Unsupervised Learning

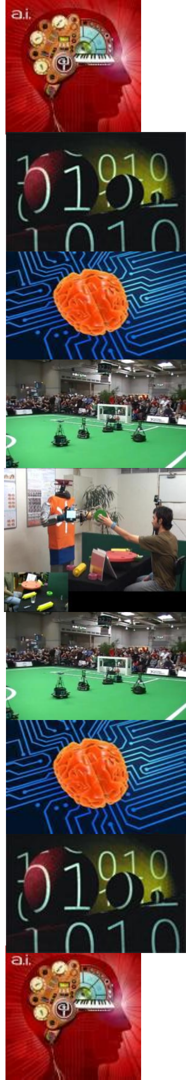
Introduction

- Dataset does not contain labels

- Clustering vs classification

- Different clustering algorithms

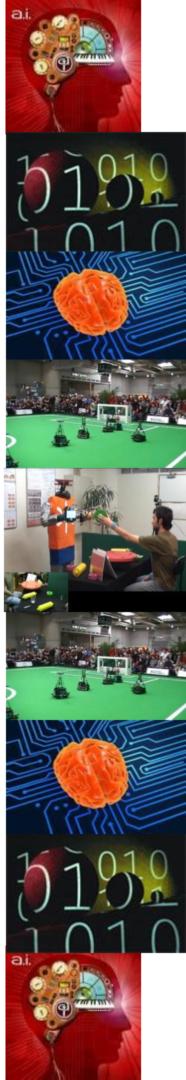
- K-Means – numerical data
- Determining K
  - Random, Elbow method



K-Means Algorithm

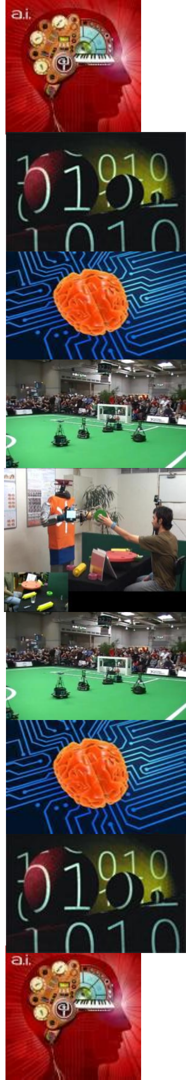
Algorithm 1 K-Means Algorithm

Given a set of data instances $D = d_1, \ldots, d_n$
Determine K (the number of clusters)
Randomly select centroids c_1 to c_k for each of the K clusters
while the algorithm has not converged **do**
    for $i \leftarrow 1$ *to* n **do**
        for $j \leftarrow 1$ *to* K **do**
            Calculate the Euclidean distance e_j of d_i from the centroid of cluster k_j
        end for
        Add d_i to the cluster k_j with the smallest e_j value
    end for
    for $j \leftarrow 1$ *to* K **do**
        for $l \leftarrow 1$ *to* m **do**
            Calculate the average a_l of the lth dimension of the data instances in the cluster j
        end for
        Update the jth centroid to the averaged values a_j for each dimension of the data instance
    end for
end while



Example Data Instances

Entity	Attr1	Attr2
1	1	1
2	1.5	2
3	3	4
4	5	7
5	3.5	5



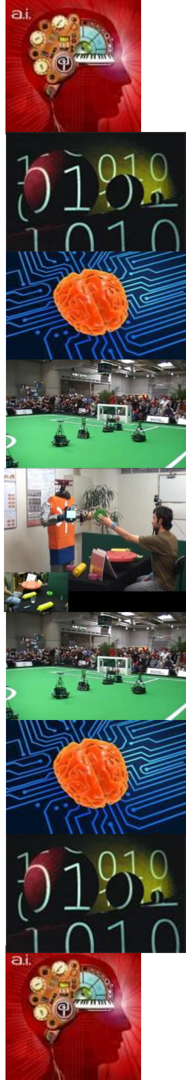
Example Data Instances

Entity 1 (c1) = $\text{sqrt}((1.5-1)^2+(2-1)^2) = 1.12$

Entity 1 (c2) = $\text{sqrt}((5-1)^2+(7-1)^2) = 7.21$

Entity 3 (c1) = $\text{sqrt}((1.5-3)^2+(2-4)^2) = 2.50$

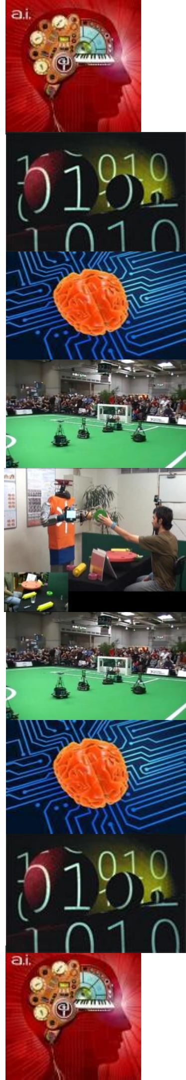
Entity 3 (c2) = $\text{sqrt}((5-3)^2+(7-4)^2) = 3.61$



Example Data Instances

Entity 5 (c1) = $\mathrm{sqrt}((1.5-3.5)^2+(2-5)^2) = 3.61$

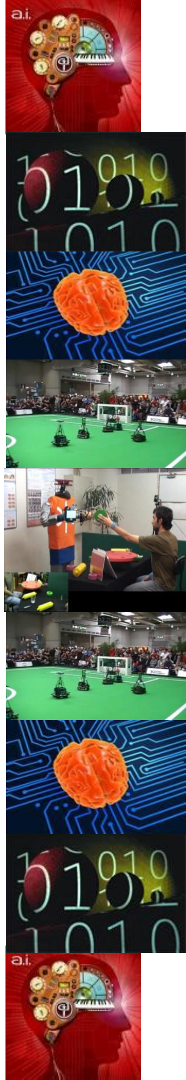
Entity 5 (c2) = $\mathrm{sqrt}((5-3.5)^2+(7-5)^2) = 2.5$



Euclidean Distance – Iteration 1

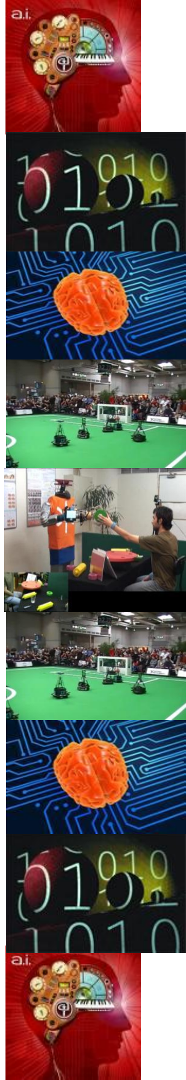
Table 2: Euclidean distance-Iteration 1

Entity	Attr1	Attr2	Cluster 1	Cluster 2
1	1	1	1.12	7.21
3	3	4	2.5	3.61
5	3.5	5	3.61	2.5



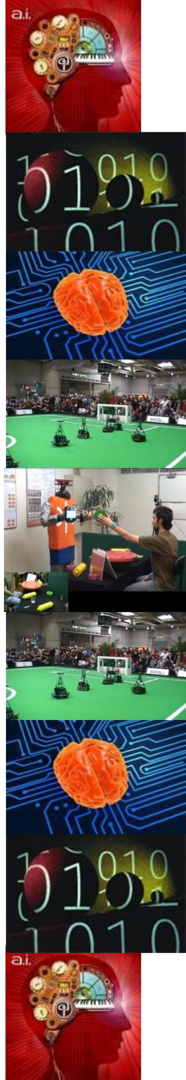
Clusters 1 Updated Centroids

	Attr 1	Attr 2
Entity 1	1	1
Entity 2	1.5	2
Entity 3	3	4
Updated Centroid	**1.83**	**2.33**



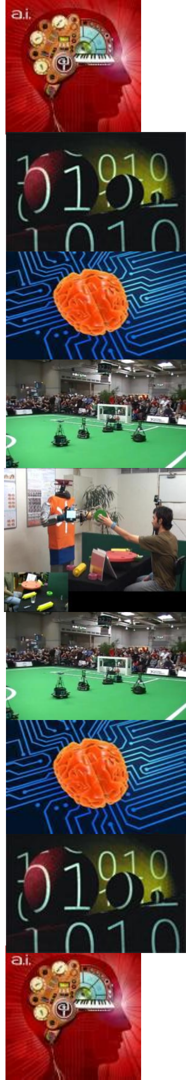
Clusters 2 Updated Centroids

	Attr 1	**Attr 2**
Entity 4	5	7
Entity 5	3.5	5
Updated Centroid	**4.25**	**6**



Euclidean Distance – Iteration 2

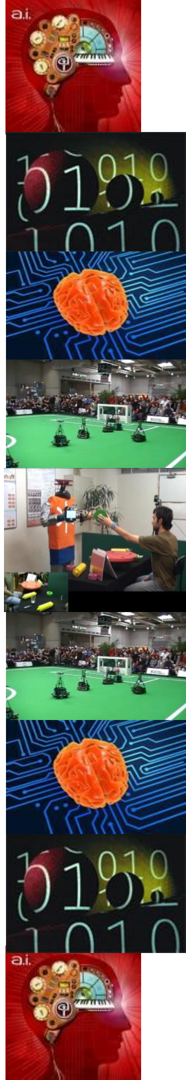
Entity	Attr 1	Attr 2	Cluster (1.83, 2.33)	Cluster2 (4.25, 6)
Entity 1	1	1	**1.57**	5.96
Entity 2	1.5	2	**0.46**	4.85
Entity 3	3	4	**2.04**	2.35
Entity 4	5	7	5.64	**1.25**
Entity 5	3.5	5	3.15	**1.25**



k-mediods

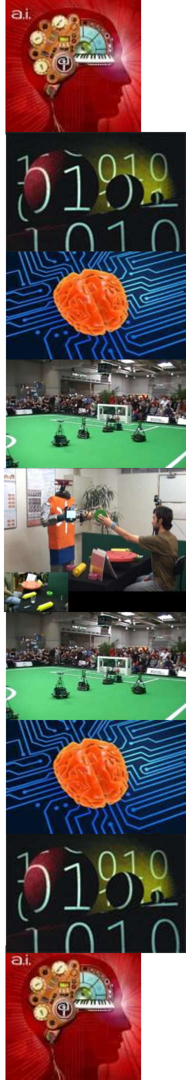
- **Clustering algorithm** similar to k-means but with some key differences.

1. Initialization: Randomly select initial medoids.
2. Iteration: Assign data points to closest medoids based on dissimilarity.
3. Medoid Reassignment: Check if swapping a data point with the current medoid improves the cluster's total distance. If so, swap them.
4. Stopping Criterion: Stop iterating if no medoid swaps occur (indicating stability).
5. Output: Return the final cluster assignments and medoids.



Unsupervised Applications

- k-Means uses the mean (average) of points within a cluster as the centroid.
- k-Medoids uses an actual data point from the cluster as the medoid, making it potentially more robust to outliers.
- k-Means requires data points to have numerical values for distance calculations.
- k-Medoids can work with other dissimilarity measures, making it more flexible for some data types.

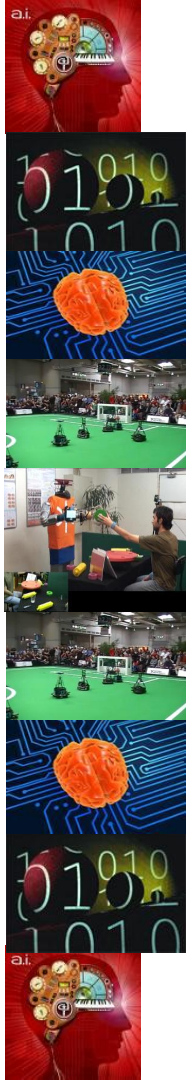


Unsupervised Applications

- Clustering

- Association - rules to associate variables e.g market basket analysis.

- Dimensionality reduction.



Unsupervised Applications

- Clustering
- Association - rules to associate variables e.g market basket analysis.
- Dimensionality reduction.