Hybrid Intelligent Systems

Lecture 5. Part 2.

Unsupervised NN for clustering.
Adaptive Resonance Theory

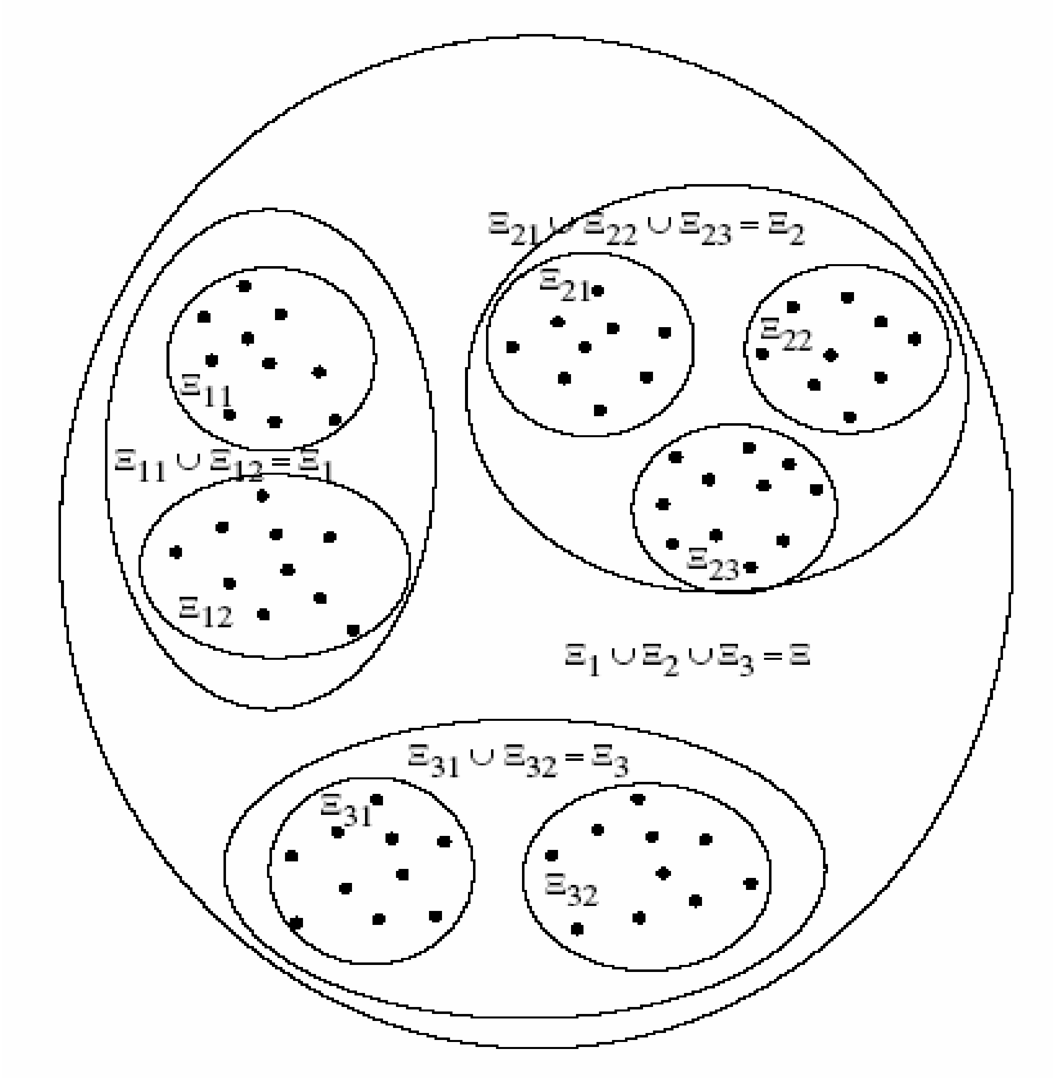
Definition

- Clustering is the process of partitioning a set of objects into subsets based on some measure of similarity (or dissimilarity) between pairs of the objects.

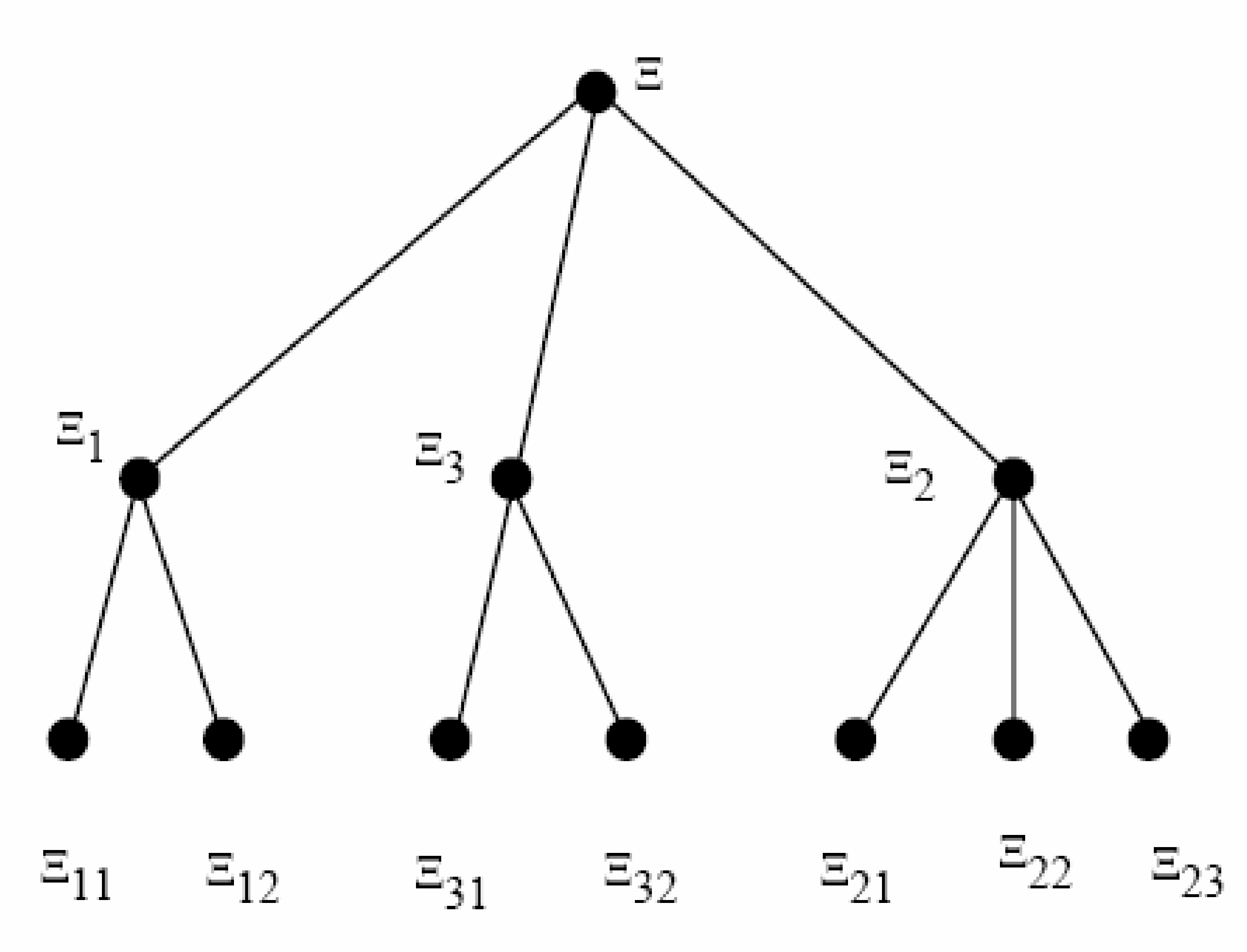
Cluster Analysis

- Goals
  - Organize information about data so that relatively homogeneous groups (clusters) are formed and describe their unknown properties.
  - Find useful and interesting groupings of samples.
  - Find representatives for homogeneous groups.
- Two components of cluster analysis.
  - The (dis)similarity measure between two data samples.
  - The clustering algorithm.

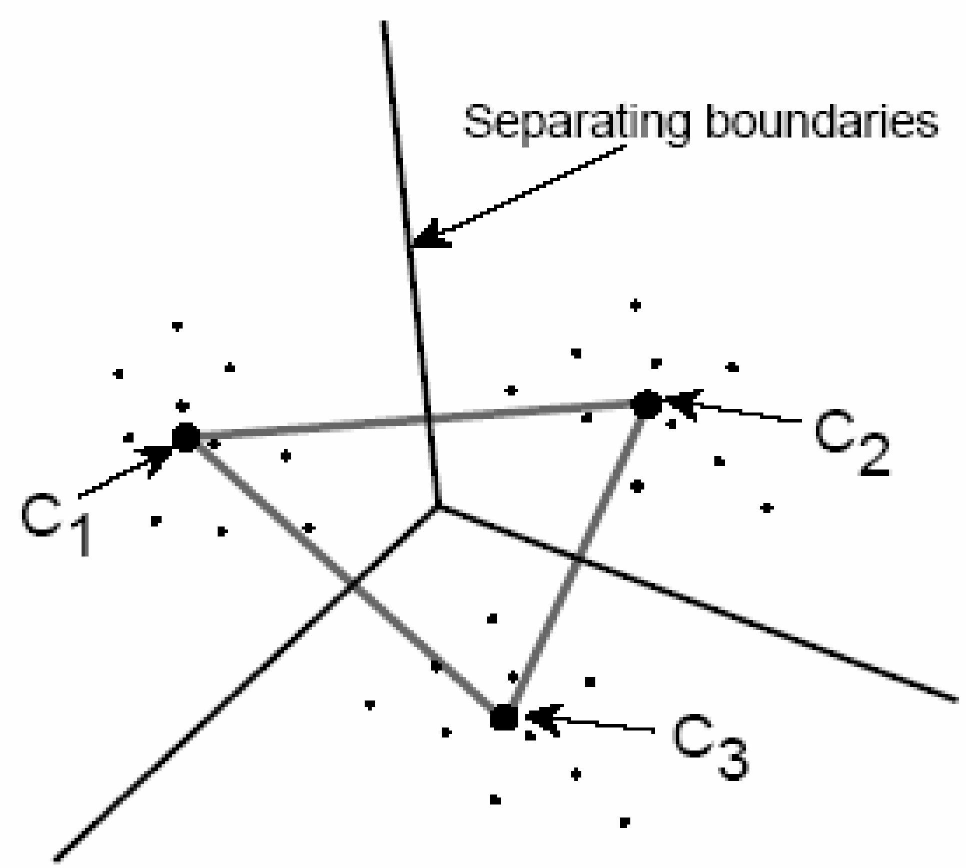
Hierarchy of clusters



Display of hierarchy as tree



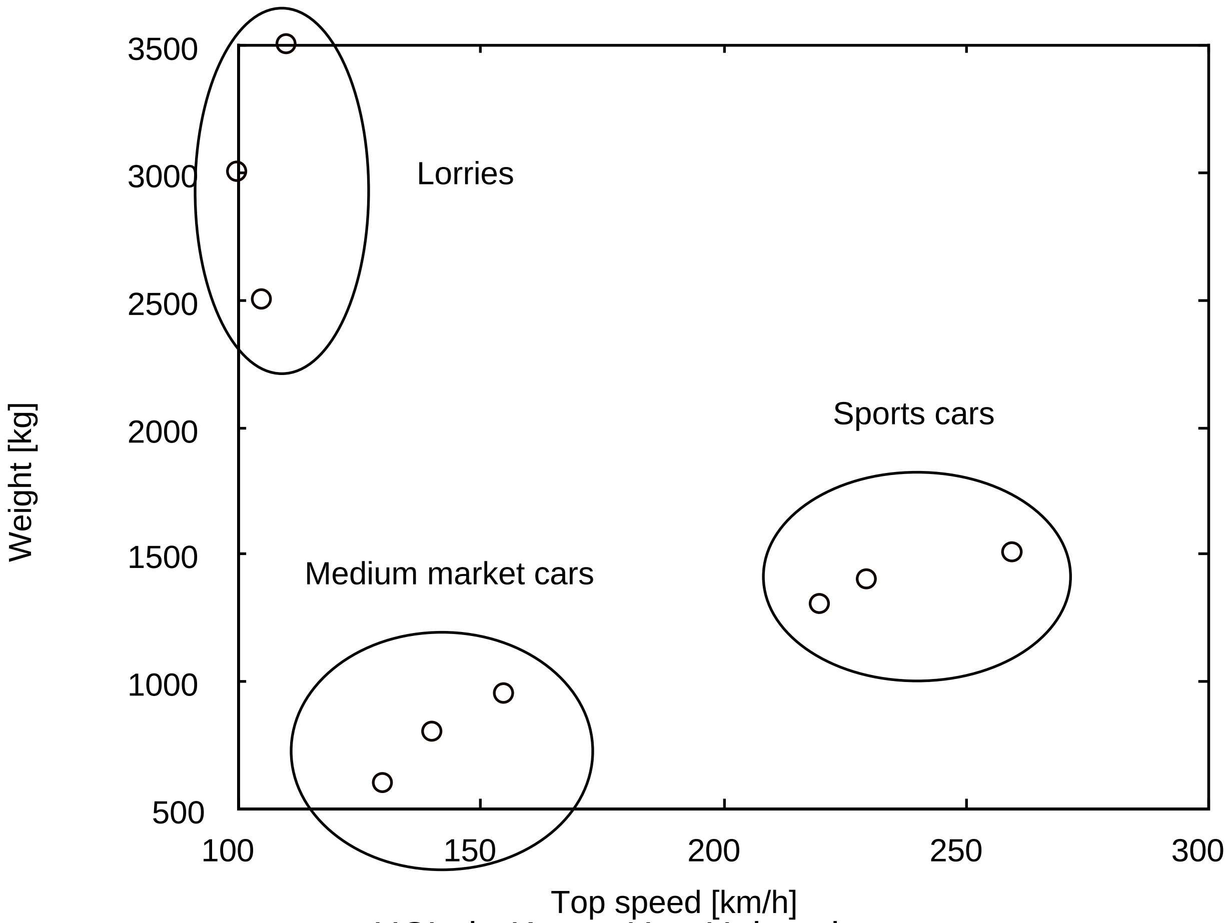
Minimum-Distance Clustering



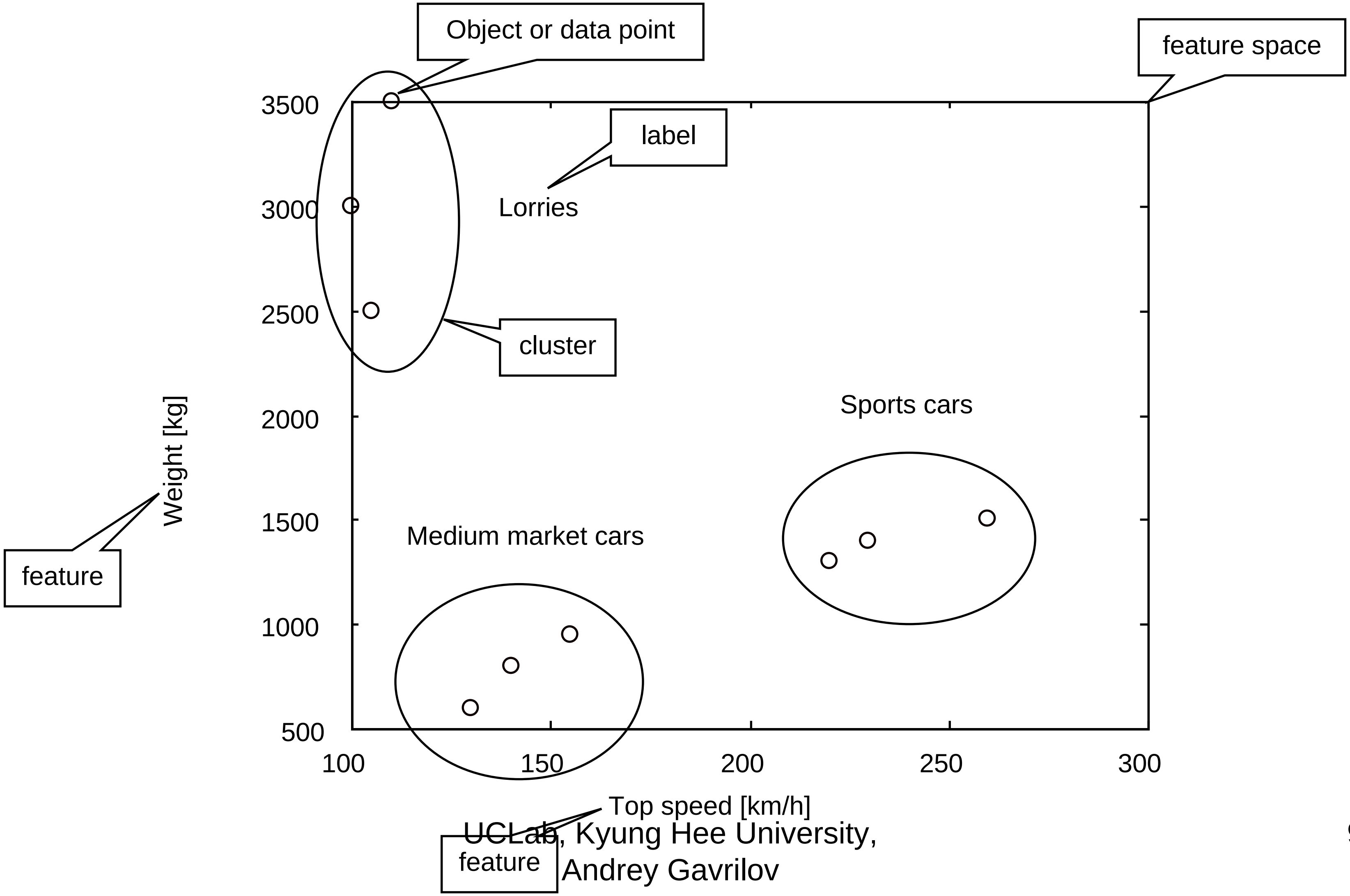
Vehicle Example

Vehicle	Top speed km/h	Colour	Air resistance	Weight Kg
V1	220	red	0.30	1300
V2	230	black	0.32	1400
V3	260	red	0.29	1500
V4	140	gray	0.35	800
V5	155	blue	0.33	950
V6	130	white	0.40	600
V7	100	black	0.50	3000
V8	105	red	0.60	2500
V9	110	gray	0.55	3500

Vehicle Clusters



Terminology



Distance/Similarity Measures

- **Minkowski** metric

$$dist(\mathbf{x}_i, \mathbf{x}_k) = \left(\sum_{j=1}^{d} \left| x_{ij} - x_{kj} \right|^r \right)^{1/r} \qquad \mathbf{x}_i = \left(x_{i1}, x_{i2}, \ldots, x_{id} \right)^T$$

1. $\forall i,\ dist(i,i) = 0$
2. $\forall (i,k) \quad dist(i,k) \geq 0$
3. $\forall (i,k) \quad dist(i,k) = dist(k,i)$
4. $\forall (i,k,m) \quad dist(i,k) \leq dist(i,m) + dist(m,k)$

Distance/Similarity Measures (2)

- Common Minkowski metrics
  - Euclidean distance: $r = 2$

    $$dist(i,k) = \left[\sum_{j=1}^{d} (x_{ij} - x_{kj})^2 \right]^{1/2}$$

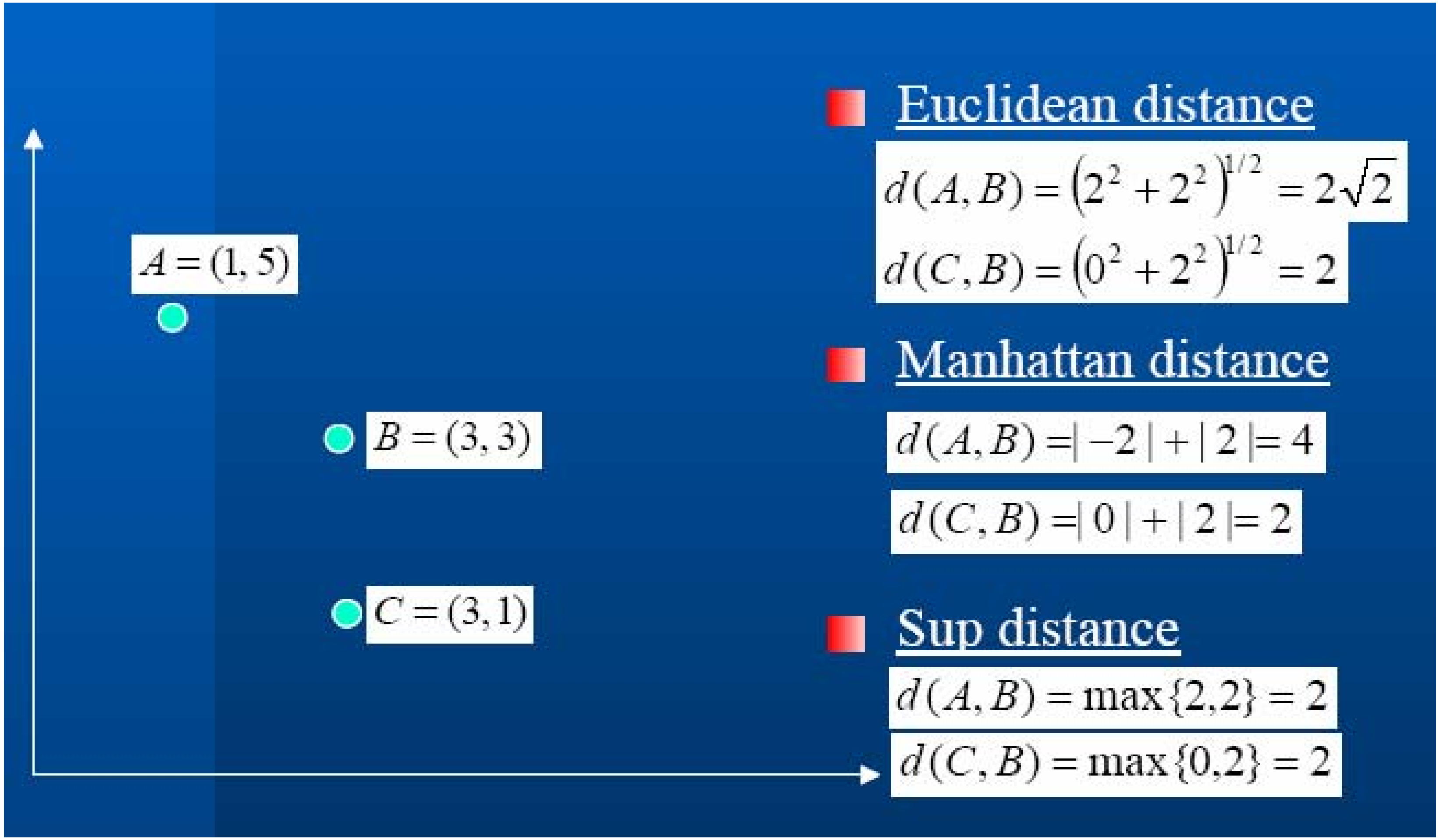
  - Manhattan distance: $r = 1$

    $$dist(i,k) = \sum_{j=1}^{d} \left| x_{ij} - x_{kj} \right|$$

    - If all the features are binary, Hamming distance.
  - Sup distance: $r \to \infty$

    $$dist(i,k) = \max_{1 \le j \le d} \left| x_{ij} - x_{kj} \right|$$

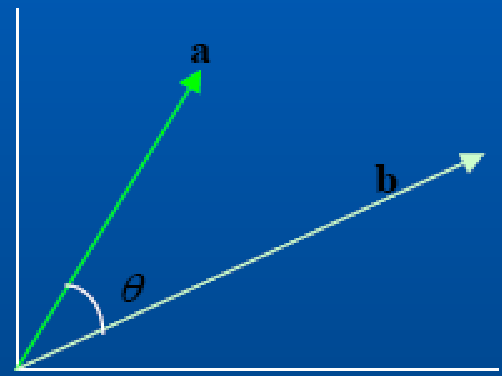
Distance/Similarity Measures (3)



Distance/Similarity Measures (4)

- Cosine similarity
  - Estimate the angle between two data items represented in vector space.

$$SIM(\mathbf{a},\mathbf{b}) = \cos\theta = \frac{\mathbf{a}\bullet\mathbf{b}}{|\mathbf{a}|\times|\mathbf{b}|}$$



- Pearson correlation coefficient
  - Represent linear relationship between two data items.

$$SIM(\mathbf{a},\mathbf{b}) = \frac{\sum_{i=1}^{d}\mathbf{a}_i\mathbf{b}_i - \frac{1}{d}\sum_{i=1}^{d}\mathbf{a}_i\sum_{j=1}^{d}\mathbf{b}_j}{\sqrt{\sum_{i=1}^{d}\mathbf{a}_i^2 - \frac{1}{d}\left(\sum_{i=1}^{d}\mathbf{a}_i\right)^2}\sqrt{\sum_{i=1}^{d}\mathbf{b}_i^2 - \frac{1}{d}\left(\sum_{i=1}^{d}\mathbf{b}_i\right)^2}}$$

Adaptive Resonance Theory (ART)

- One of the nice features of human memory is its ability to learn many new things without necessarily forgetting things learned in the past.
- Stephen Grossberg: Stability-Plasticity dilemma

(1) How can a learning system remain adaptive (plastic) in response to significant input, yet remain stable in response to irrelevant input?

(2) How does the system known to switch between its plastic and its stable modes?

(3) How can the system retain previously learned information while continuing to learn new things?

Basic Concept of ART

- A key to solving the stability-plasticity dilemma is to add a feedback mechanism between the competitive layer and the input layer of a network.
- Grossberg and Carpenter: ART model
- ART is one of the unsupervised learning models.
- This kind of model was first established in the early 1960.
- Grossberg introduced the ART in 1976.
- G.A. Carpenter continued the research in ART.
- Now many modifications of ART exist: ART-1, ART-2, FuzzyART, ARTMAP,

ART Based Architectures

- Adaptive Resonance Theory by Grossberg (1976)
- Family of ART neural network architectures
  - unsupervised data classification
    - ART 1 → binary patterns (1987)
    - ART 2 → analog patterns (1987)
    - fuzzy ART → generalization of ART1 in fuzzy set domain
  - supervised mapping
    - ART-MAP
    - fuzzy ART-MAP (1992)

Basic Concept of ART (2)

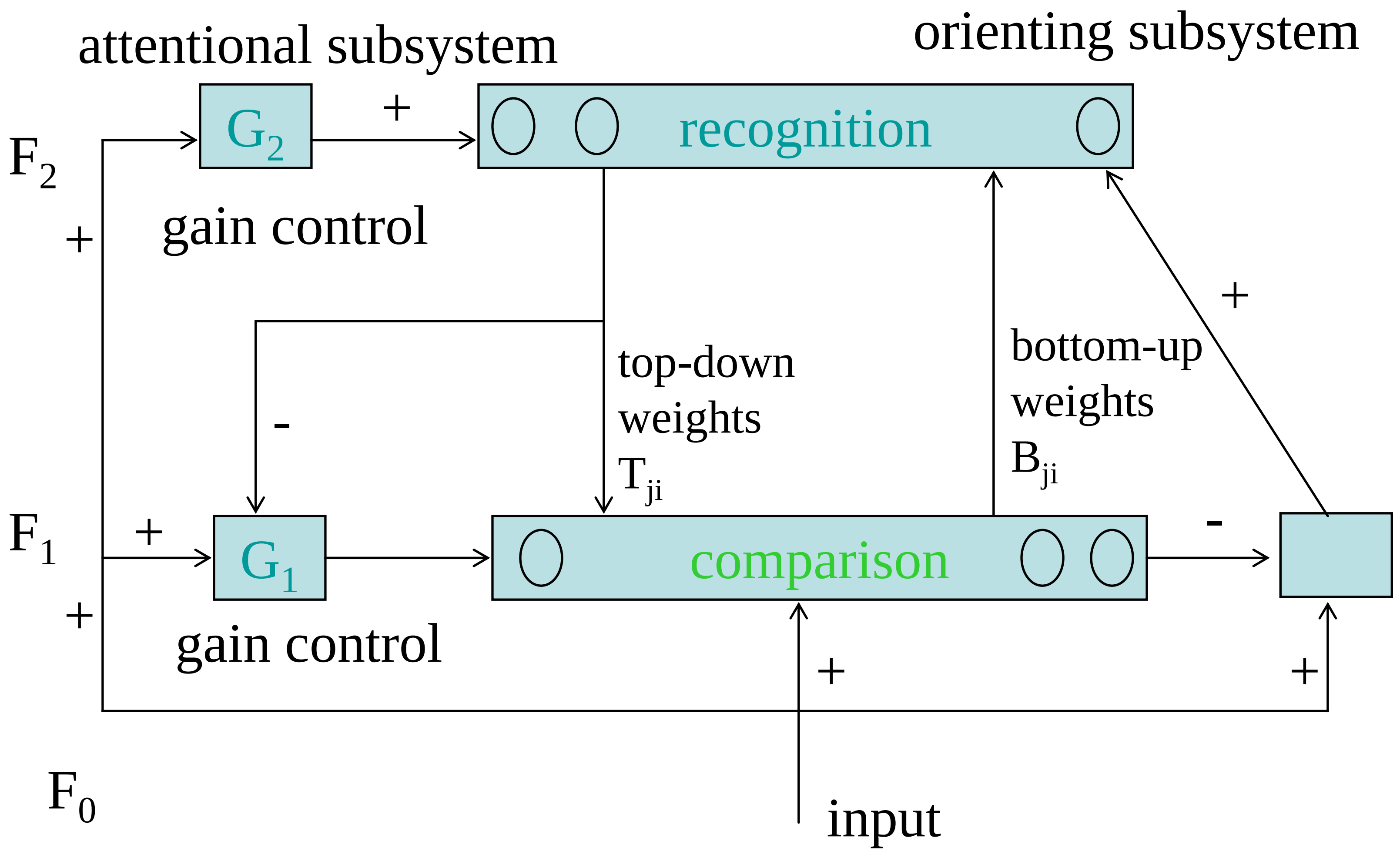
- ART 1: requires that the input vectors be binary
- ART 2: is suitable for processing analog, or gray scale, patterns
- ART gets its name from the particular way in which learning and recall interplay in the network.
- In physics, resonance occurs when a small-amplitude vibration of the proper frequency causes a large-amplitude vibration in an electrical or mechanical system.

Basic Concept of ART (3). Basic algorithm of ART

- Step 1: Initialization. Start with no cluster prototype vectors
- Step 2: Apply new input vector I
- Step 3: Find the closest cluster prototype vector (if any) P
- Step 4: If P is too far from I then, create a new cluster, returning to step 2
- Step 5: Update the matched prototype vector (update P by moving it closer to I)

Basic Concept of ART (4).

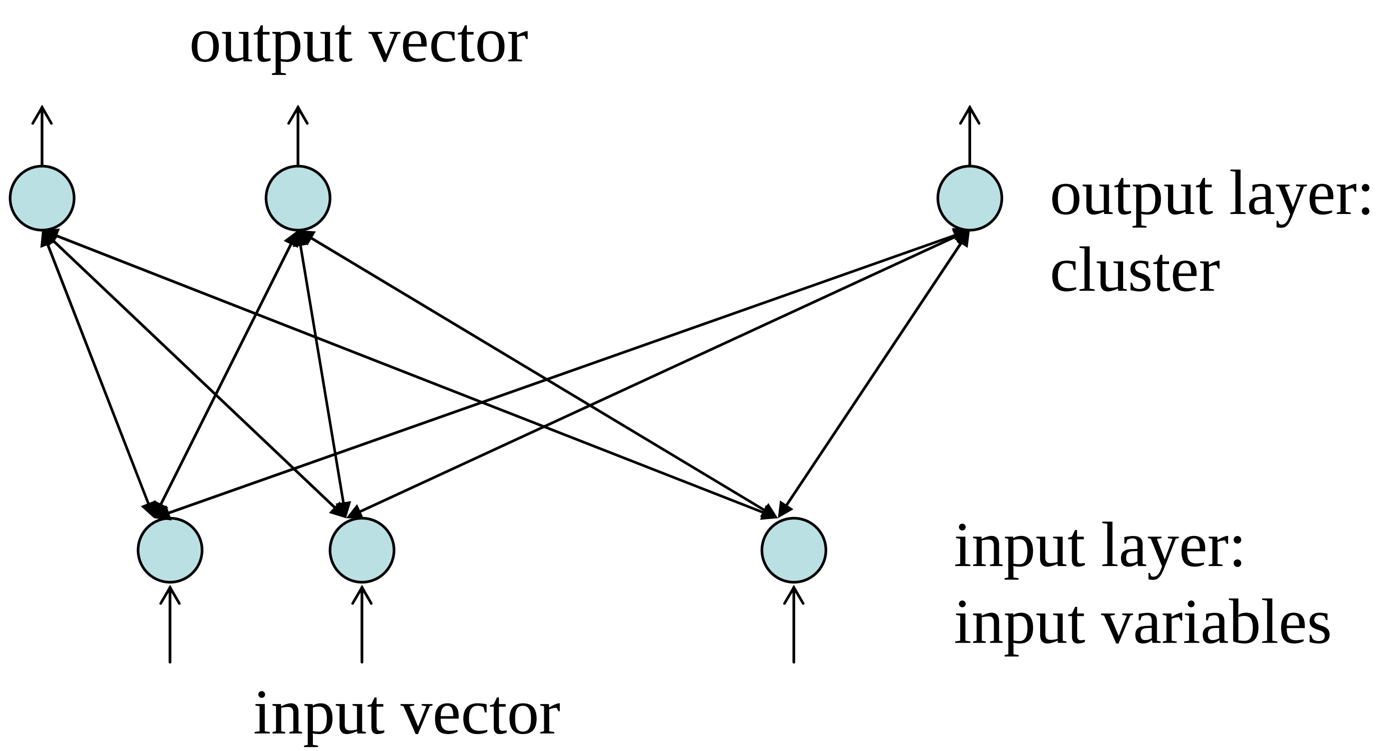
The ART network (Carpenter and Grossberg 1988).



Basic Concept of ART (5)

- B_{ji}: Forward the output from F_1 to F_2 for competition.
- T_{ji}: Forward the pattern of winner neuron to F_1 for comparison.
- G_1: To distinguish the feature of input pattern with stored patterns.
- G_2: To reset the depressed neurons in F_2 (i.e., reset losers).
- attentional subsystem: to rapidly classify the recognized patterns.
- orienting subsystem: to help attentional subsystem learn new patterns.

ART-1 Model


output vector
output layer:
cluster
input layer:
input variables
input vector

ART-1 Model (2)

- Input layer: input patterns or characteristic vectors. Activation function f(x)=x. inputs are binary values.
- Output layer: representing the clustering of training patterns. This is similar to SOFM except that SOFM has the neighborhood concept. Initially, there is only one output node. The number of output nodes increases when learning proceeds. When the stability is achieved, the learning process stops.

ART-1 Model (3). Algorithm

1. Set the network parameter: Nout=1.
2. Set the initial weighting matrices: $w^{t}[i][1] = 1$

   $$w^{b}[i][1] = \frac{1}{1+N}$$

3. Input the training vector *X*.
4. Calculate the matching value: $net[j] = \sum_{i} w^{b}[i][j] \cdot X[i]$

   $$I_{count} = 0$$

5. Find the max matching value in the output nodes:

   $$net[j^{*}] = \max_{j} net[j]$$

ART-1 Model (4). Algorithm (continue)

(6) Calculate the similarity value:

$$\| X \| = \sum_i X[i]$$

$$\| w^t_{j^*} \cdot X \| = \sum_i w^t[i][j^*] \cdot X[i]$$

$$V_{j^*} = \frac{\| w^t_{j^*} \cdot X \|}{\| X \|}$$

(7) Test the similarity value:

If $V < \rho\,(\text{vigilance})$ then go to step (8).

Otherwise go to step (9).

(8) Test whether there are output nodes applicable to the rule.

If *Icount*<*Nout*, then try the second max matching value in the output nodes.

ART-1 Model (5). Algorithm (continue)

Set *Icount*=*Icount*+1; net[j*]=0, go to step (5).

otherwise

(a) generate new cluster:

$$\text{set } N_{out} = N_{out} + 1$$

$$\text{set new weighting matrix } w$$

$$w^t[\mathrm{i}][\mathrm{N}_{\mathrm{out}}] = \mathrm{x}$$

$$\mathrm{w}^{\mathrm{b}}[\mathrm{i}][\mathrm{N}_{\mathrm{out}}] = \frac{\mathrm{x}}{0.5 + |\,\mathrm{w}^{\mathrm{t}} \cdot X\,|}$$

(b) set the output values for output nodes:

if *j*=*j**, then *Y[j]*=1

else *Y[j]*=0.

(c) go to step (3) (input new vector *X*)

ART-1 Model (6). Algorithm (continue)

(9) Adjust the weighting matrix

(a) adjust the weights:

$$w^t[\mathrm{i}][\mathrm{j}^*] = w^t[i][j^*] \cdot X[i]$$

$$\mathrm{w}^{\mathrm{b}}[\mathrm{i}][\mathrm{j}^*] = \frac{\mathrm{w}^{\mathrm{t}}[i][j^*] \cdot X[i]}{0.5 + \sum_i w^t[i][j^*] \cdot X[i]}$$

(b) set the output values for output nodes:

if *j*=*j**, then *Y[j]*=1

else *Y[j]*=0.

(c) go to step (3) (input new vector *X*)

ART-1 Model (7)

- Given an input vector:
  X=[1 1 1 1 1 0 0 0 0 0]
- Assume 5 output nodes.
  3 cases for comparisons.
- Case 1:

$$w_1^b = \frac{1}{5.5}[1,1,1,1,1,0,0,0,0,0],\ w_1^t = [1,1,1,1,1,0,0,0,0,0]$$

$$w_2^b = \frac{1}{4.5}[1,1,1,1,0,0,0,0,0,0],\ w_2^t = [1,1,1,1,0,0,0,0,0,0]$$

$$w_3^b = \frac{1}{3.5}[1,1,1,0,0,0,0,0,0,0],\ w_3^t = [1,1,1,0,0,0,0,0,0,0]$$

$$w_4^b = \frac{1}{2.5}[1,1,0,0,0,0,0,0,0,0],\ w_4^t = [1,1,0,0,0,0,0,0,0,0]$$

$$w_5^b = \frac{1}{1.5}[1,0,0,0,0,0,0,0,0,0],\ w_5^t = [1,0,0,0,0,0,0,0,0,0]$$

ART-1 Model (8)

- Node 1: matching value=5/5.5=0.909, similarity value=5/5=1.0.
- Node 2: matching value=4/4.5=0.888, similarity value=4/5=0.8.
- Node 3: matching value=3/3.5=0.857, similarity value=3/5=0.6.
- Node 4: matching value=2/2.5=0.8, similarity value=2/5=0.4.
- Node 5: matching value=1/1.5=0.667, similarity value=1/5=0.2.
- The matching value is proportional to similarity value.

ART-1 Model (9)

- Case 2:
- Assume 6 output nodes.

$$w_1^b = \frac{1}{3.5}[1,1,1,0,0,0,0,0,0,0],\ w_1^t = [1,1,1,0,0,0,0,0,0,0]$$

$$w_2^b = \frac{1}{4.5}[1,1,1,0,0,1,0,0,0,0],\ w_2^t = [1,1,1,0,0,1,0,0,0,0]$$

$$w_3^b = \frac{1}{5.5}[1,1,1,0,0,1,1,0,0,0],\ w_3^t = [1,1,1,0,0,1,1,0,0,0]$$

$$w_4^b = \frac{1}{6.5}[1,1,1,0,0,1,1,1,0,0],\ w_4^t = [1,1,1,0,0,1,1,1,0,0]$$

$$w_5^b = \frac{1}{7.5}[1,1,1,0,0,1,1,1,1,0],\ w_5^t = [1,1,1,0,0,1,1,1,1,0]$$

$$w_6^b = \frac{1}{8.5}[1,1,1,0,0,1,1,1,1,1],\ w_6^t = [1,1,1,0,0,1,1,1,1,1]$$

ART-1 Model (10)

- Node 1: matching value=3/3.5=0.857, similarity value=3/5=0.6.
- Node 2: matching value=3/4.5=0.666, similarity value=3/5=0.6.
- Node 3: matching value=3/5.5=0.545, similarity value=3/5=0.6.
- Node 4: matching value=3/6.5=0.462, similarity value=3/5=0.6.
- Node 5: matching value=3/7.5=0.4, similarity value=3/5=0.6.
- Node 6: matching value=3/8.5=0.353, similarity value=3/5=0.6.

ART-1 Model (11)

- The same similarity value but different matching value.
- If the number of corresponding bits of output vectors to input vector are the same, the one with less ones in output vector will be selected for vigilance test.

ART-1 Model (12)

- Case 3:
- Assume 3 output nodes.

$$w_1^b = \frac{1}{3.5}[1,1,1,0,0,0,0,0,0,0],\ w_1^t = [1,1,1,0,0,0,0,0,0,0]$$

$$w_2^b = \frac{1}{3.5}[0,1,1,1,0,0,0,0,0,0],\ w_2^t = [0,1,1,1,0,0,0,0,0,0]$$

$$w_3^b = \frac{1}{3.5}[0,0,1,1,1,0,0,0,0,0],\ w_3^t = [0,0,1,1,1,0,0,0,0,0]$$

$$w_4^b = \frac{1}{3.5}[0,0,0,1,1,1,0,0,0,0],\ w_4^t = [0,0,0,1,1,1,0,0,0,0]$$

ART-1 Model (13)

- Node 1: matching value=3/3.5=0.857, similarity value=3/5=0.6.
- Node 2: matching value=2/3.5=0.571, similarity value=2/5=0.4.
- Node 3: matching value=1/3.5=0.286, similarity value=1/5=0.2.
- Node 4: matching value=0/3.5=0.0, similarity value=0/5=0.0.
- Although the number of 1's in the output vector are the same, the matching value and similarity values are all different. But the matching value is proportional to similarity value.

Continuous-Valued ART (ART-2)

- Procedures:
- Given a new training pattern, a MINNET (min net) is adopted to select the winner, which yields the min distance $\| x - w_j \|$.
- Vigilance test: A neuron j* passes the vigilance test if $\| x - w_{j^*} \| < \rho$
- where the vigilance value ρ determines the radius of a cluster.
- If the winner fails the vigilance test, a new neuron unit k is created with $w_k = x$.

Continuous-Valued ART (ART-2) (2)

- If the winner passes the vigilance test, adjust the weight of the winner j* by
- $$w_{j^*}^{new} = \frac{x + w_{j^*}^{(old)} \| cluster_{j^*}^{(old)} \|}{1 + \| cluster_{j^*}^{(old)} \|}$$
- where ||*clusteri*|| denotes the number of members in cluster *i*.

Continuous-Valued ART (ART-2) (3)

- Effect of different order of pattern presentation:
  - The ART is sensitive to the presenting order of the input patterns.
- Effect of vigilance thresholds:
  - The smaller vigilance threshold leads to the more clusters are generated.
- Effect of re-clustering:
  - Use the current centroids as the initial reference for clustering.
  - Re-cluster one by one each of the training patterns.
  - Repeat the entire process until there is no change of clustering during one entire sweep.

order	pattern	winner	test value	decision	cluster 1 centroid	cluster 2 centroid	cluster 3 centroid
1	(1.0,0.1)	-	-	new cluster	(1.0,0.1)		
2	(1.3,0.8)	1	1.0	pass test	(1.15,0.45)		
3	(1.4,1.8)	1	1.6	fail⇒ new cluster		(1.4,1.8)	
4	(1.5,0.5)	1	0.4	pass test	(1.27,0.47)		
5	(0.0,1.4)	2	1.8	fail⇒ new cluster			(0.0,1.4)
6	(0.6,1.2)	3	0.8	pass test			(0.3,1.3)
7	(1.5,1.9)	2	0.2	pass test		(1.45,1.85)	
8	(0.7,0.4)	1	0.63	pass test	(1.13,0.45)		
9	(1.9,1.4)	2	0.9	pass test		(1.6,1.7)	
10	(1.5,1.3)	2	0.5	pass test		(1.58,1.6)	

(a) Original order.

The execution sequence of the ART-2 with the vigilance threshold 1.5.

order	pattern	winner	test value	decision	cluster 1 centroid	cluster 2 centroid	
1	(1.5,1.3)	-	-	new cluster	(1.5,1.3)		
2	(1.9,1.4)	1	0.5	pass test	(1.7,1.35)		
3	(0.7,0.4)	1	1.95	fail⇒ new cluster		(0.7,0.4)	
4	(1.5,1.9)	1	0.75	pass test	(1.63,1.53)		
5	(0.6,1.2)	2	0.9	pass test		(0.65,0.8)	
6	(0.0,1.4)	2	1.25	pass test		(0.43,1.0)	
7	(1.5,0.5)	1	1.17	pass test	(1.6,1.28)		
8	(1.4,1.8)	1	0.72	pass test	(1.56,1.38)		
9	(1.3,0.8)	1	0.84	pass test	(1.52,1.28)		
10	(1.0,0.1)	2	1.47	pass test		(0.58,0.78)	

(b) Reverse order