CLUSTERING I

- In DM, **classification** deals with partitioning data based on a pre-classified training sample
- Often it is useful to partition data without having a training sample **unsupervised learning**
- Example 1: determine groups of customers who have similar buying patterns
- Example 2: determine groups of patients who exhibit similar reactions to prescribed drugs

CLUSTERING II

- **Goal of clustering:**
  place data (records) into groups, such that records on each group are similar to each other and dissimilar from records in other groups.
- The groups are disjoint.
- "similar" is defined via a similarity function
- For numerical data we can use the **Euclidean distance**

  $D([a_1,\ldots,a_n],[b_1,\ldots,b_n]) = \Sigma\ |a_i\text{-}b_i|^2$
- Small distance → greater similarity

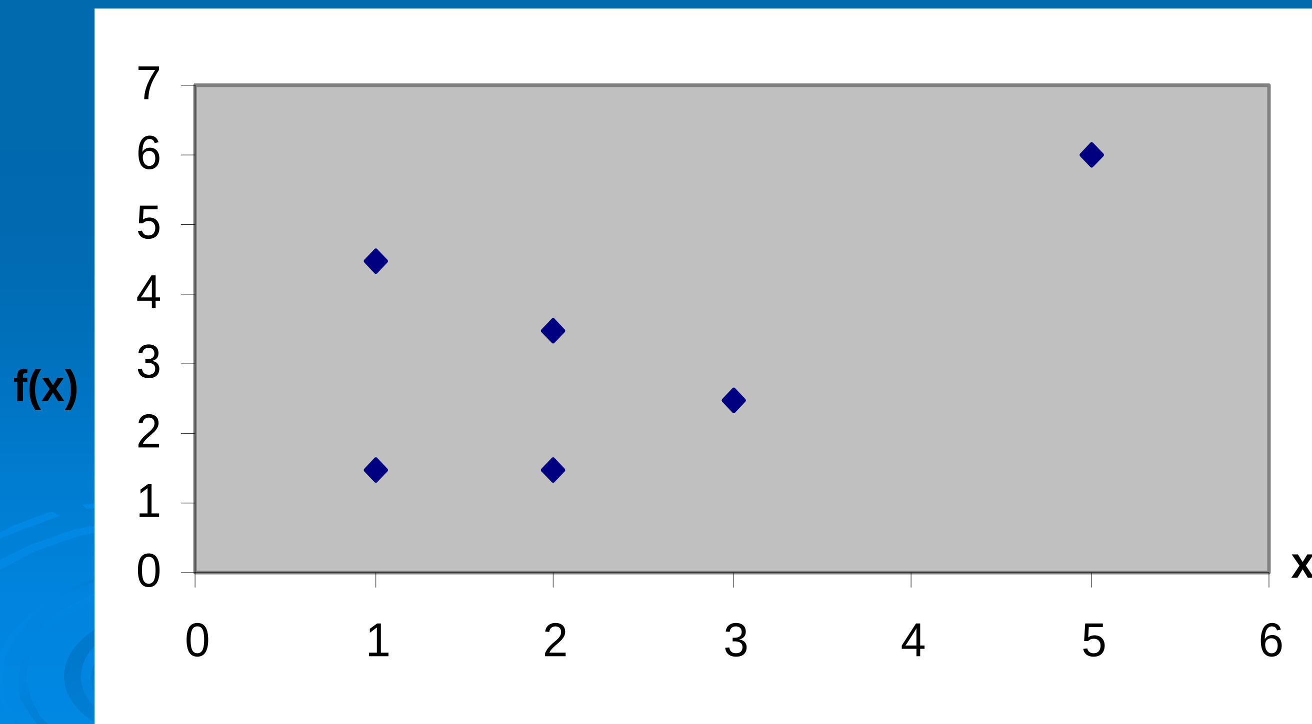
The K-Means Algorithm

1. Choose a value for *K*, the total number of clusters.
2. Randomly choose *K* points as cluster centers.
3. Assign the remaining points to their closest cluster center.
4. Calculate a new cluster center for each cluster.
5. Repeat steps 3-4 until the cluster centers stabilize

An Example Using K-Means

Table 3.6 • **K-Means Input Values**

Instance	X	Y
1	1.0	1.5
2	1.0	4.5
3	2.0	1.5
4	2.0	3.5
5	3.0	2.5
6	5.0	6.0



- Choose K = 2 & C1=(1.0,1.5), C2=(2.0,1.5)
- Compute D(C1,i), D(C2,i) for i=1,2,3,4,5,6
- Create the clusters {1,2} and {3,4,5,6}
- Recompute each cluster center:

  x = (x1+x2)/2 y = (y1+y2)/2

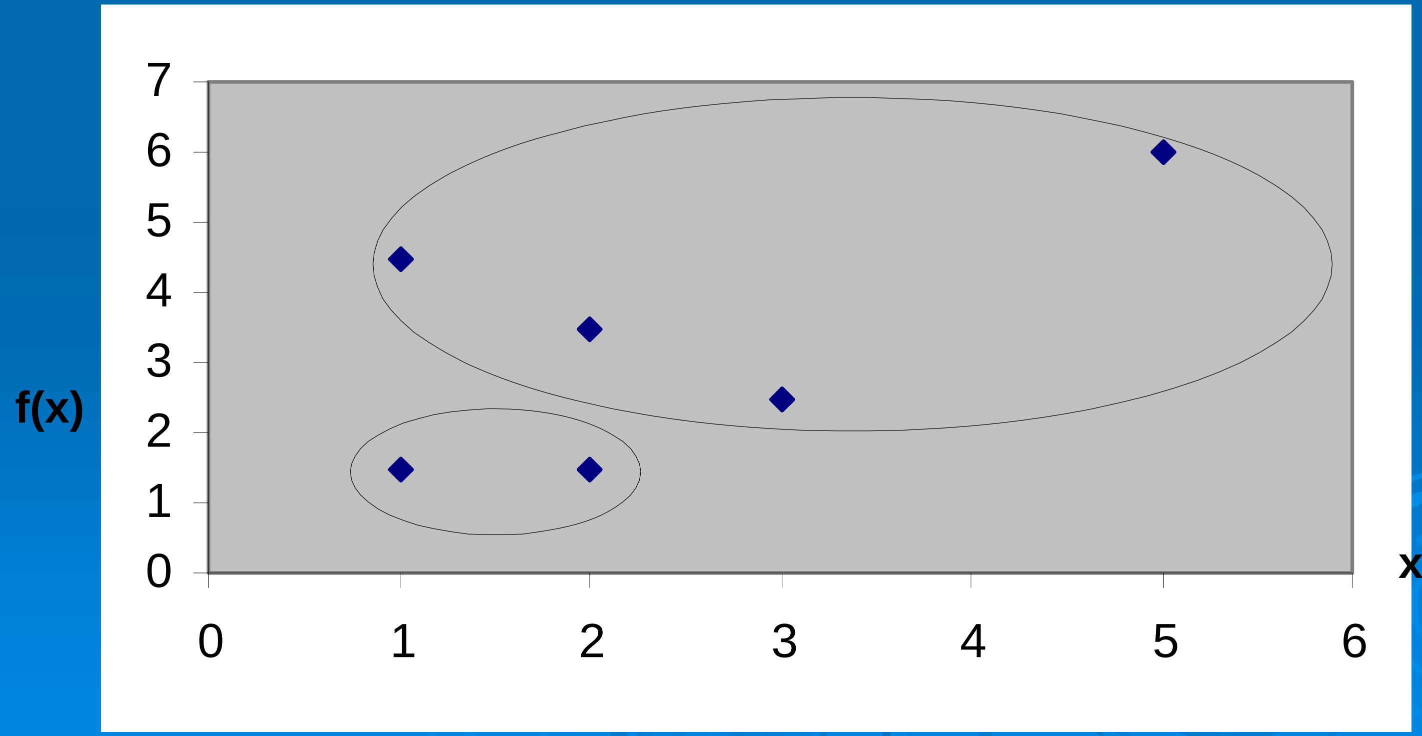
  x = (x3+x4+x5+x6)/4 y = (y3+y4+y5+y6)/4

  new cluster centers: C1 = (1,3), C2 = (3,3.375)
- Compute D(C1,i), D(C2,i) for i=1,2,3,4,5,6
- Create the clusters {1,2,3} and {4,5,6}
- Recompute each cluster center:

  new cluster centers: (1.33,2.5), (3.33,4)

Table 3.7 • **Several Applications of the K-Means Algorithm (K = 2)**

Outcome	Cluster Centers	Cluster Points	Squared Error
1	(2.67,4.67)	2, 4, 6	14.50
	(2.00,1.83)	1, 3, 5	
2	(1.5,1.5)	1, 3	15.94
	(2.75,4.125)	2, 4, 5, 6	
3	(1.8,2.7)	1, 2, 3, 4, 5	9.60
	(5,6)	6	



Subtleties of K-Means I

- We may see a different final cluster configuration for each alternative choice of the initial cluster centers.
- The algorithm is guaranteed to produce a stable clustering, but not necessarily an optimal one
- An **optimal clustering** for K-Means is defined as a clustering for which the summation of the squared error differences between the data points and their corresponding cluster center is minimum.
- Often we use the squared error as a **termination criterion**, instead of running K-Means several times

Subtleties of K-Means II

- It only works with real-valued data. Categorical attributes, must be converted to numerical values (RGB)
- We must select the number of clusters in advance. Run the algorithm with several different values of K.
- Works best when the clusters in the data are of approximately equal size.
- There is no way to tell which attributes are significant in determining the cluster formation.
- The lack of an intuitive explanation about the nature of the clusters formed doesn’t help us interpret the findings