Classification by Distance Functions.

- M-pattern classes with prototype patterns $z_1, z_2, \cdots, z_M$.

Euclidean distance between an arbitrary pattern vector x and the ith prototype:

$$D_i = \|x - z_i\| = \sqrt{(x - z_i)^T (x - z_i)}$$

x is assigned to class w_i if $D_i < D_j$ for all $i \neq j$.

-
$$\begin{aligned} D_i^2 &= \|x - z_i\|^2 \\ &= (x - z_i)^T (x - z_i) \\ &= x^T x - x^T z_i - z_i^T x + z_i^T z_i \qquad (x^T z_i = z_i^T x \\ &= x^T x - 2 x^T z_i + z_i^T z_i \\ &= x^T x - 2\left(\underline{x^T z_i - \tfrac{1}{2} z_i^T z_i}\right) \end{aligned}$$

- $\mathrm{Min}\{D_i^2\} \Rightarrow \mathrm{Min}\{D_i\} \Rightarrow \mathrm{Max}\{x^T z_i - \frac{1}{2} z_i^T z_i\}$

- The decision functions

$$d_i(x) = x^T z_i - \frac{1}{2} z_i^T z_i, \quad i = 1, 2, 3, \cdots, M$$

※ A pattern x is assigned to class w_i
if
$$d_i(x) > d_j(x) \text{ for all } i \neq j.$$

* Multi prototypes

- A single prototype $\rightarrow z_i \Rightarrow w_i$

Several prototypes $\rightarrow z_i \Rightarrow w_i$ $\quad$ N_i : the # of prototypes in the ith pattern class.

$$z_i^1, \; z_i^2, \; \vdots, \; z_i^{N_i}$$

- The distance function between an arbitrary pattern x and class w_i

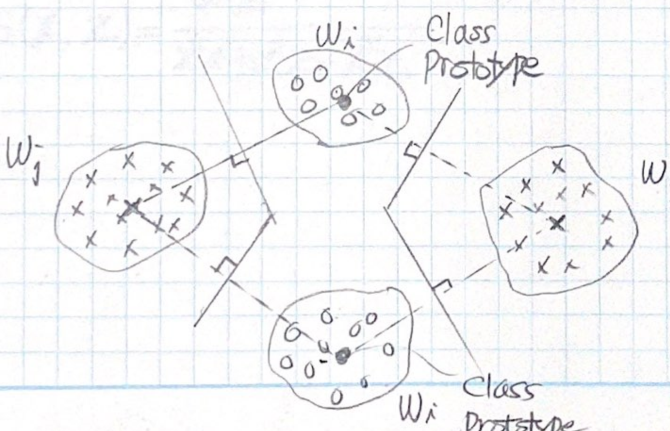
$$D_i = \min_k \| x - z_i^k \|, \quad k = 1, 2, \cdots, N_i$$

- The distances D_i, $i = 1, 2, \cdots, M$, are computed and the unknown is classified into w_i if $D_i < D_j$ for all $i \neq j$.

- The decision functions

$$d_i(x) = \max_k \left\{ x^T z_i^k - \frac{1}{2} z_i^{k^T} z_i^k \right\} \quad k = 1, 2, \cdots, N_i$$

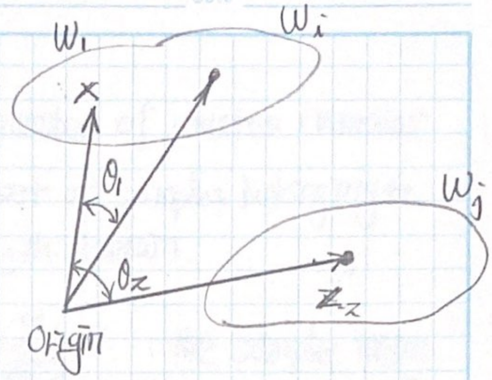
where x is placed in class w_i if $d_i(x) > d_j(x)$, for all $i \neq j$.



Piecewise-linear decision boundaries for two classes, each of which is characterized by two prototypes

Cluster Seeking

Measure of Similarity



$$S(x, z_1) = \cos\theta_1 = \frac{x^T z_1}{\|x\| \|z_1\|}$$

$$S(x, z_2) = \cos\theta_2 = \frac{x^T z_2}{\|x\| \|z_2\|}$$

[A similarity measure]

- Euclidean distance

$$D = \|x - z\|$$

- Mahalanobis distance from x to m

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

↑ the mean vector

↑ The covariance matrix

- The nonmetric similarity function;

$$s(x, z) = \frac{x^T z}{\|x\| \|z\|}$$

- Tanimoto measure

$$S(x, z) = \frac{x^T z}{x^T x + z^T z - x^T z}$$

Clustering Criteria

$$J = \sum_{j=1}^{N_c} \sum_{x \in S_j} \| x - m_j \|^2$$

N_c : the number of cluster domains

S_j : the set of samples belonging to the jth domain.

$m_j = \frac{1}{N_j} \sum_{x \in S_j} x$: the sample mean vector of set S_j.

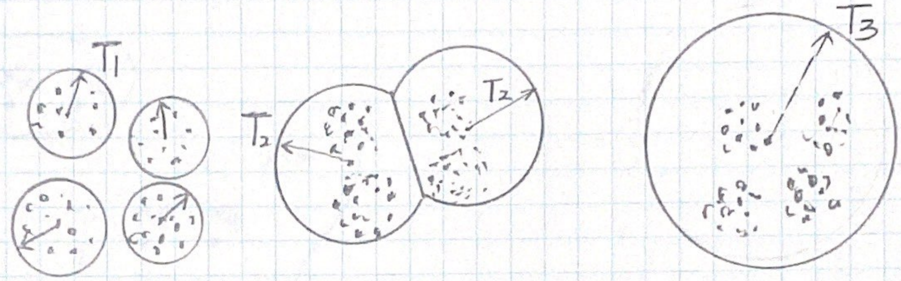
N_j : the number of samples in S_j.

$\Rightarrow$ The overall sum of the squared errors between the samples of a cluster domain and their corresponding mean.

Course ______ Sheet ______ of ______

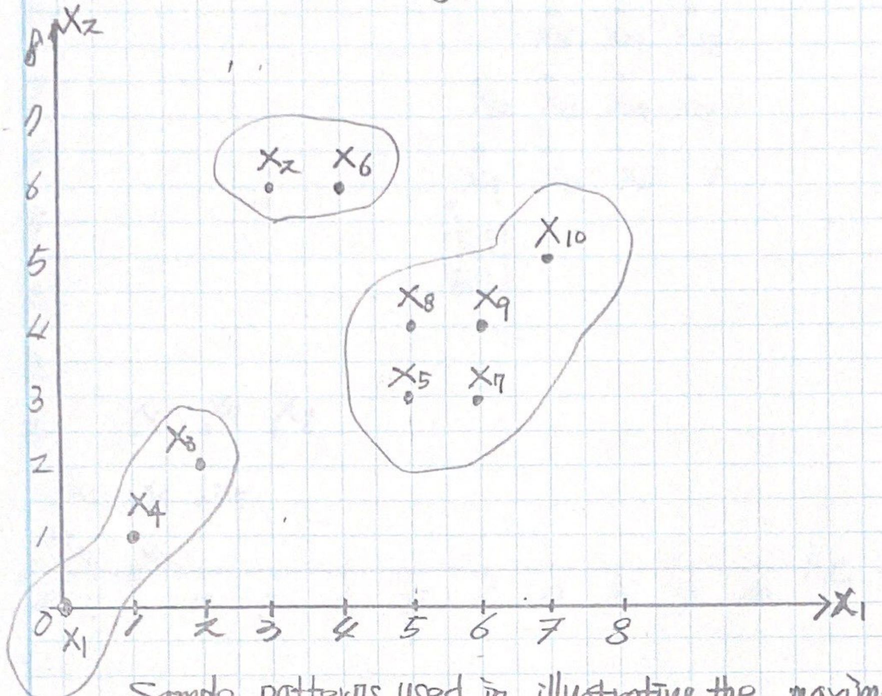
By ______ Date ______

A simple cluster-seeking Algorithm

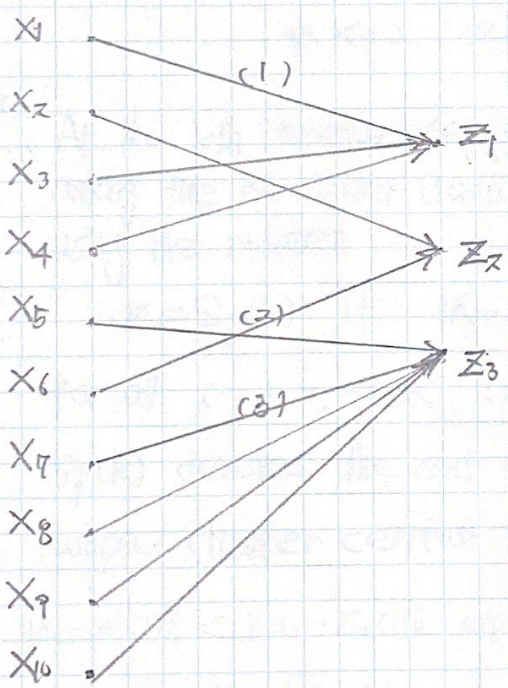


Effects of the threshold and starting points in a simple cluster seeking scheme.

* Maximin - Distance Algorithm (Maximum-Minimum) distance algorithm

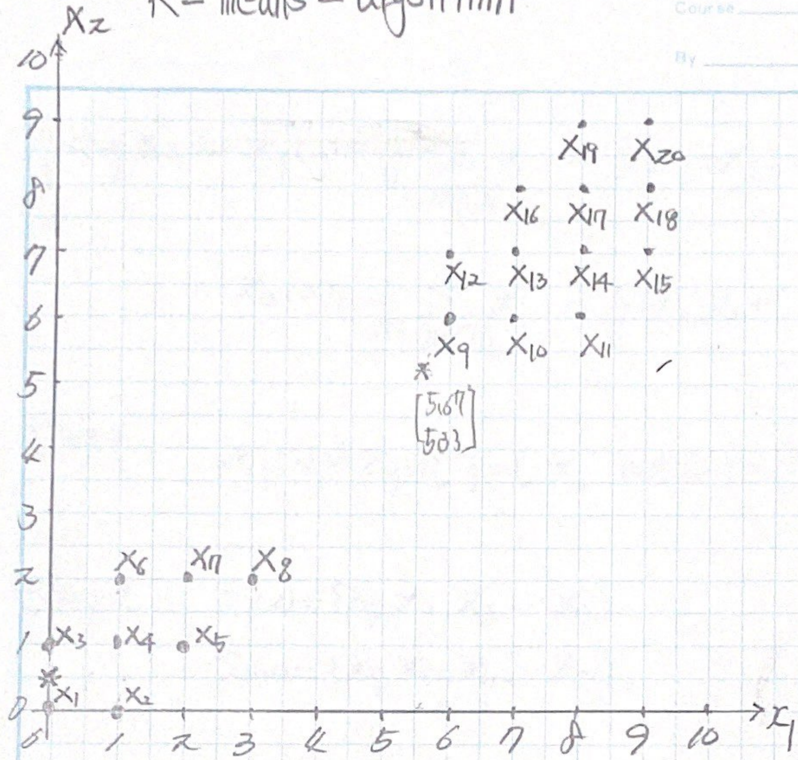


Sample patterns used in illustrating the maximum distance algorithm



Sample & Category tables

K- means - algorithm



Step 1. Choose K initial cluster centers $z_1(1), z_2(1), \cdots, z_K(1)$.

$$* \ K=2, \quad z_1(1) = x_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad z_2(1) = x_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Step 2, At the kth iterative step distribute the samples $\{x\}$ among the K cluster domains, using the relation

$$x \in S_j(k) \quad \text{if} \quad \|x - z_j(k)\| < \|x - z_i(k)\|$$

for all $i = 1, 2, \cdots, K$, $i \neq j$,

$S_j(k)$ denotes the set of samples whose cluster center is $z_j(k)$.

* $\|x_1 - z_1(1)\| < \|x_1 - z_i(1)\|$ and $\|x_3 - z_1(1)\| < \|x_3 - z_i(1)\|$,

$i = 2$, $S_1(1) = \{x_1, x_3\}$.

Similarly, the remaining patterns are closer to $z_2(1)$,

so $S_2(1) = \{x_2, x_4, x_5, \cdots, x_{20}\}$

Course ______ Sheet ______ of ______

By ______ Date ______

Step 3. $Z_1(2) = \frac{1}{N_1} \sum_{X \in S_1(1)} X = \frac{1}{2}(X_1 + X_2) = \frac{1}{2}\left[\begin{pmatrix}0\\0\end{pmatrix} + \begin{pmatrix}0\\1\end{pmatrix}\right] = \begin{bmatrix}0\\0.5\end{bmatrix}$

$Z_2(2) = \frac{1}{N_2} \sum_{X \in S_2(1)} X = \frac{1}{18}(X_3 + X_4 + \cdots + X_{20}) = \begin{bmatrix}5.67\\5.33\end{bmatrix}$

Step 4. $Z_j(2) \neq Z_j(1)$, $j = 1, 2$, return to Step 2

Step 2. $S_1(2) = \{X_1, X_2, X_3 \cdots X_8\}$

$S_2(2) = \{X_9, X_{10}, X_{11}, \cdots X_{20}\}$.

Step 3, Update the cluster centers:

$Z_1(3) = \frac{1}{8}(X_1 + \cdots + X_8) = \begin{bmatrix}1.25\\1.13\end{bmatrix}$

$Z_2(3) = \frac{1}{12}(X_9 + \cdots + X_{20}) = \begin{bmatrix}7.67\\7.33\end{bmatrix}$

Step 4. $Z_j(3) \neq Z_j(2)$, Return to step 2.

Step 2; $Z_1(4) = Z_1(3)$ and $Z_2(4) = Z_2(3)$

$\Rightarrow$ Cluster centers: $Z_1 = \begin{bmatrix}1.25\\1.13\end{bmatrix}$, $Z_2 = \begin{bmatrix}7.67\\7.33\end{bmatrix}$ ▨