Knowledge Discovery and Data Mining

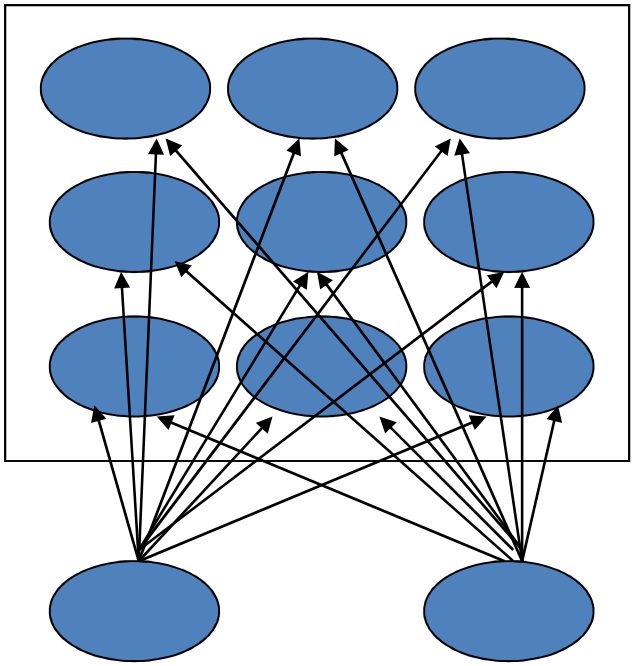
Unit # 9

Clustering Techniques

- K-Means ✔
- Agglomerative ✔
- Kohonen Self-Organizing Maps
- Adaptive Resonance Theory

Kohonen Self-Organizing Maps

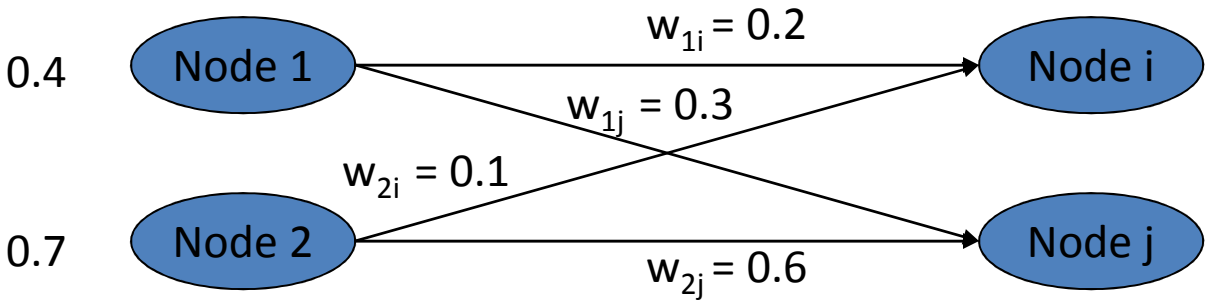
- Formalized by Teuvo Kohonen in 1982 for unsupervised clustering.
- Supports two layers:
  - The input layer contains one node for each input attribute.
  - Nodes of the input layer have a weighted connection to all nodes in the output layer.
  - The output layer is commonly organized as a two dimensional grid.



Working of Kohonen Maps

- During the network learning, input instances are presented to each output layer node. A simple feed-forward algorithm computes the activation of each output node.
- A winner-takes-all approach then decides which particular output activation is chosen as the correct classification. But instead of choosing the largest activation, we choose the lowest activation, as it represents the class with the closest Euclidean distance to the input prototype vector.
- The node is rewarded by having its weights changed to more closely match the instance.
- The output nodes winning the most instances during the final pass of the data through the network are saved.
- The number of output layer nodes saved corresponds to the number of clusters believed to be in the data.

Working of Kohonen Maps (Cont'd)



- The score for classifying a new instance with output node j is given by

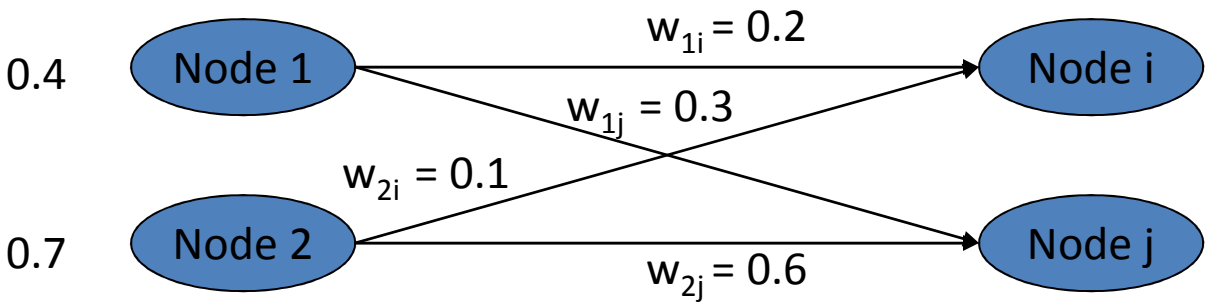
$$\text{sqrt}\left(\sum (n_i - w_{ij})^2\right)$$

  - n_i is the attribute value for the current instance at input i.
  - w_{ij} is the weight associated with the ith input node and output node j.
- Weights are updated according the following formula:

$$w_{ij}\ (\text{new}) = w_{ij}\ (\text{current}) + \Delta w_{ij}$$

  - where $\Delta w_{ij} = r(n_i - w_{ij})$, r is the learning parameter and $0<r<1$.

Working of Kohonen Maps (Cont'd)



- Score of Node i: sqrt(0.4-0.2)^2 + (0.7-0.3)^2 = 0.45
- Score of Node j: sqrt(0.4-0.3)^2 + (0.7-0.6)^2 = 0.14

- Thus, the record belongs to Cluster j.

- Next we update the weights of incoming links to node j.
- $\Delta w_{1j} = 0.8 \times (0.4 - 0.3) = 0.08$
- $\Delta w_{2j} = 0.8 \times (0.7 - 0.6) = 0.08$
- $w_{1j} = 0.3 + 0.08 = 0.38$
- $w_{2j} = 0.6 + 0.08 = 0.38$

Summary

- The simplicity of this algorithm makes it a great choice for clustering.
- One primary disadvantage of the algorithm is that the number of output classes must be defined upfront.
- This is significant because it assumes that we have some general knowledge of the data and how it should be classified.

Adaptive Resonance Theory

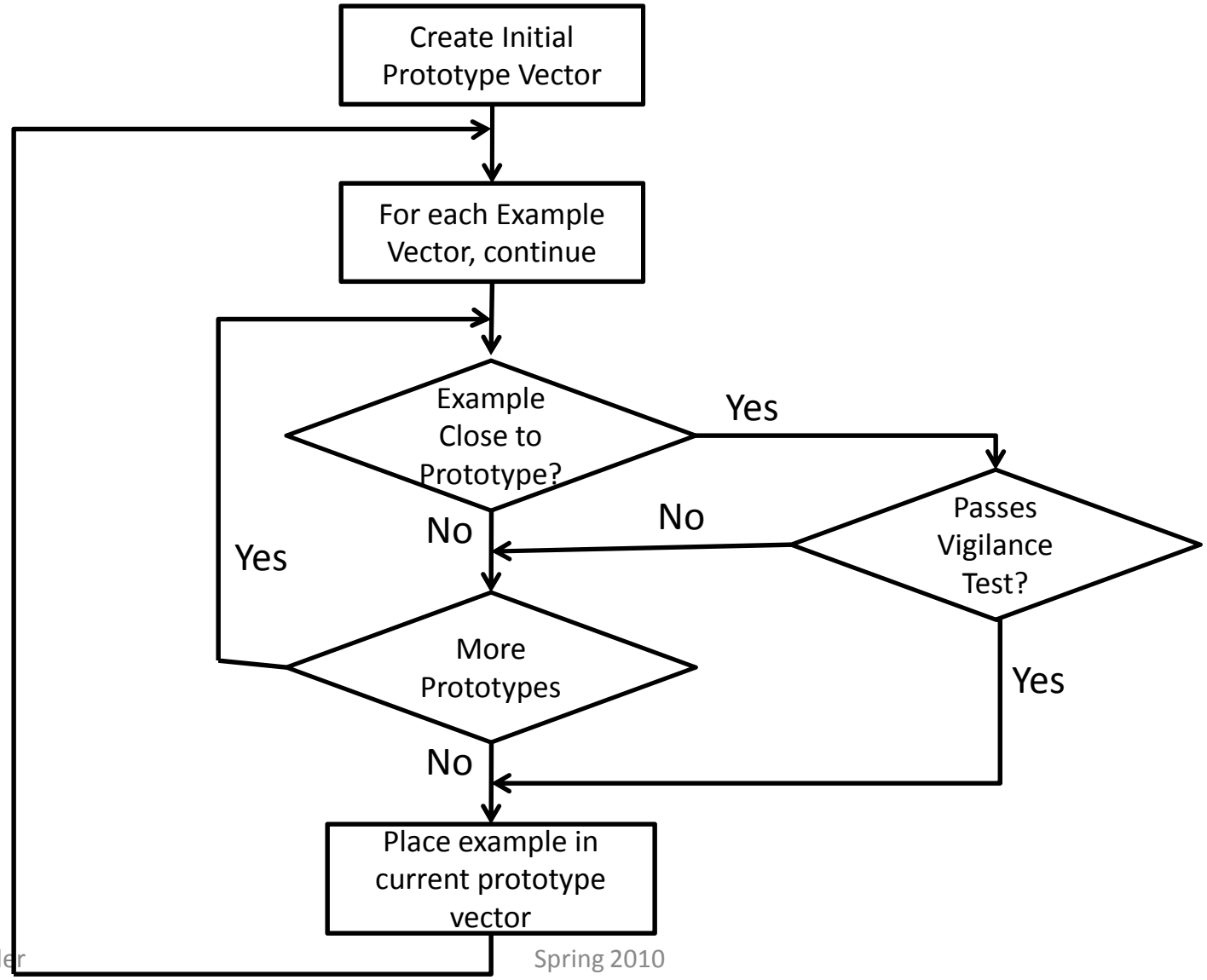
- The ART1 (adaptive resonance theory) algorithm is a simple, unsupervised learning algorithm with biological motivation.
- It works with objects called feature vectors.
- A feature vector is nothing more than a collection of binary values that represent some type of information.
- An example of a feature vector is a customer's purchase data

Hammer	Paper	Pen	Kit-Kat	Pencil	Binder	Snickers
1	0	0	1	0	0	1

Working of ART1

- We being with a set of feature vectors and a set of initialized prototype vectors (P1….PN).
- The prototype vector is the center of the cluster.
- The number of prototype vector, N, is the maximum number of clusters that can be supported.
- The d parameter represents the length of the vector.
- We initialize a vigilance parameter (ρ) to a small value between 0 and 1 and a beta parameter to a small positive integer.
- The parameter (d) represents the dimension of the vectors.

ART1 Algorithm Flow



ART1 Conditions

- Initially, no prototype vectors exist, so at the start fo the algorithm an initial prototype vector is created with the first example vector.
- We then check all subsequent example feature vectors against each existing prototype vector for its proximity.
- Proximity Test
  - $||P_i \cap E||/(\beta+||Pi||) > ||E||/(\beta+d)$
- Vigilance Test
  - $||P_i \cap E||/(||E||) > \rho$

Example

- P0: {1, 0, 0, 1, 1, 0, 1}
- P1: {1, 1, 0, 0, 0, 1, 0}
- E: {1, 1, 1, 0, 0, 1, 0}
- $\beta = 1.0$, $\rho = 0.6$, $d = 7$
- Proximity Test with P0 ?
- Proximity Test with P1 ?
- Vigilance Test with P1 ?
- P1 AND E {1,1,0,0,0,1,0} AND {1,1,1,0,0,1,0} = {1,1,0,0,0,1,0}

Termination Condition

- As example feature vectors are tested against the prototype vectors, new clusters are created or existing clusters are modified at the inclusion of an example.
- This process, known as “resonance”, indicates the process of learning within the algorithm.
- When the algorithm reaches equilibrium (that is, no further changes occur with the prototype vectors), learning is complete and the data set is classified.

Summary

- ART1 is both conceptually simple and easy to implement.
- Earlier algorithms, such a k-means clustering algorithm, though much simpler, have some significant drawbacks.
- For example, k-means does not allow the creation of new clusters (the clusters are statically defined at the start).
- Also, no parameter exists within k-means to adjust the class size of the result clusters.
- A drawback to both algorithms (ART1 and k-means) is that the final set of clusters (and prototype vectors) can be influenced based on the order in which training is performed.