属性的相似性，在使用特征向量表示时，体现为同类样本在特征空间中靠的很近，因此可以用各种方法度量样本数据间的差异。一般说来，使用欧氏距离是最常用的，它表示两个向量的差向量的模，这种计算在衡量几何距离时最为合适，例如各城市之间的距离。

距离度量

- 但在模式识别中特征向量的各个分量的含义往往是不同的，就像苹果的例子中，一个表示重量，一个表示直径，两者的单位都不一样，因此使用欧氏距离并不合理。一般来说样本的各个分量的分布范围在数量级上比较相近为好。使用分量差的绝对值总和表示距离往往是对欧氏距离的简化，将平方计算改为了绝对值计算。

分类决策规则

- 3.该算法中的分类决策规则往往是多数表决，即由输入实例的 K 个最临近的训练实例中的多数类决定输入实例的类别

优点

- 方法简单，不需要估计概率
- 对所有未知输入都可以进行分类
- 物理意义明确
- 是一种经典的模式识别方法，是衡量其它方法的标尺。

不足

- 当样本不平衡时，如一个类的样本容量很大，而其他类样本容量很小时，有可能导致当输入一个新样本时，该样本的K个邻居中大容量类的样本占多数。 该算法只计算“最近的”邻居样本，某一类的样本数量很大，那么或者这类样本并不接近目标样本，或者这类样本很靠近目标样本。

不足

- 该方法的另一个不足之处是计算量较大，因为对每一个待分类的文本都要计算它到全体已知样本的距离，才能求得它的K个最近邻点。目前常用的解决方法是事先对已知样本点进行剪辑，事先去除对分类作用不大的样本。该算法比较适用于样本容量比较大的类域的自动分类，而那些样本容量较小的类域采用这种算法比较容易产生误分。

不足

- 实现 K 近邻算法时，主要考虑的问题是如何对训练数据进行快速 K 近邻搜索，这在特征空间维数大及训练数据容量大时非常必要。

近邻法常见的改进与完善措施

- 尽管近邻法有其优良品质，但是它的一个严重弱点与问题是需要存储全部训练样本，以及繁重的距离计算量。
- 但以简单的方式降低样本数量，只能使其性能降低，这也是不希望的。
- 为此要研究既能减少近邻法计算量与存储量，同时又不明显降低其性能的一些改进算法。

近邻法常见的改进与完善措施

- 改进的方法大致分为两种原理。一种是对样本集进行组织与整理，分群分层，尽可能将计算压缩到在接近测试样本邻域的小范围内，避免盲目地与训练样本集中每个样本进行距离计算。
- 另一种原理则是在原有样本集中挑选出对分类计算有效的样本，使样本总数合理地减少，以同时达到既减少计算量，又减少存储量的双重效果。

近邻法常见的改进与完善措施

（1）快速搜索近邻法

（2）剪辑近邻法

（3）最佳距离度量近邻法

（4）近邻法的其它改进措施

••••••••••••

相关的改进与完善已使近邻法成为一个庞大的家族。

快速搜索近邻法

- 这种方法着眼于只解决减少计算量，但没有达到减少存储量的要求。

- 基本思想：
  - 将样本集按邻近关系分解成组，给出每组的质心所在，以及组内样本至该质心的最大距离。这些组又可形成层次结构，即组又分子组。
  - 因而待识别样本可将搜索近邻的范围从某一大组，逐渐深入到其中的子组，直至树的叶结点所代表的组，确定其相邻关系。

最近邻法平均错误率P和最小错误率贝叶斯分类器的平均错误率P*之间有关系式（当 $N \to +\infty$ 时）：

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

粗略地说，当 时，最近邻法错误率比最小错误率贝叶斯分类器的错误率P*略大，但不会大于2P*。由于贝叶斯分类器是所有分类器中错误率最小的，用最近邻法能得到接近贝叶斯分类器的性能应该说是不错的。

最近邻法的错误率

- 由于X'与所用训练样本集有关，因此错误率有较大偶然性。
- 但是如果所用训练样本集的样本数量N极大，即N→∞时，可以想像X'将趋向于X，或者说处于以X为中心的极小邻域内，此时分析错误率问题就简化为在X样本条件下X与一个X(X'的极限条件)分属不同类别的问题。
- 如果样本X的两类别后验概率分别为$P(\omega_1|X)$与$P(\omega_2|X)$，那么对X值，在N→∞条件下，发生错误决策的概率为：

$$\lim_{n\to\infty} P_N(e|X) = 1 - \sum_{i=1}^{c} P^2(\omega_i|X)$$

最近邻法的错误率

而在这条件下的平均错误率

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

$$= \int \lim_{N\to\infty} P_N(e \mid X) p(X) dX = \int [1 - \sum_{i=1}^{c} P^2(\omega_i \mid X)] p(X) dX$$

P称为渐近平均错误率，是$P_N(e)$在$N\to\infty$的极限。

为了与基于最小错误率的贝叶斯决策方法对比，下面写出贝叶斯错误率的计算式：　　$P^* = \int P^*(e \mid X) p(X) dx$

其中　$P^*(e \mid X) = 1 - P(\omega_m \mid X)$　　$P^*(\omega_m \mid X) = \max_i [P(w_i \mid X)]$

最近邻法的错误率

若是两类问题，则

贝叶斯错误率：　$P^*(e|X)=1-P(\omega_1|X)$

最近邻法错误率：　$\lim_{N\to\infty} P_N(e|X)=1-P^2(\omega_1|X)-P^2(\omega_2|X)$

$$\Delta P=P(\omega_1|X)[1-P(\omega_1|X)]-P^2(\omega_2|X)$$
$$=P(\omega_2|X)[P(\omega_1|X)-P(\omega_2|X)]$$

- 可见在一般情况下△P是大于零的值，只要$P(\omega_1|X)>P(\omega_2|X)>0$。

最近邻法的错误率

$$\Delta P = P(\omega_1 \mid X)[1 - P(\omega_1 \mid X)] - P^2(\omega_2 \mid X)$$
$$= P(\omega_2 \mid X)[P(\omega_1 \mid X) - P(\omega_2 \mid X)]$$

- 有以下两种例外情况△P＝0：
  - P(ω1|X)＝1
  - P(ω1|X)＝P(ω2|X)＝1/2。

最近邻法的错误率

请想一下，什么情况下$P(\omega_1|X)=1$或$P(\omega_2|X)=1$？ $P(\omega_1|X)=P(\omega_2|X)$会出现什么什么情况？

最近邻法的错误率

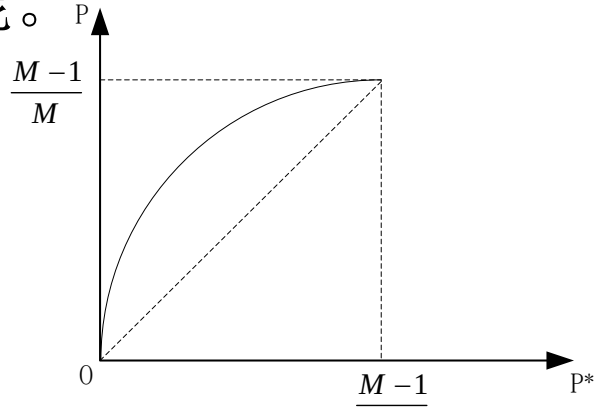
请想一下，什么情况下$P(\omega_1|X)=1$或$P(\omega_2|X)=1$？ $P(\omega_1|X)=P(\omega_2|X)$会出现什么什么情况？

- 一般来说，在某一类样本分布密集区，某一类的后验概率接近或等于1。此时，基于最小错误率贝叶斯决策基本没错，而近邻法出错可能也很小。
- 而后验概率近似相等一般出现在两类分布的交界处，此时分类没有依据，因此基于最小错误率的贝叶斯决策也无能为力了，近邻法也就与贝叶斯决策平起平坐了。

- 从以上讨论可以看出，当$N\to\infty$时，最近邻法的渐近平均错误率的下界是贝叶斯错误率，这发生在样本对某类别后验概率处处为1的情况或各类后验概率相等的情况。

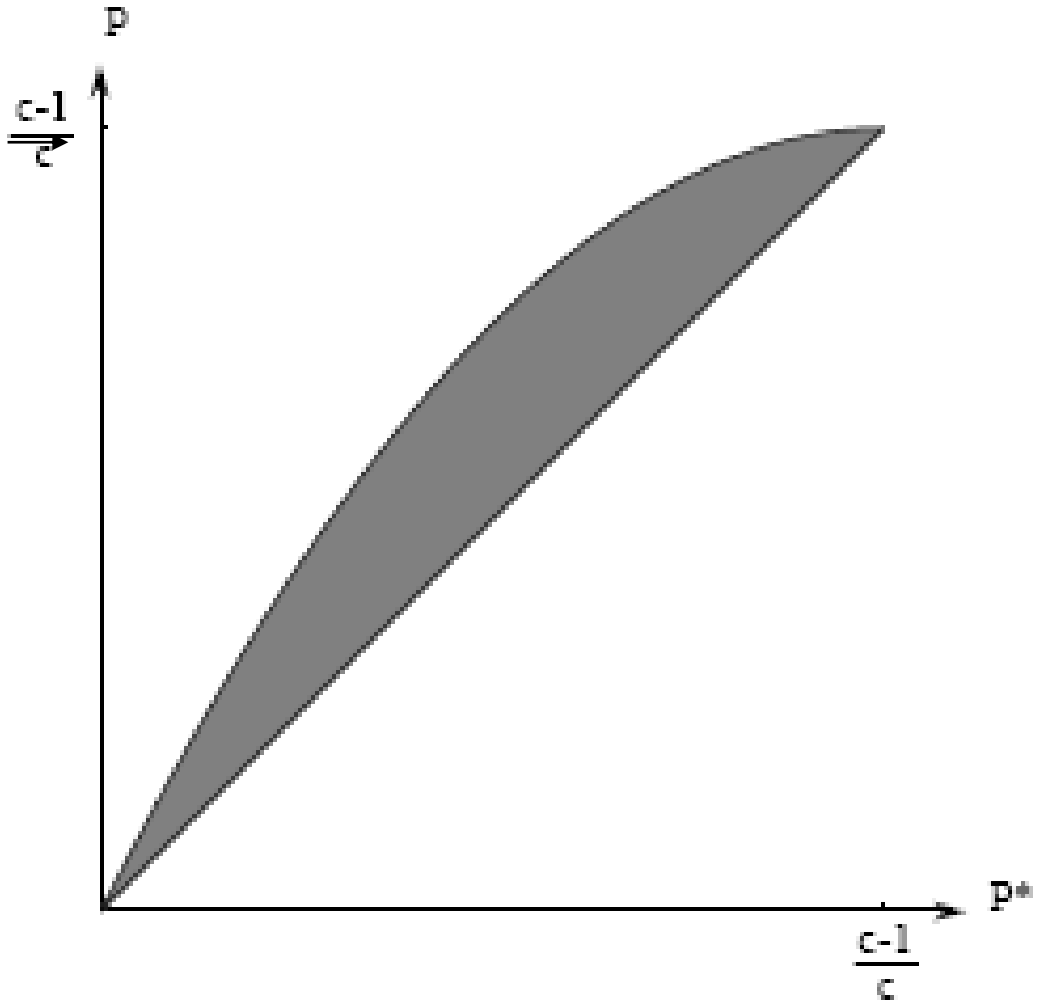
最近邻法的错误率分析

下图示出最近邻法错误率P与贝叶斯分类器错误率 间的关系。在 和 这两个极端情况下，在一般情况下P略大于，P有一个变化范围，它与M个类的相互构成情况有关。上面已经说过，当M个后验概率中有一个最大，其余M-1个都相等时是最近邻法最容易出错的情况。即使在这种情况下，最近邻法性能也是不错的。所有的这些结论都是在N很大的条件下得到的，若N不够大，造成最近邻 与X离得较远，这时不会有好的分类性能。

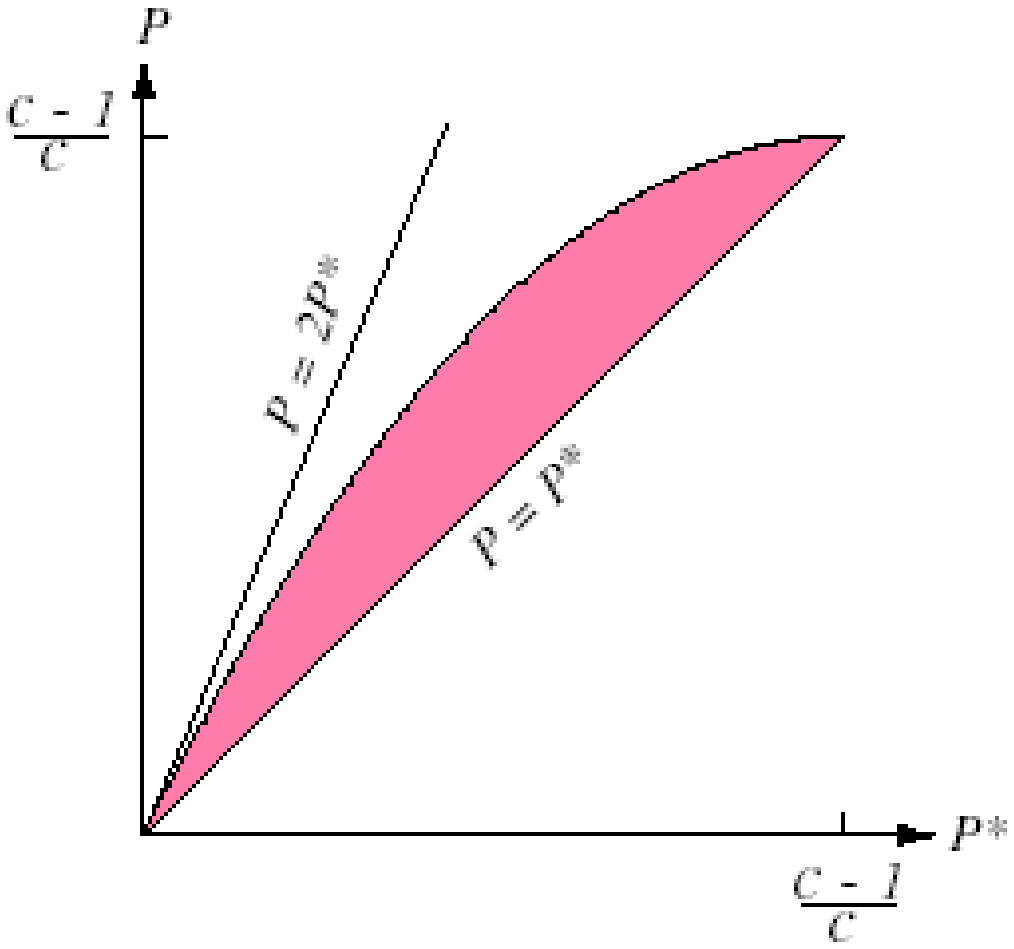


$N \to +\infty$ 时近邻法性能

最近邻法错误率上下界与贝叶斯错误率的关系

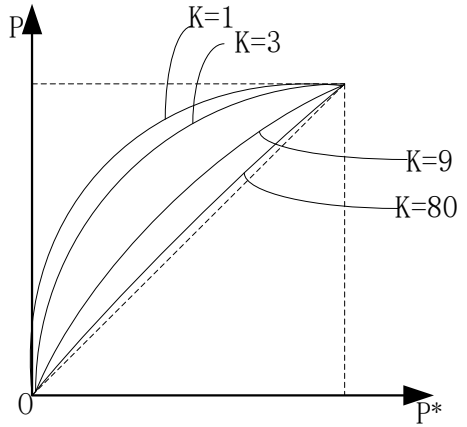


近邻法的错误率



K近邻法

以上所述的即是K近邻法的基本规则，这样做的所得分类器的性能应该比只根据单个近邻点就做决策的最近邻法来得好。进一步分析的结果证明了这一点，前提是$N \to +\infty$，下图示出 $N \to +\infty$ 时K近邻法错误率与贝叶斯分类器错误率之间的关系。从不同 K(包括K=1)的 $P - P^*$ 曲线可以看出当 $N \to +\infty$ 时的情况，K愈大，K近邻法的性能愈接近贝叶斯分类器。由此似乎应该得出结论：K愈大愈好。但图5-4所示的结果是在 $N \to +\infty$ 时的情况。随着K变大，要想使K个最近邻点都落在X点附近所需的样本数N比最近邻法中所需的样本数多的多。因此，我们可以发现k-近邻法在实际场合中无法实现，并且当错误代价很大时，也会产生较大的风险。

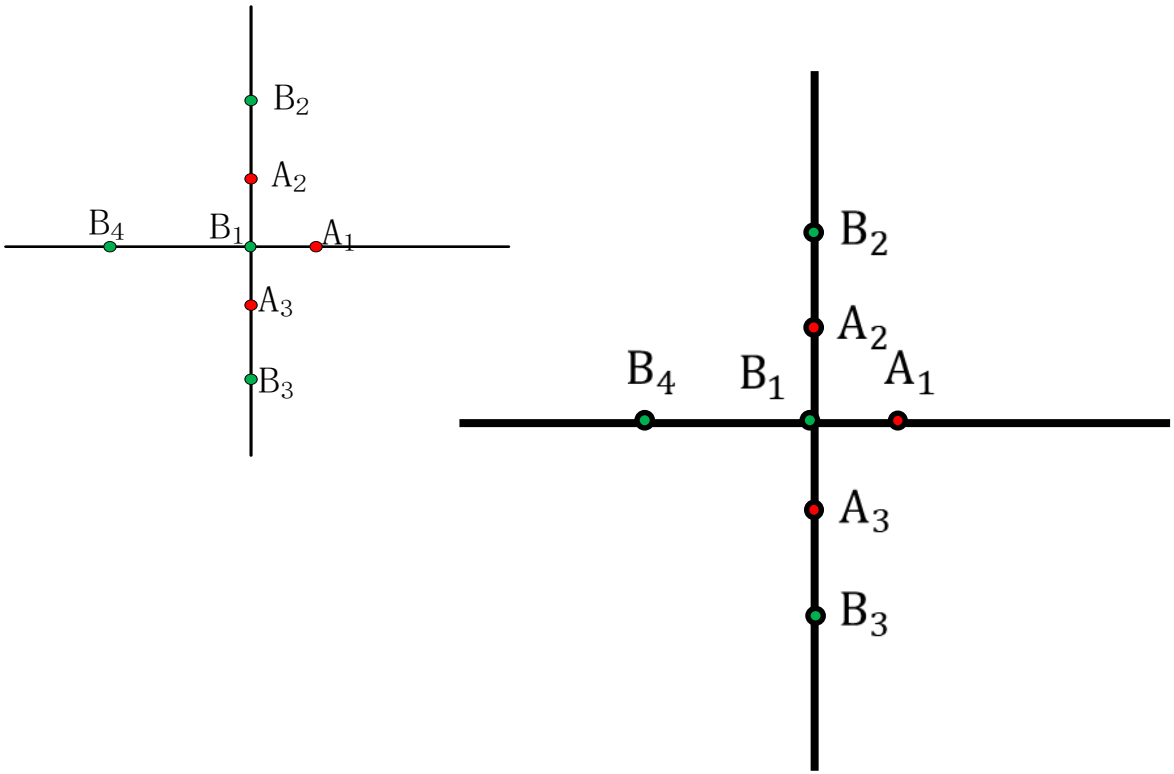


$N \to +\infty$ 时近邻法与贝叶斯分类器性能的对比

K近邻法

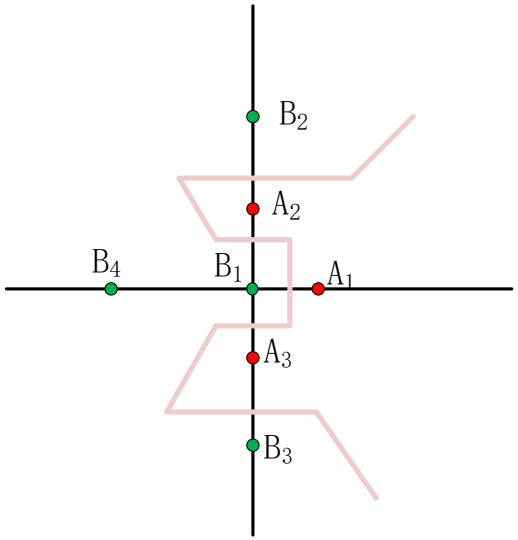
【例5-1】设在一个二维空间，A类有三个训练样本，图中用红点表示，B类四个样本，图中用蓝点表示。

试问：(1)按近邻法分类，这两类最多有多少个分界面；(2)画出实际用到的分界面



K近邻法

解: 按近邻法，对任意两个由不同类别的训练样本构成的样本对，如果它们有可能成为测试样本的近邻，则它们构成一组最小距离分类器，它们之间的中垂面就是分界面，因此由三个A类与四个B类训练样本可能构成的分界面最大数量为3×4＝12。实际分界面如下图所示，由9条线段构成。



剪辑近邻法

- 目的：去掉靠近两类中心的样本？
- 基本思想：当不同类别的样本在分布上有交迭部分的，分类的错误率主要来自处于交迭区中的样本。当我们得到一个作为识别用的参考样本集时，由于不同类别交迭区域中不同类别的样本彼此穿插，导致用近邻法分类出错。因此如果能将不同类别交界处的样本以适当方式筛选，可以实现既减少样本数又提高正确识别率的双重目的。为此可以利用现有样本集对其自身进行剪辑。

剪辑近邻方法

对于两类问题，设将已知类别的样本集 $X^{(N)}$ 分成参照集 $X^{(NR)}$ 和测试集 $X^{(NT)}$ 两部分，这两部分没有公共元素，们的样本数各为**NR**和**NT**，**NR+NT=N**。利用参照集 $X^{(NR)}$ 中的样本 $y_1, y_2, \cdots, y_{NR}$ 采用最近邻规则对已知类别的测试集 $X^{(NT)}$ 中的每个样本 $x_1, x_2, \cdots, x_{NT}$ 进行分类，剪辑掉 $X^{(NT)}$ 中被错误分类的样本。

若 $y^0(x) \in X^{(NR)}$ 是 $x \in X^{(NT)}$ 的最近邻元，剪辑掉不与 $y^0(x)$ 同类的 x ，余下的判决正确的样本组成剪辑样本集 $X^{(NTE)}$ ，这一操作称为剪辑。

剪辑近邻方法

获得剪辑样本集 $X^{(NTE)}$ 后，对待识模式 x 采用最近邻规则进行分类。

$$d_i(x) = \underbrace{\min}_{j=1,2,\cdots,N_i} \left\| x - x_j^{(i)} \right\| \qquad i = 1,2,\cdots,c$$

如果 $d_m(x) = \underbrace{\min}_{i=1,2,\cdots,c} d_i(x)$ 则 $x \in \omega_m$

这里 $x_j \in X^{(NTE)}$

剪辑近邻方法

可以证明下面的定理：当样本数 $N \to +\infty$,$\frac{NT}{NR} \to \frac{\alpha}{1-\alpha}$ ， 0<α <1,如果x是$p(x|w_1)$和 $p(x|w_2)$ 的连续点，设x在 $X^{(NT)}$ 中的最近邻为 x^0 ，则 在 $X^{(NR)}$ 中的最近邻 $y^0(x^0)$ 有

$$\lim_{N\to\infty} y^0(x^0) = x$$

且

$$\lim_{N\to\infty} P(w_i \mid y^0(x^0) = x) = P(w_i \mid x) \ \ (i = 1,2)$$

以该定理为基础，可以证明x的最近邻 x^0 属于 w_1 类的渐近概率为

$$\varphi(w_1 \mid x^0) \overset{def}{=} \lim_{N\to\infty} P(w_1 \mid x^0) = \frac{P(w_1 \mid x)^2}{1 - 2P(w_1 \mid x)P(w_2 \mid x)}$$

在给定x条件下的渐进误判概率为

$$\begin{aligned} P_{1-NN}^{E}(e \mid x) &= P(w_1 \mid x)\varphi(w_2 \mid x^0) + P(w_2 \mid x)\varphi(w_1 \mid x^0) \\ &= P(w_1 \mid x)\varphi(w_2 \mid x) + P(w_2 \mid x)\varphi(w_1 \mid x) \\ &= P(w_1 \mid x) + \varphi(w_2 \mid x) - 2\varphi(w_1 \mid x)P(w_1 \mid x) \\ &= P(w_1 \mid x) + \frac{P(w_1 \mid x)^2}{1 - 2P(w_1 \mid x)P(w_2 \mid x)} - \frac{2P(w_1 \mid x)^3}{1 - 2P(w_1 \mid x)P(w_2 \mid x)} \\ &= \frac{P(w_1 \mid x)P(w_2 \mid x)}{1 - 2P(w_1 \mid x)P(w_2 \mid x)} \end{aligned}$$

剪辑近邻方法

误判的情况是，x属于w_1 类而其近邻元属于w_2 类但其近邻元属于w_1 ；或x属于w_2 但其近邻元属于w_1 类，因此没有剪辑的最近邻法的渐近条件误判概率还可以表示成

$$P_{1-NN}(e|x) = 2P(w_1|x)P(w_2|x)$$

将上式代入可得

$$P^E_{1-NN}(e|x) = \frac{P_{1-NN}(e|x)}{2[1-P_{1-NN}(e|x)]}$$

由于上式分母中$P_{1-NN}(e|x) \le \frac{1}{2}$，从而分母不小于1，上式表明，剪辑近邻法的渐进条件误判概率小于或等于没有剪辑的最近邻法，即

$$P^E_{1-NN}(e|x) \le P_{1-NN}(e|x)$$

从而有

$$P^E_{1-NN}(e) \le P_{1-NN}(e)$$

当$P_{1-NN}(e|x)$ 很小时，可推知

$$P^E_{1-NN}(e) \approx P_{1-NN}(e)/2$$

剪辑近邻方法

由于没有剪辑的最近邻法渐近误差判断概率 $P_{1-NN}(e)$ 的上界为 $2P_B(e)$ ，因此经过剪辑的最近邻法的渐近误判概率 $P_{1-NN}^{E}(e)$ 接近贝叶斯误判概率 $P_B(e)$ ，即

$$P_{1-NN}^{E}(e) \approx P_B(e)$$

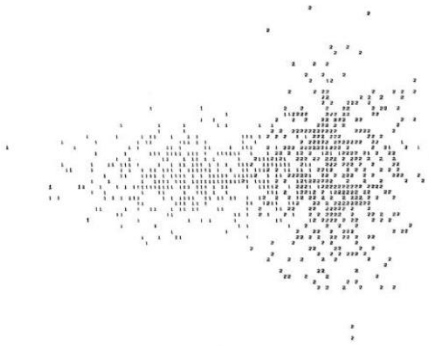
剪辑最近邻法可以推广至*k*—近邻法中，具体的做法是：第一步用*k—NN*法进行剪辑，第二步用*1—NN*法进行分类。

如果样本足够多，就可以重复地执行剪辑程序，以进一步提高分类性能。称为重复剪辑最近邻法。

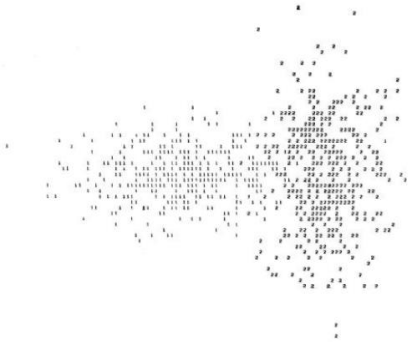
剪辑近邻方法

【例5-2】观察下列两个样本集使用剪辑近邻法后的分类状况。

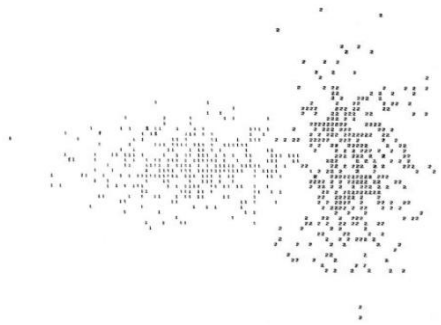
解：第一个样本集



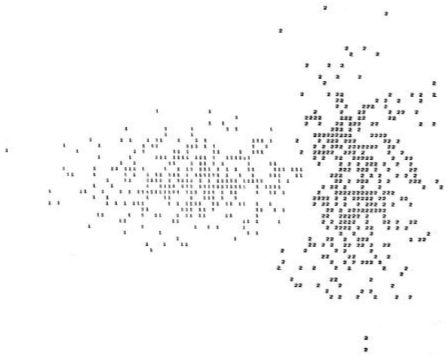
(a) 原始样本集



(b)第一次迭代以后



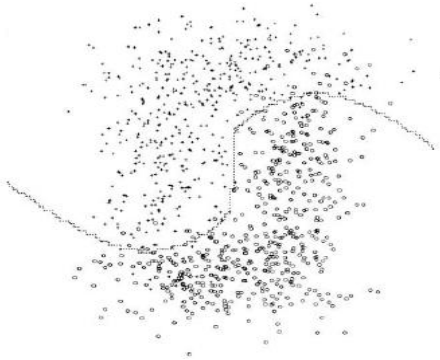
(c)第三次迭代以后



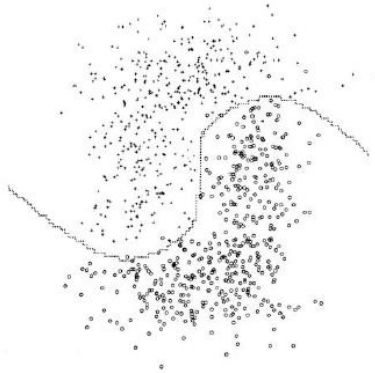
(d)最终的样本集

剪辑近邻方法

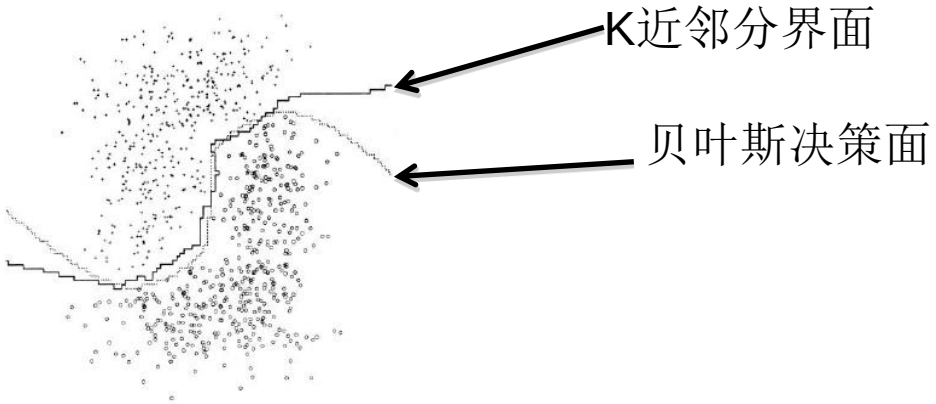
第二个样本集



(a) 原始样本集



(b)第一次迭代以后



(c)最终的样本集