第三章 判别域代数界面方程法

3.1 判别域界面方程分类的概念

3.2 线性判别函数

3.3 判别函数值的鉴别意义、权空间及解空间

3.4 fisher线性判别

3.5 线性可分条件下判别函数权矢量算法

3.6 一般情况下的判别函数权矢量算法

3.7 非线性判别函数

3.8 最近邻方法

3.6 最近邻方法

聚类分析： 无先验知识，按最近距离原则进行分类。

代数界面方法： 有先验知识，要进行训练，按判别函数值符号或大小进行分类。

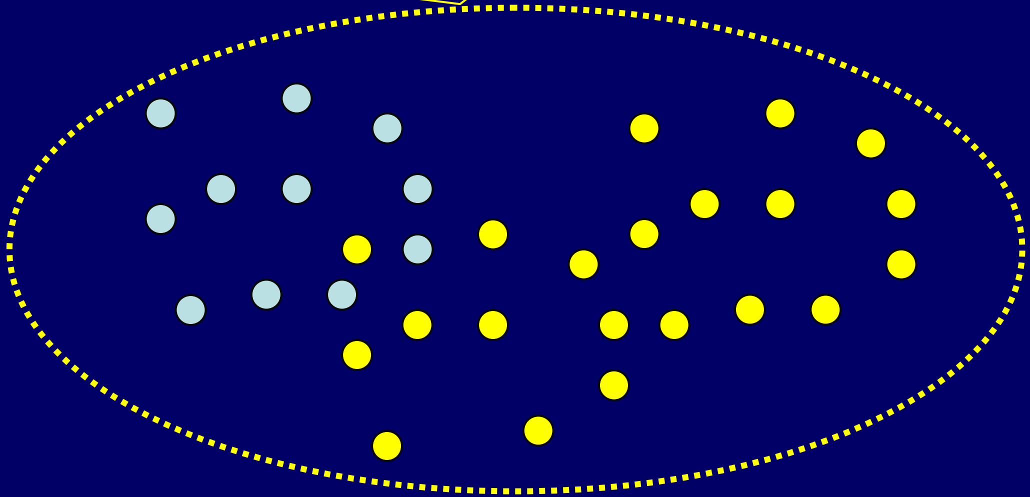
最近邻方法： 有先验知识，但不进行训练，按最近距离原则进行分类。

最近邻方法特点：

思想直观、方法简单、效果较好。

3.6 最近邻方法

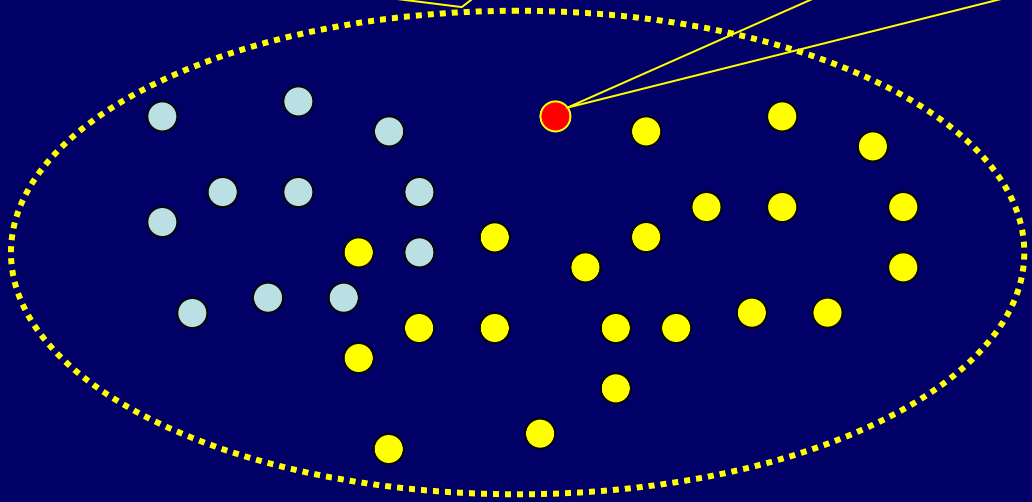
(1) 已知N个已知类别样本X



3.6 最近邻方法　1-NN

(1) 已知N个已知类别样本X

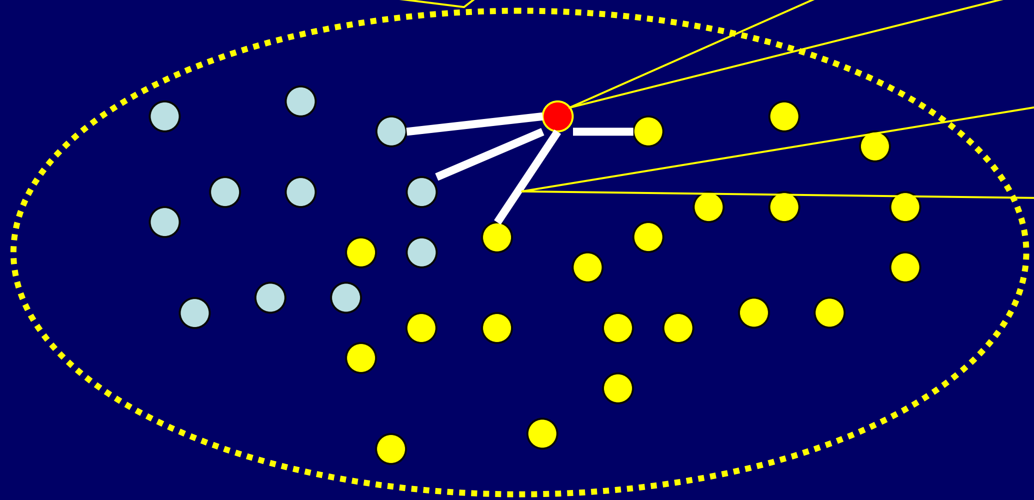
(2) 输入未知类别样本x



3.6 最近邻方法　　1-NN

(1) 已知N个已知类别样本X

(2) 输入未知类别样本x



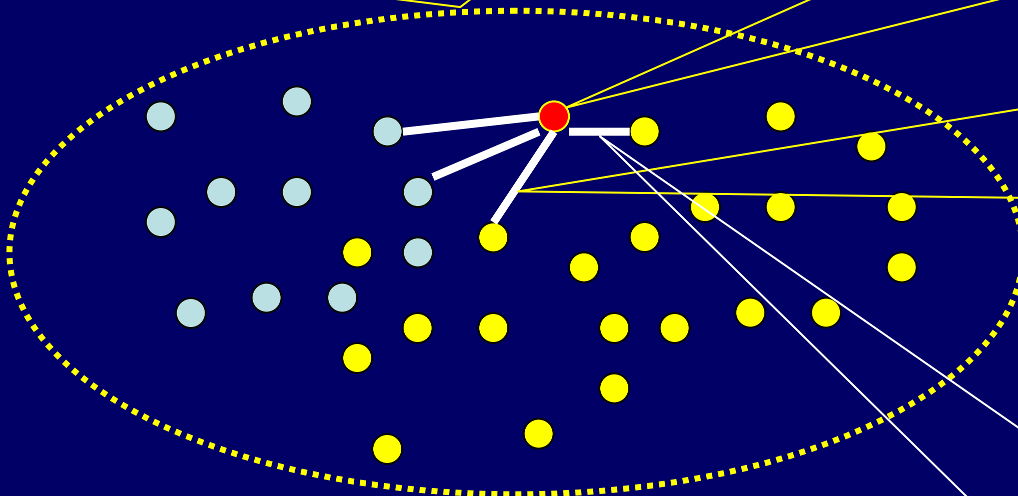
(3) 计算x到 $x_i \in X$, $(i=1, 2, \ldots, N)$ 的距离 $d_i(x)$

3.6 最近邻方法

1-NN

(1) 已知N个已知类别样本X

(2) 输入未知类别样本x



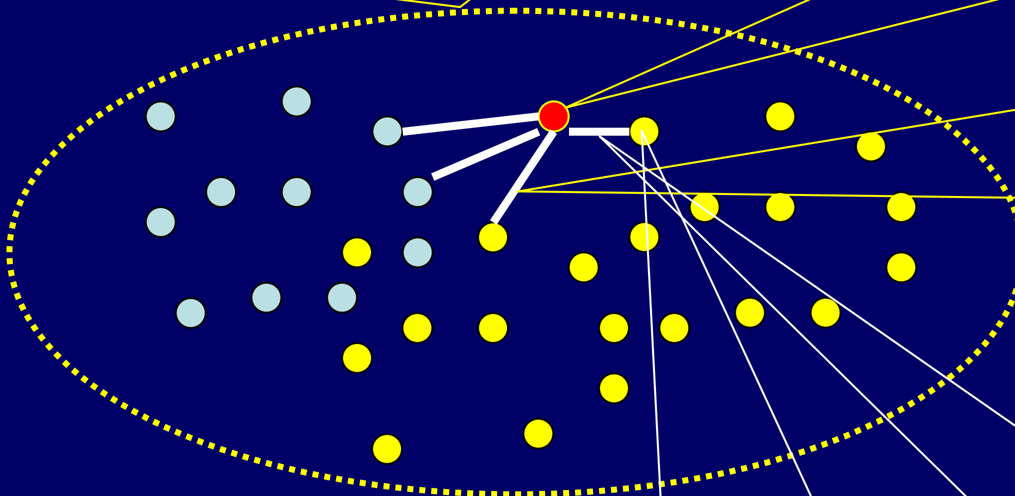
(3) 计算x到 $x_i \in X$, $(i=1,2,\ldots,N)$ 的距离 $d_i(x)$

(4) 找出最小距离 $d_m(x)=\min\{d_i(x)\}$

3.6 最近邻方法 1-NN

(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 $x_i \in X$, $(i=1,2,\ldots,N)$ 的距离 $d_i(x)$

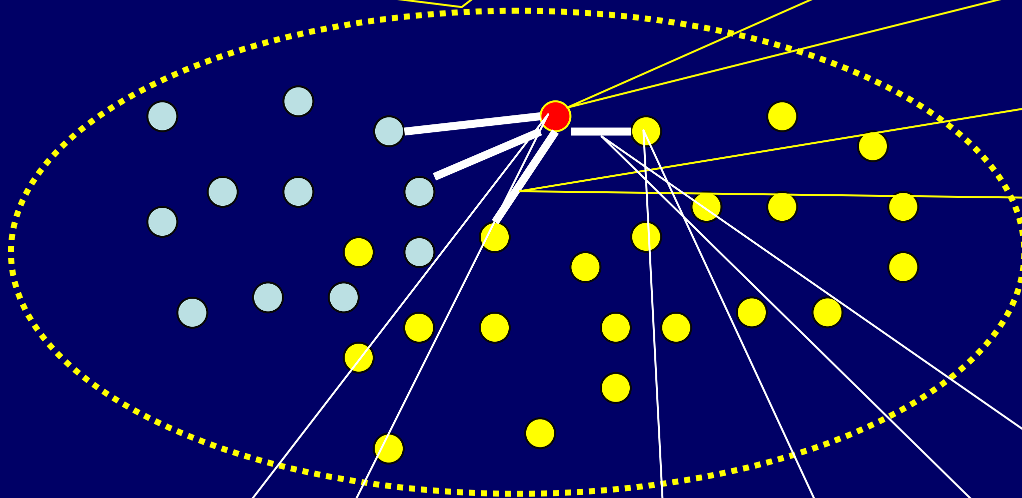
(4)找出最小距离 $d_m(x)=\min\{d_i(x)\}$

(5)看x_m属于哪一类：$x_m \in \omega_2$

3.6 最近邻方法　　1-NN

(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$x_i \in X$，$(i=1,2,\ldots,N)$的距离$d_i(x)$

(4)找出最小距离$d_m(x)=\min\{d_i(x)\}$

(5)看x_m属于哪一类：$x_m \in \omega_2$

(6) 判$x \in \omega_2$

3.6 最近邻方法

3.6.1最近邻决策规则—1-NN

对于C类问题,设类ω_i(i=1,2,…,c)有N_i个样本 $x_j^{(i)}$

分类的思想是: $(j=1,2,\cdots,N_i)$

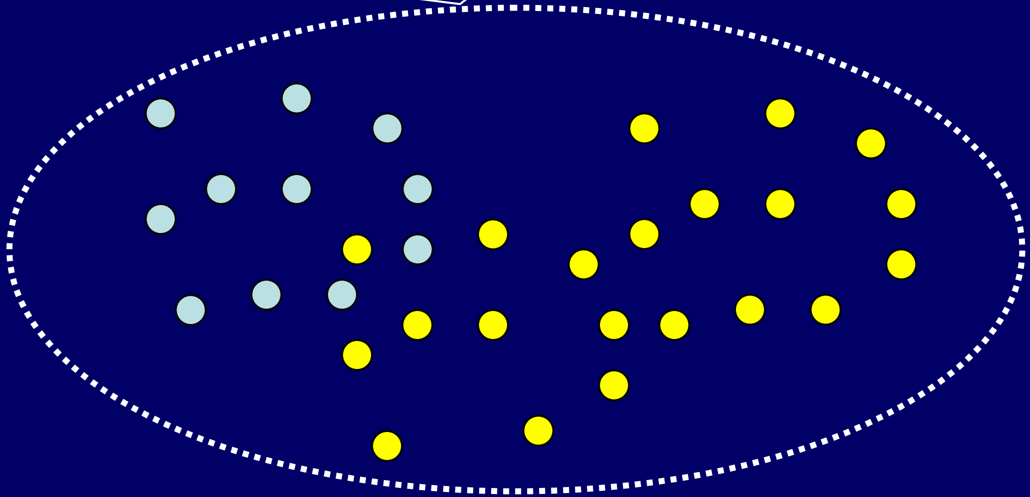
对于一个待识别模式**x**, 分别计算它与 $N=\sum_{i=1}^{c}N_i$ 个已知类别的样本 $x_j^{(i)}$的距离,将它判为距离最近的那个样本所属的类。即:

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

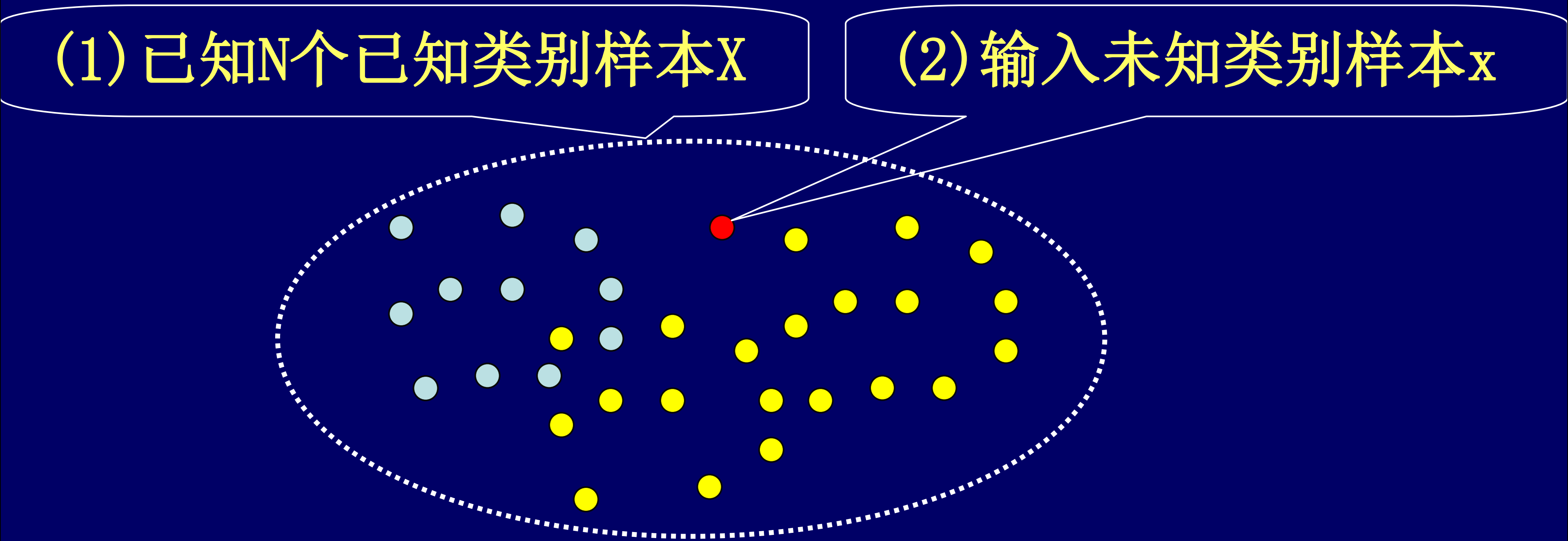
如果 $d_m(x)=\underbrace{\min}_{i=1,2,\cdots,c}d_i(x)$ 则 $x\in\omega_m$

3.6 最近邻方法　　K-NN

(1) 已知N个已知类别样本X



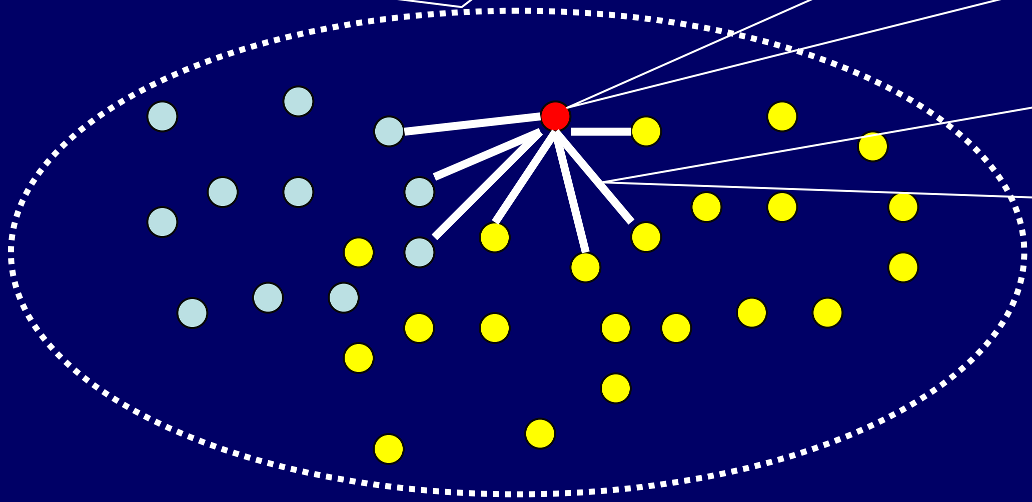
3.6 最近邻方法 K-NN



3.6 最近邻方法　K-NN

(1) 已知N个已知类别样本X

(2) 输入未知类别样本x

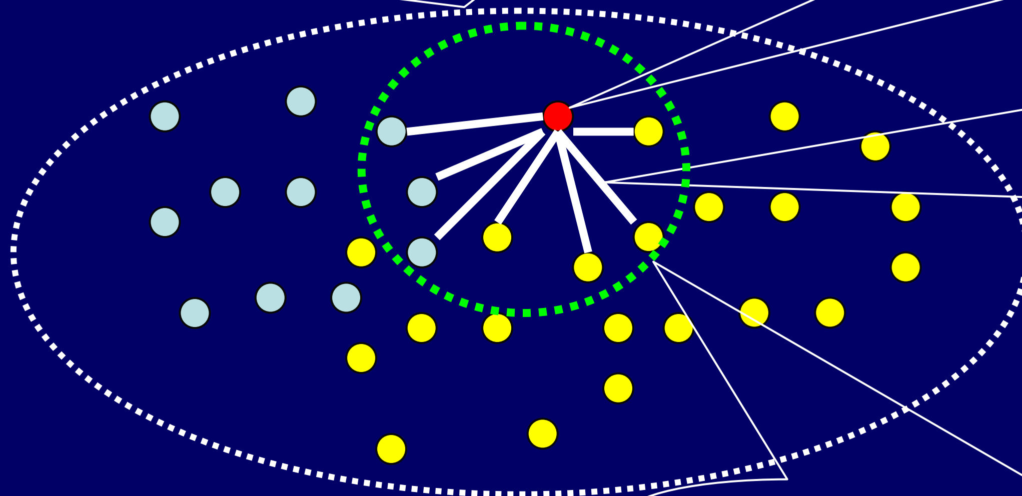


(3) 计算x到 $x_i \in X$, $(i=1, 2, \ldots, N)$ 的距离 $d_i(x)$

3.6 最近邻方法 K-NN

(1) 已知N个已知类别样本X

(2) 输入未知类别样本x



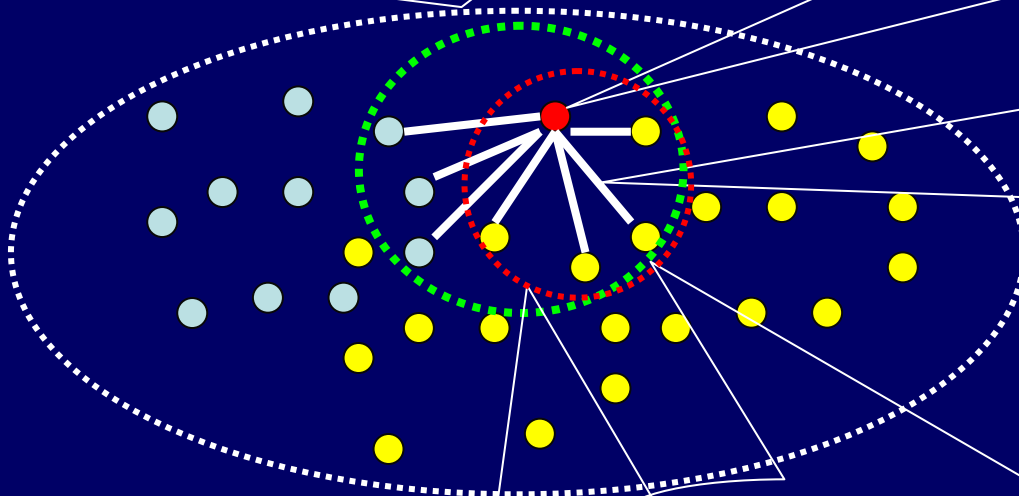
(3) 计算x到 $x_i \in X$, $(i=1, 2, \ldots, N)$ 的距离 $d_i(x)$

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

3.6 最近邻方法 K-NN

(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$x_i \in X$，$(i=1,2,\ldots,N)$的距离$d_i(x)$

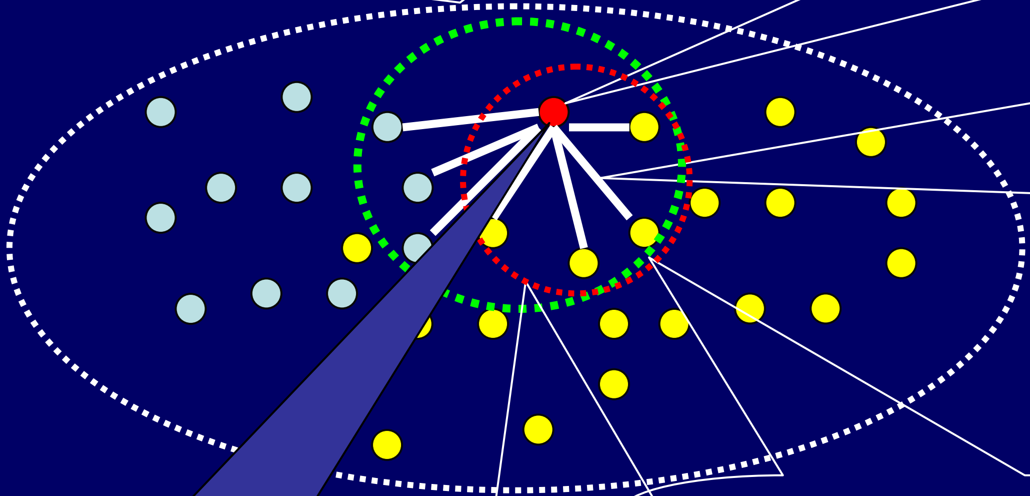
(4)找出x的k个最近邻元 $X_k=\{x_i, i=1,2,\ldots,k\}$

(5)看X_k中属于哪一类的样本最多 $k_1=3<k_2=4$

3.6 最近邻方法 K-NN

(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 $x_i \in X$, $(i=1, 2, \ldots, N)$ 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 $x \in \omega_2$

3.6 最近邻方法

3.6.1最近邻决策规则—K-NN

对于一个待识别模式**x**, 分别计算它与 $N=\sum_{i=1}^{c}N_i$ 个已知类别的样本 $x_j^{(i)}$ 的距离, 取k个最近邻样本, 这k个样本中哪一类最多, 就判属哪一类。即:

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

显然 $$\sum_{i=1}^{c}k_i=k$$

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

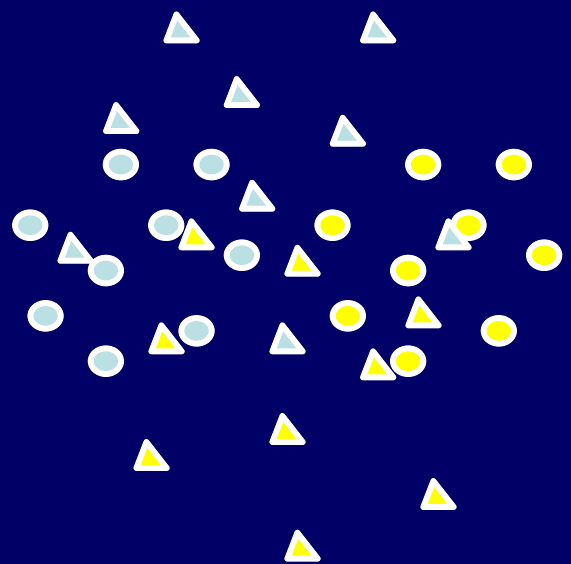
3.6 最近邻方法

■ $\in \omega_1$

■ $\in \omega_2$

○ $\in X^{(NR)}$

△ $\in X^{(NT)}$



3.6 最近邻方法

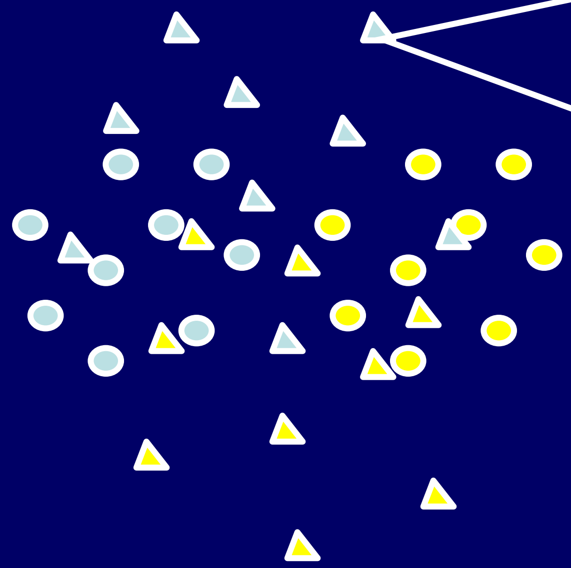
剪辑最近邻方法

■ $\in \omega_1$

■ $\in \omega_2$

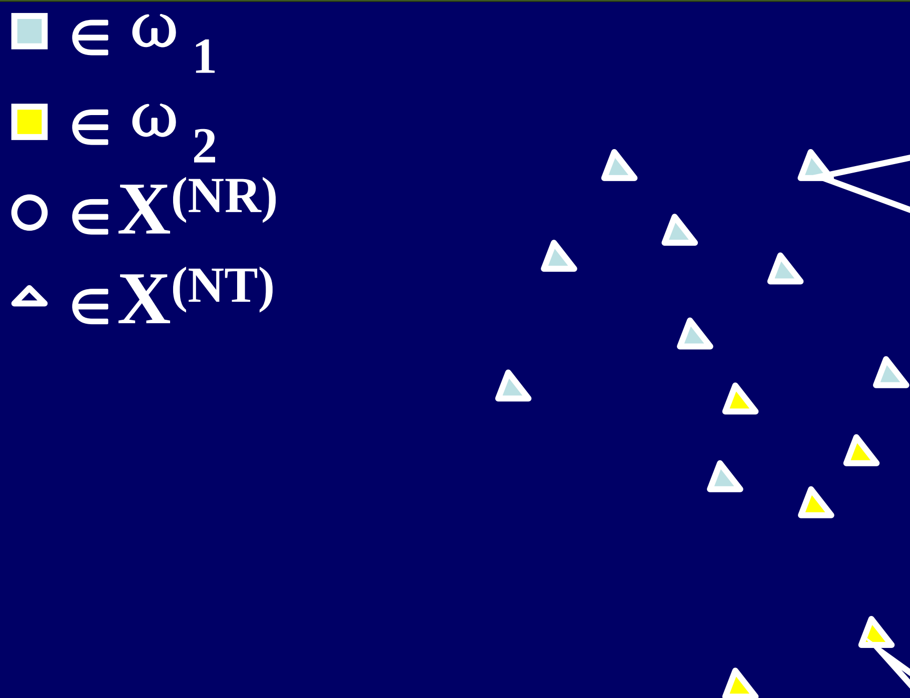
○ $\in X^{(NR)}$

△ $\in X^{(NT)}$



用$X^{(NR)}$中的样本采用最近邻规则对$X^{(NT)}$中的每个样本分类，剪辑掉$X^{(NT)}$中被错误分类的样本。

3.6 最近邻方法



用$X^{(NR)}$中的样本采用最近邻规则对$X^{(NT)}$中的每个样本分类，剪辑掉$X^{(NT)}$中被错误分类的样本。

余下判决正确的样本组成剪辑样本集$X^{(NTE)}$。

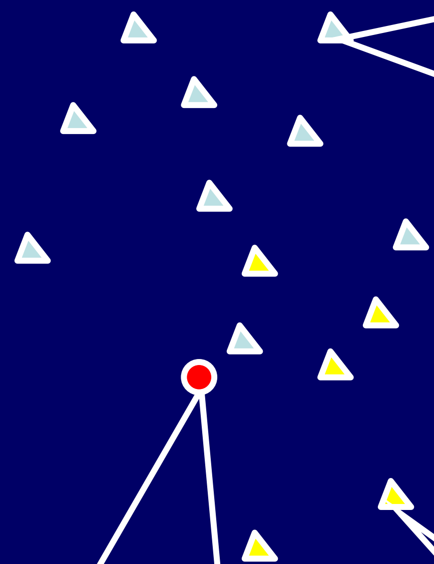
3.6 最近邻方法

■ $\in \omega_1$

■ $\in \omega_2$

○ $\in X^{(NR)}$

△ $\in X^{(NT)}$



用$X^{(NR)}$中的样本采用最近邻规则对$X^{(NT)}$中的每个样本分类，剪辑掉$X^{(NT)}$中被错误分类的样本。

余下判决正确的样本组成剪辑样本集$X^{(NTE)}$。

用$X^{(NTE)}$ 对输入的未知样本做1-NN分类。

3.6 最近邻方法

3.6.2 剪辑最近邻方法

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，它们的样本数各为*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。

3.6.2 剪辑最近邻方法

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。

3.6 最近邻方法

3.6.2 剪辑最近邻方法

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。

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

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

这里 $x_j \in X^{(NTE)}$

3.6 最近邻方法

3.6.2 剪辑最近邻方法

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。

3.6 最近邻方法 MULTIEDIT实用算法

（1） 将样本集 $X^{(N)}$ 随机地划分为**s**个子集：

$$X^{(N)}=\{X_1, X_2, \cdots, X_s\} \qquad (s \geq 3)$$

（2） 用最近邻法，以 $X_{(i+1)\,\mathrm{mod}\,s}$ 为参照集，对 X_i 中的样本进行分类，其中***i*** **=1,2,…,s;**

（3）去掉（2）中被错误分类的样本；

（4）用所留下的样本构成新的样本集 $X^{(NE)}$；

（5）如果经过***k*** 次迭代再没有样本被剪辑掉则停止；否则转至（1）。

有七个二维矢量：

$$X^{(1)} = \{\vec{x}_1 = (1,0)', \vec{x}_2 = (0,1)', \vec{x}_3 = (0,-1)'\} \in \omega_1$$

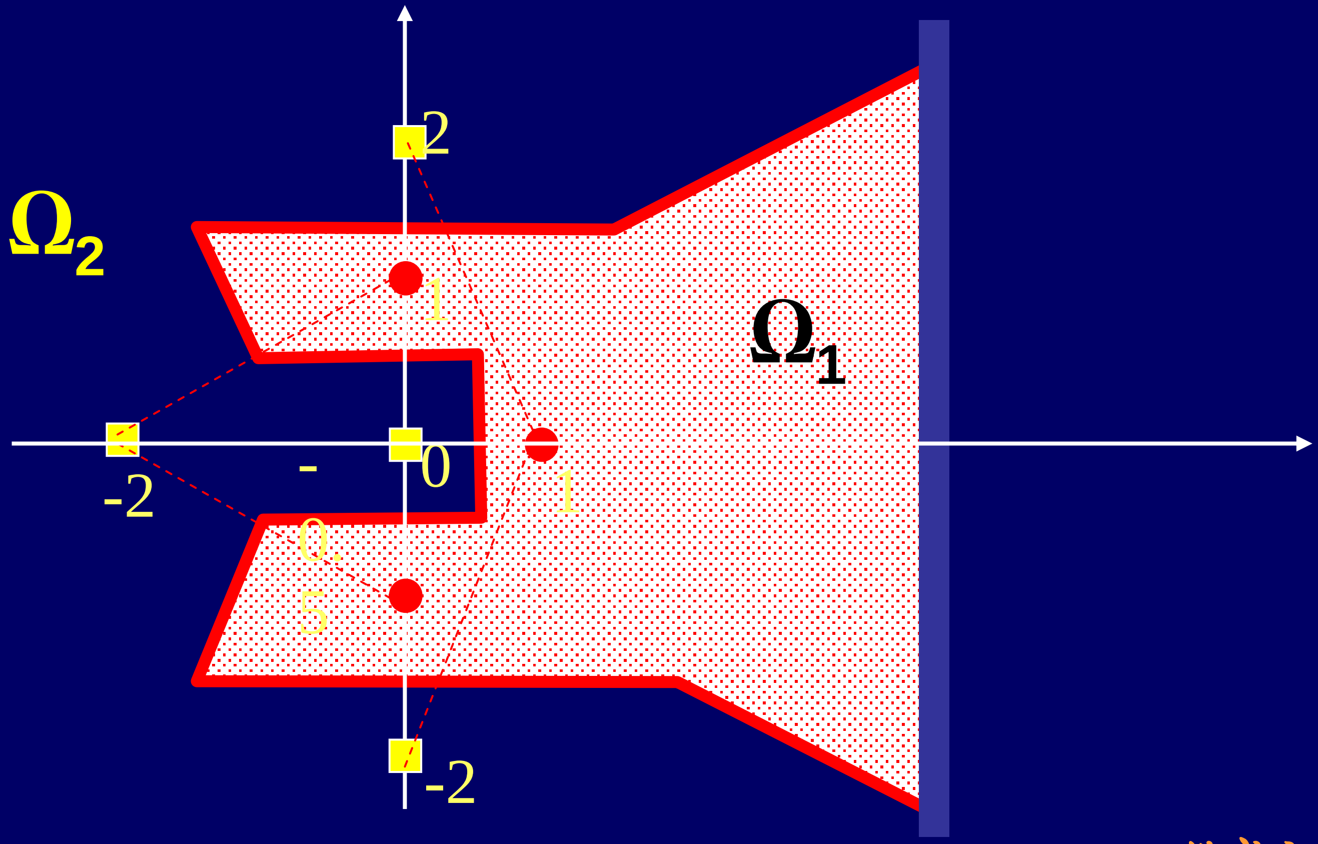
$$X^{(2)} = \{\vec{x}_4 = (0,0)', \vec{x}_5 = (0,2)', \vec{x}_6 = (0,-2)', \vec{x}_7 = (-2,0)'\} \in \omega_2$$

（1）画出最近邻法1-NN决策面；
（2）求样本均值，若按离样本均值距离的大小进行分类，试画出决策面。

例子

（1）若按1-NN进行分类，则决策面由 $\forall \vec{x}_i \in X^{(1)}$

与 $\forall \vec{x}_j \in X^{(2)}$ **连线的中垂面构成。**



3.6 最近邻方法

例子

（2）
$$\vec{m}_1=\frac{1}{3}(\vec{x}_1+\vec{x}_2+\vec{x}_3)=(\frac{1}{3},0)'$$
$$\vec{m}_2=\frac{1}{4}(\vec{x}_4+\vec{x}_5+\vec{x}_6+\vec{x}_7)=(-\frac{1}{2},0)'$$

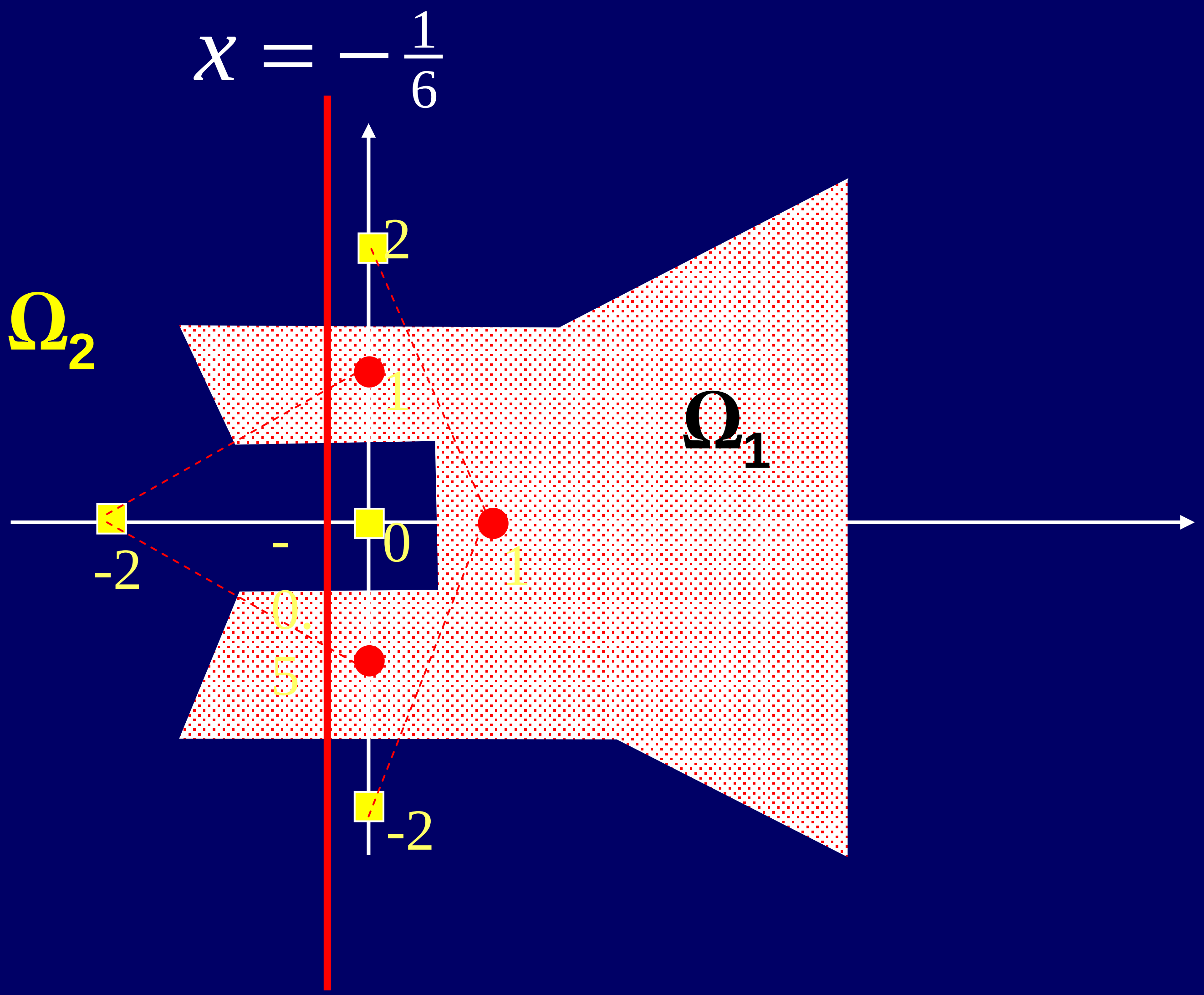
则按离样本均值距离的大小进行分类的决策面是

$\vec{m}_1$与 $\vec{m}_2$连线的中垂面。

即：$x=-\frac{1}{6}$

3.6 最近邻方法

例子



谢谢！

30