Sınıflandırma (Classification)

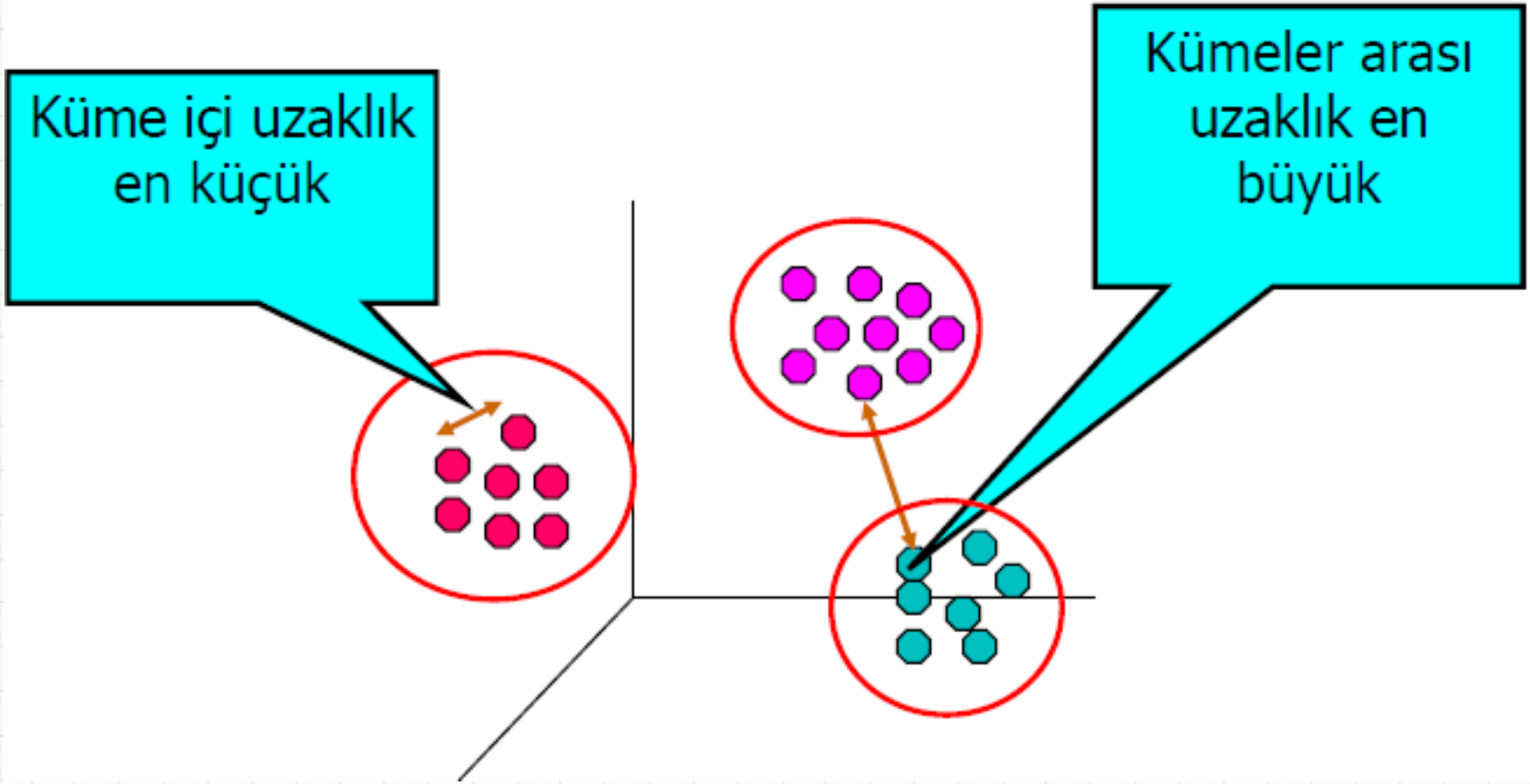
- Eğiticili (supervised) sınıflandırma:
  Sınıflandırma: Sınıf sayısı ve bir grup örneğin hangi sınıfa ait olduğunu bilinir

- Eğiticisiz (unsupervised) sınıflandırma:
  Kümeleme: Hangi nesnenin hangi sınıfa ait olduğu ve grup sayısı belirsizdir.

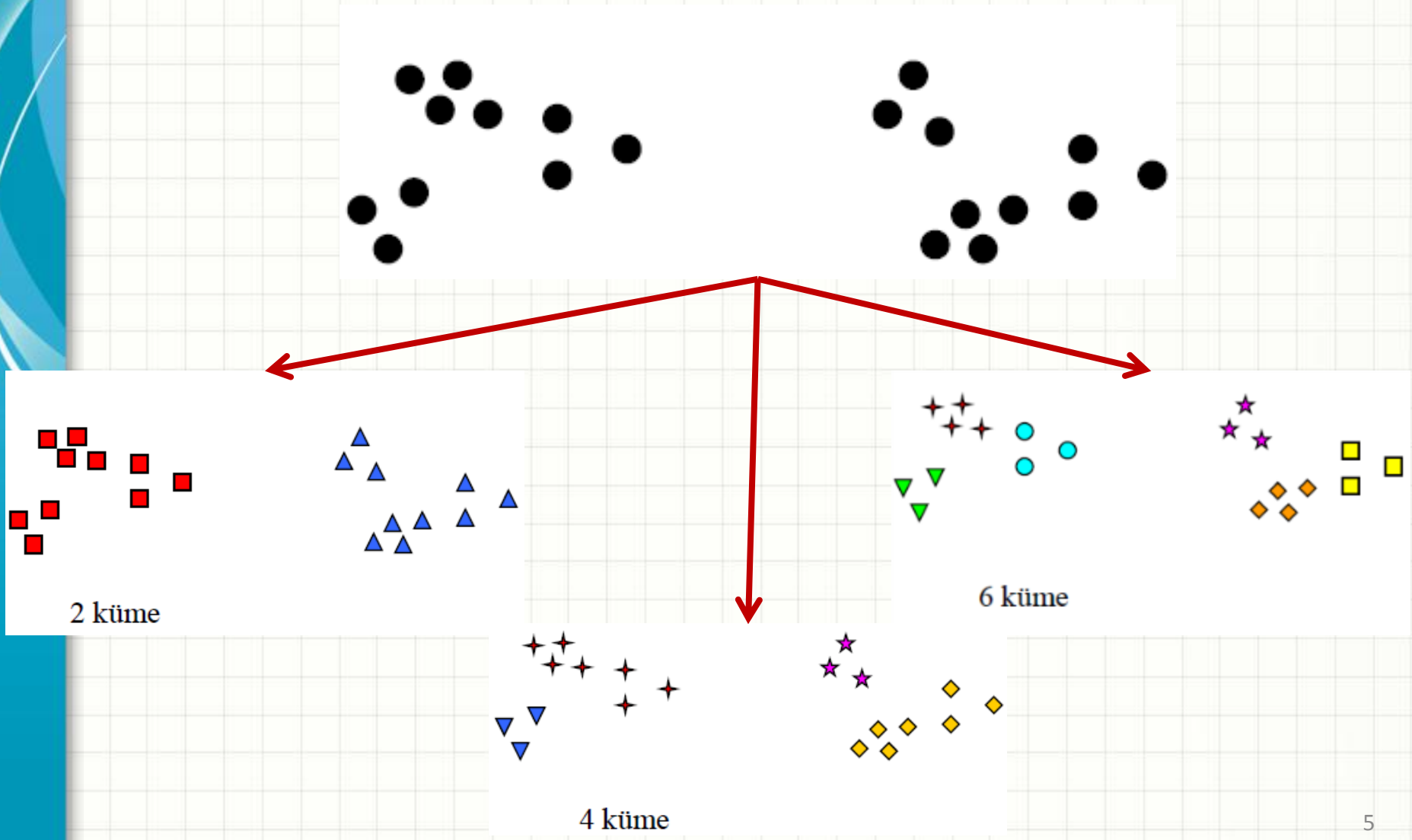
Kümeleme (Clustering)

- Kümeleme, eğiticisiz öğrenme ile gerçekleştirilir.
- Küme: Birbirine benzeyen nesnelerden oluşan gruptur.
  - Aynı kümedeki örnekler birbirine daha çok benzer
  - Farklı kümedeki örnekler birbirine daha az benzer

Benzerlik İlişkisi: Örnek



Örnekte kaç küme vardır?



Benzerlik Ölçüsü: Nümerik

- Veri kümesi içindeki nümerik örneklerin birbirine olan benzerliğini ölçmek için mesafe ölçüsü kullanılabilir.
- Ancak mesafe ölçüsü benzerlikle ters orantılıdır.
  - L1 Norm (City Block / Manhattan Distance)
  - L2 Norm (Euclidean Distance)
  - L3 Norm (*Minkowski distance)*

Mesafe Ölçüsü: L1 Norm

- L1 Norm: *City Block / Manhattan Distance*
- p boyutlu uzayda verilen i ve j noktalarının birbirine olan uzaklığı:

$$d(i,j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + \ldots + |x_{i_p} - x_{j_p}|$$

Mesafe Ölçüsü: L2 Norm

- L2 Norm: *Euclidean Distance*
- p boyutlu uzayda verilen i ve j noktalarının birbirine olan uzaklığı:

$$d(i,j)=\sqrt{(|x_{i_1}-x_{j_1}|^2+|x_{i_2}-x_{j_2}|^2+...+|x_{i_p}-x_{j_p}|^2)}$$

Mesafe Ölçüsü: L3 Norm

- L3 Norm: *Minkowski distance*
- p boyutlu uzayda verilen i ve j noktalarının birbirine olan uzaklığı:

$$d(i,j)=\sqrt[q]{(|x_{i_1}-x_{j_1}|^q+|x_{i_2}-x_{j_2}|^q+\ldots+|x_{i_p}-x_{j_p}|^q)}$$

NOT: q=2 için Euclidean uzaklığını verir

Mesafe Ölçüsü

- Mesafe ölçüsü ile ilgili özellikler:
  - $d(i,j) \geq 0$
  - $d(i,i) = 0$
  - $d(i,j) = d(j,i)$
  - $d(i,j) \leq d(i,k) + d(k,j)$

Benzerlik Ölçüsü: Binary

i ve j örneklerine ait binary (ikili) özellikler bir olasılık tablosu (contingency table) ile gösterilir:

		j Örneği	
		0	**1**
i Örneği	**0**	a	b
	1	c	d

a: i örneğinde 0, j örneğinde 0 olan özelliklerin sayısı

b: i örneğinde 0, j örneğinde 1 olan özelliklerin sayısı

c: i örneğinde 1, j örneğinde 0 olan özelliklerin sayısı

d: i örneğinde 1, j örneğinde 1 olan özelliklerin sayısı

Simple Matching Coefficient (SMA): İkili değişkenin simetrik olduğu durumlarda

$$sim(i,j) = \frac{a+d}{a+b+c+d}$$

Jaccard coefficient: İkili değişkenin asimetrik olduğu durumlarda

$$sim_{Jaccard}(i,j) = \frac{d}{b+c+d}$$

Benzerlik Ölçüsü: Binary

- i=10011011 ve j=11000110
- i ve j örneklerinin birbirlerine olan benzerlikleri;
- a=1, b=2, c=3, d=2 olduğuna göre
  - Sim $_{SMC}$(i,j)= 3/8
  - Sim $_{jaccard}$(i,j)= 2/7 olur.

Kümeleme Yöntemleri

- K-Means Kümeleme
- Hiyerarşik Kümeleme
- Yapay Sinir Ağları (SOM-Self Organized Feature Map)
- Genetik Algoritmalar

K-Means Kümeleme

- K-means algoritması basit ve etkin bir istatistiki kümeleme yöntemidir.
- K-means algoritması veri kümesini birbirinden ayrık kümelere böler.
- K küme sayısının başlangıçta bilinmesi gerekir.

K-Means Kümeleme

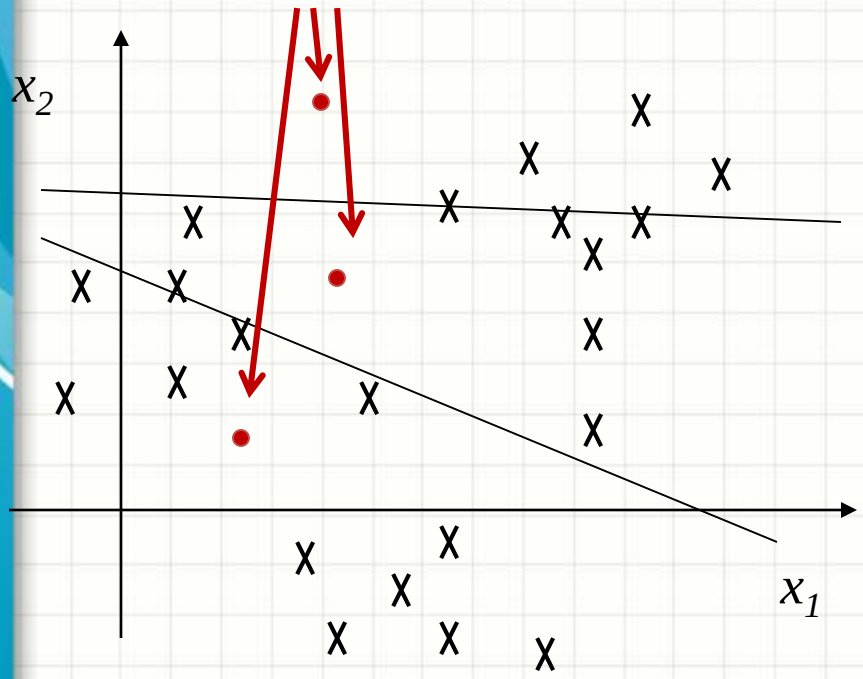
Algoritmanın adımları;

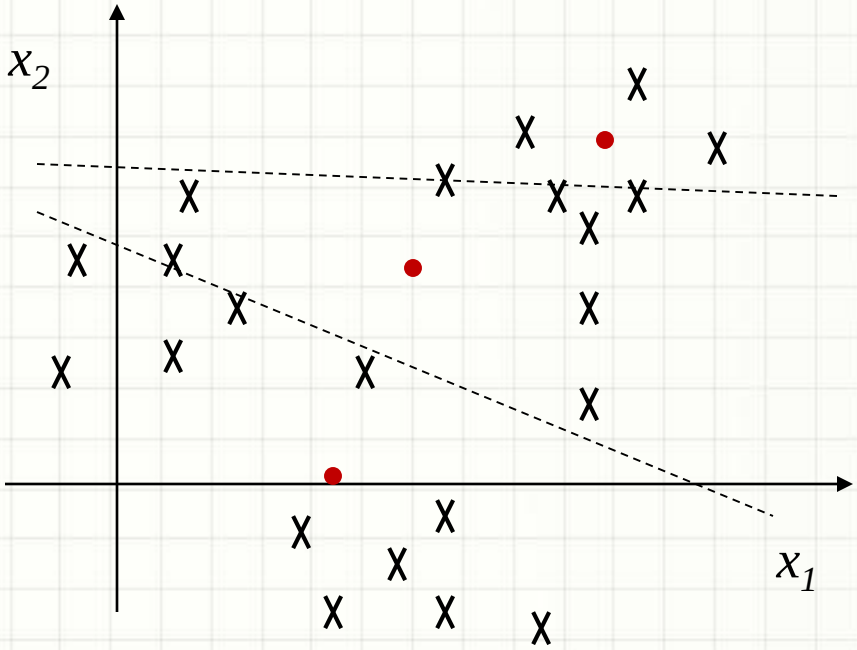
1. Belirlenecek küme sayısı (k) seçilir.
2. k adet rastgele başlangıç küme merkezi belirlenir. (Veri kümesindeki örneklerden de seçilebilir)
3. Öklid mesafesi kullanılarak kalan örneklerin en yakın olduğu küme merkezleri belirlenir.
4. Her küme için yeni örneklerle küme merkezleri hesaplanır.
5. Eğer kümelerin yeni merkez noktaları bir önceki merkez noktaları ile aynı ise işlem bitirilir.

   Değilse yeni küme merkezleri ile **3. adımdan itibaren** işlemler tekrarlanır.

Iteration *1*

Rastgele belirlenen başlangıç merkezleri: her merkez için en yakın noktaları belirle

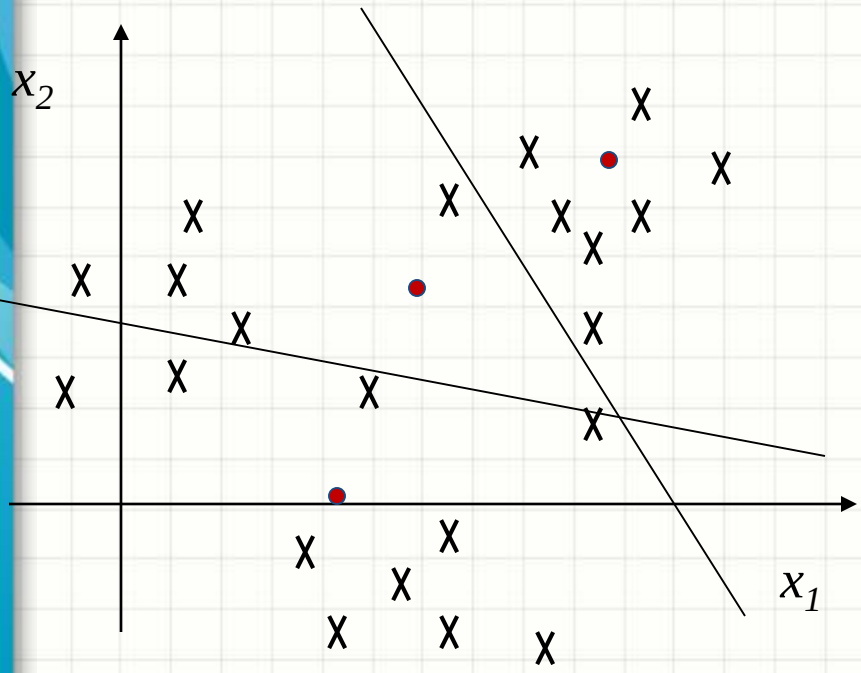


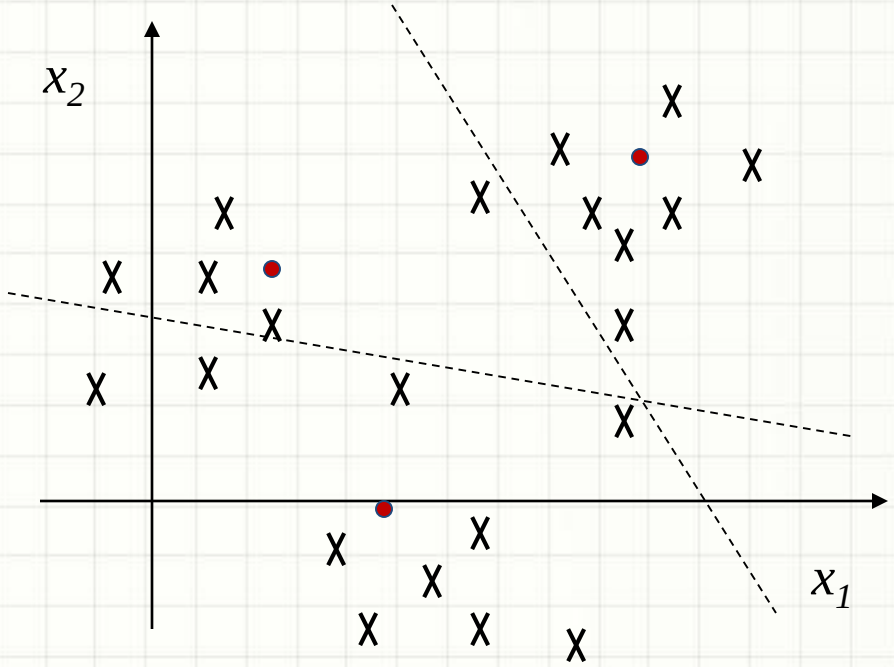


Küme merkezleri yeniden hesaplanır

Iteration *2*

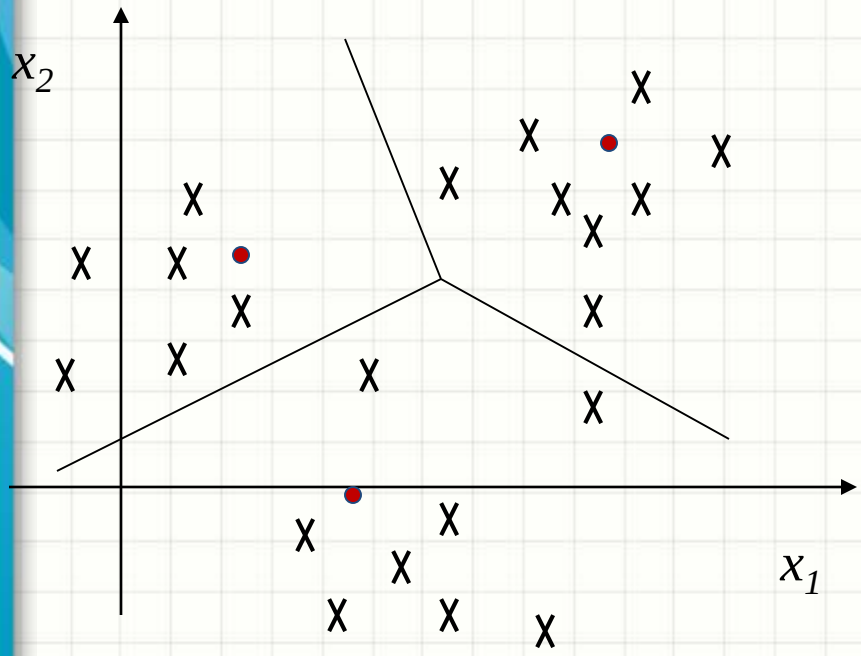
Yeni küme merkezlerine en yakın noktaları belirle

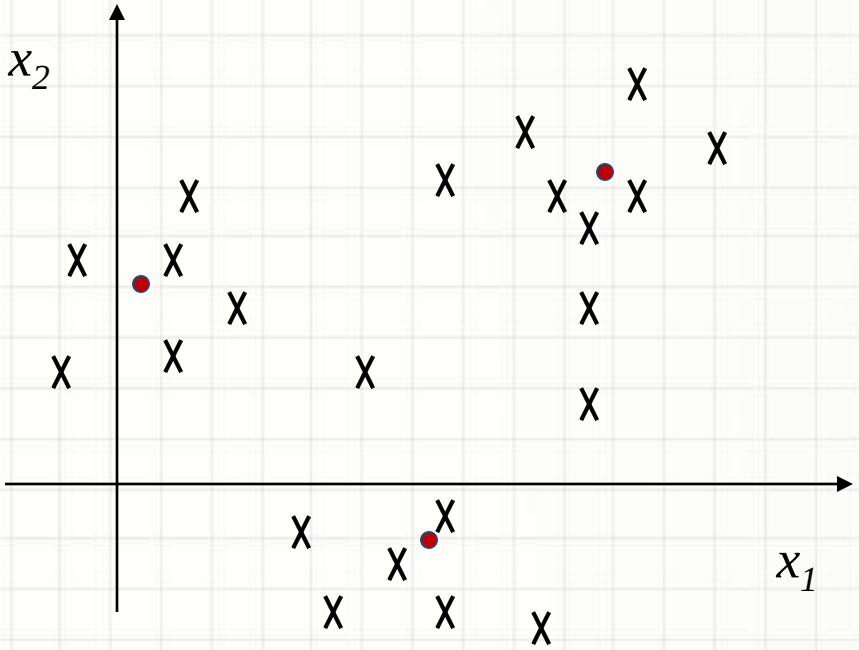




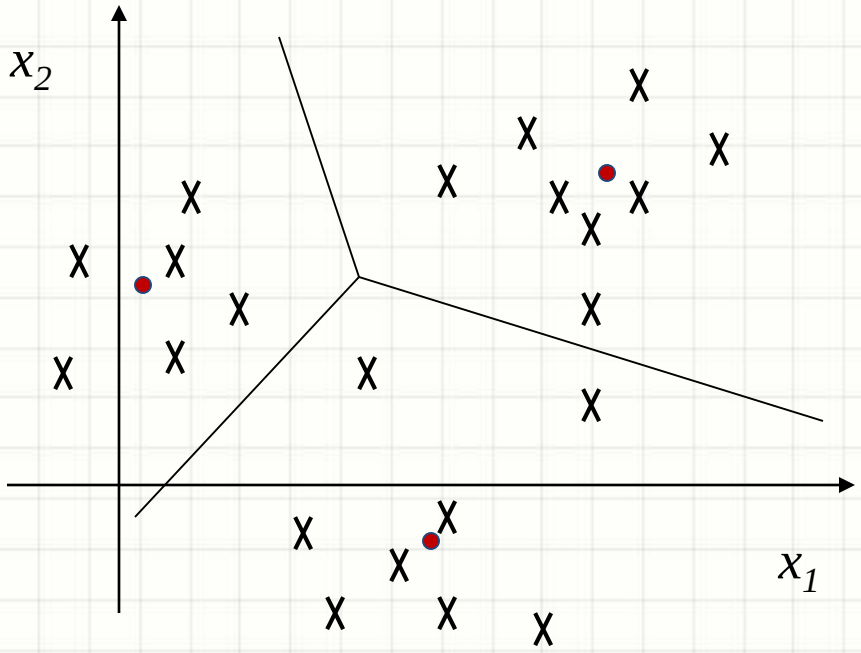
Küme merkezlerini yeniden hesapla

Iteration *3*





Iteration *4*



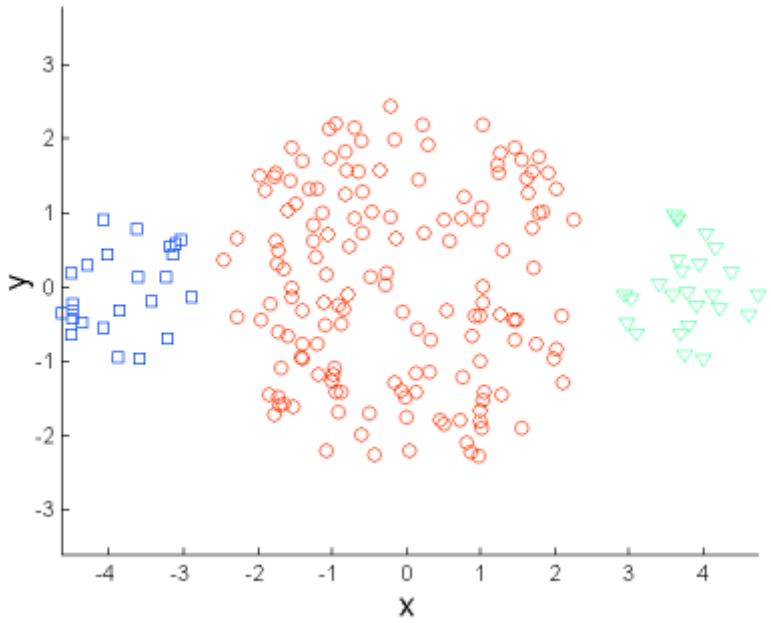
Merkezlerin yerleri değişmedi

=> DUR

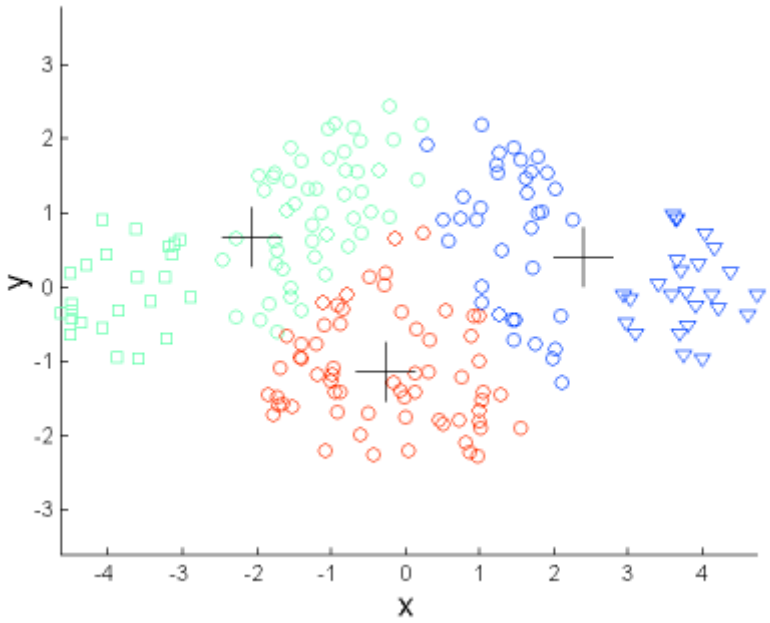
K-means: sorunları

- Kümeler, farklı
  - Büyüklük,
  - Yoğunluk ve
  - Dairesel olmayan şekillerde olduğunda
- Veride aykırı örnekler (outlier) ya da gürültü (noise) bulunduğunda

Farklı büyüklükteki kümeler

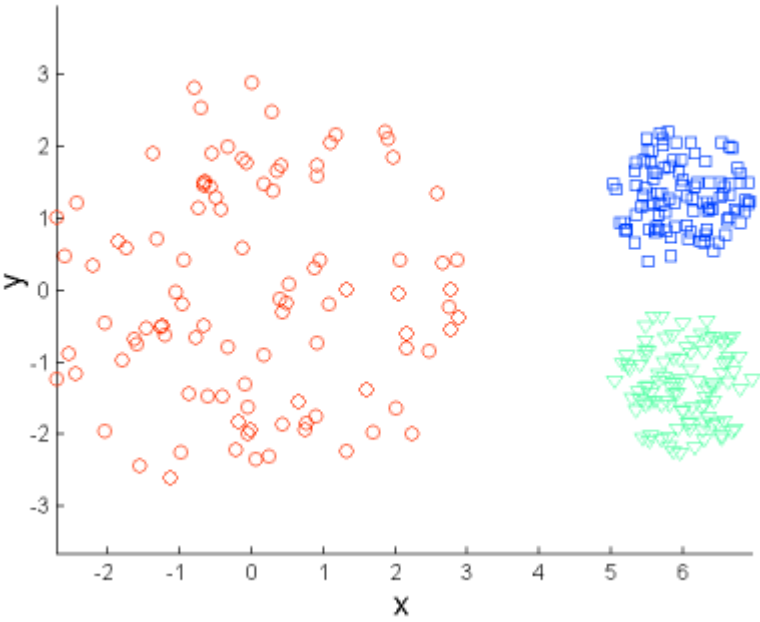


Orijinal kümeler

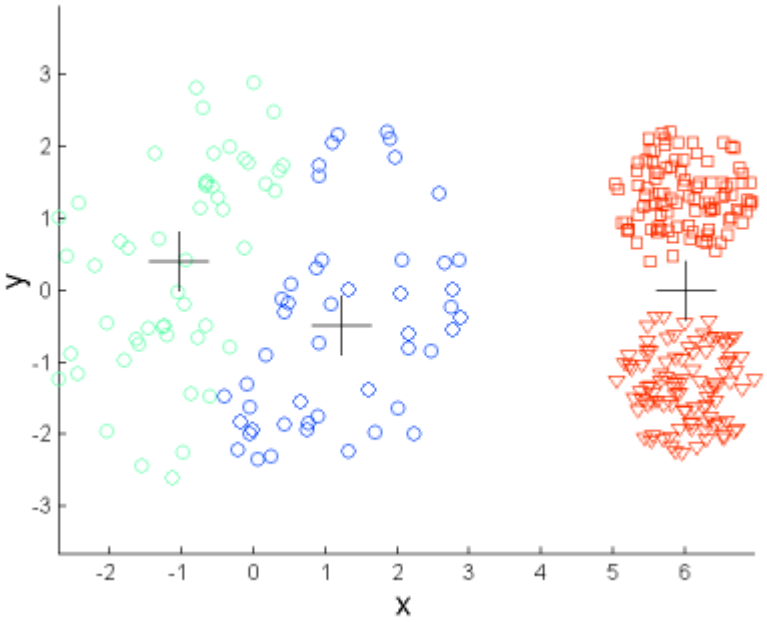


K-means (3 küme)

Farklı yoğunluktaki kümeler

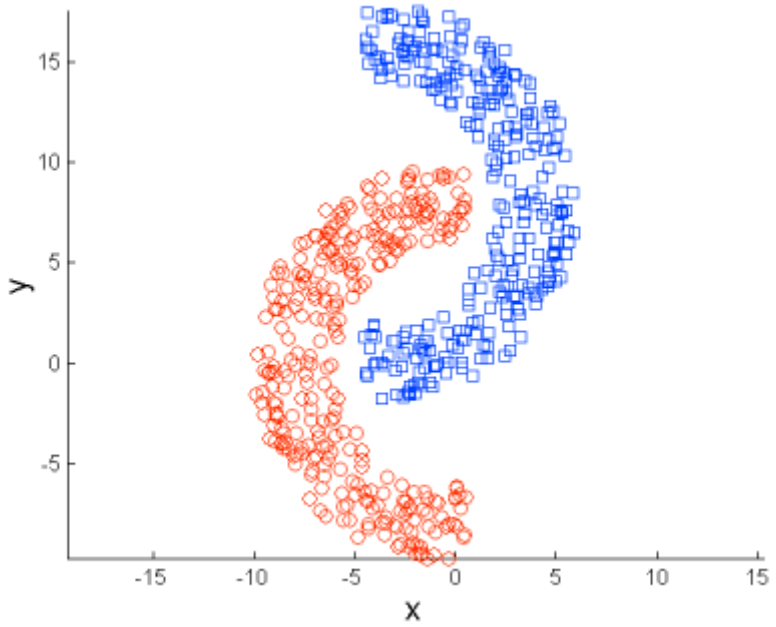


Orijinal kümeler

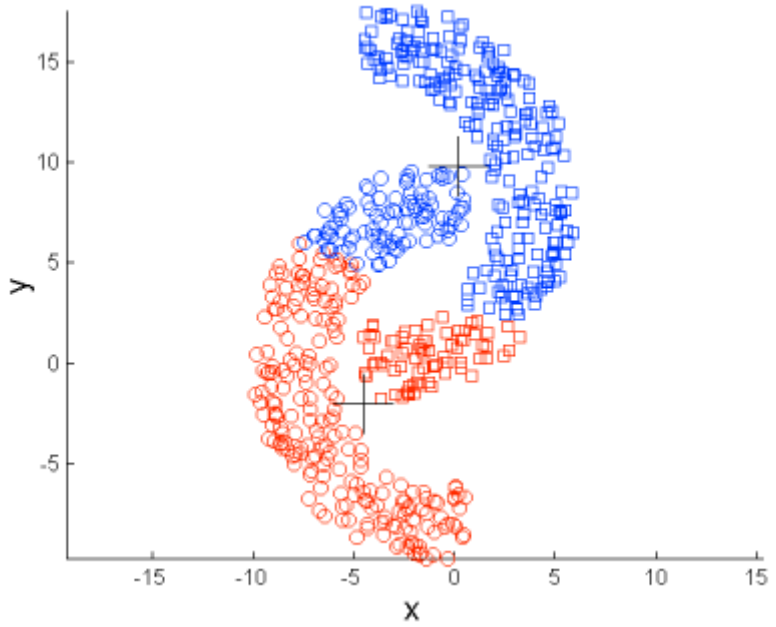


K-means (3 küme)

Dairesel olmayan kümeler

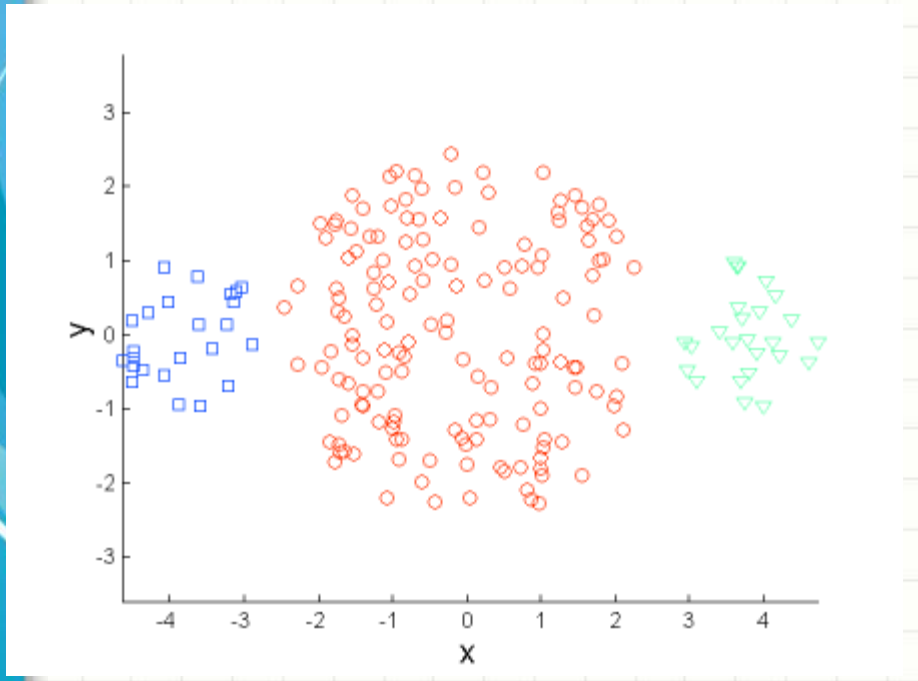


Orijinal kümeler

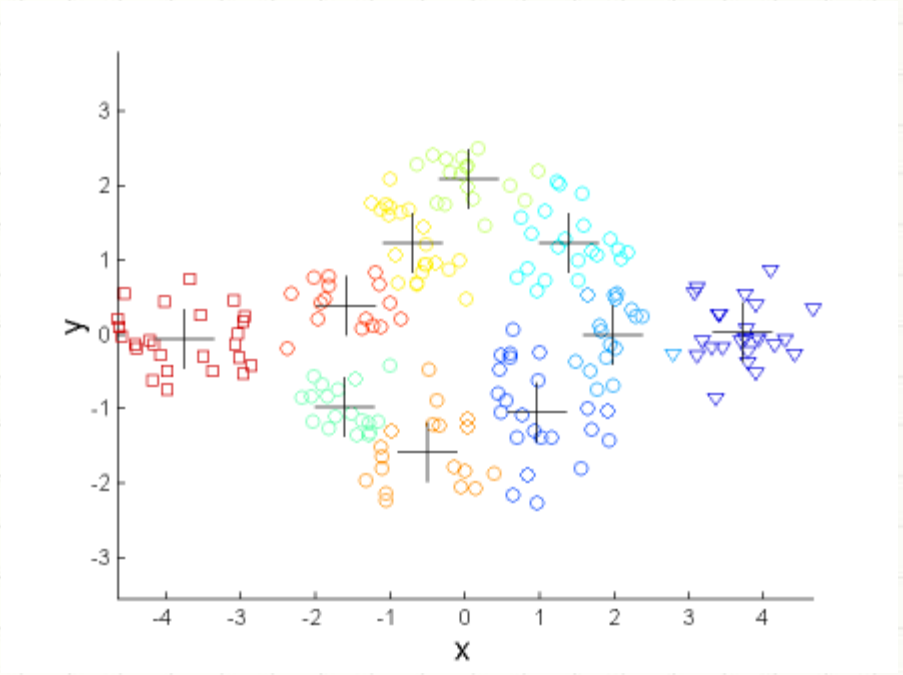


K-means (2 küme)

Çözüm



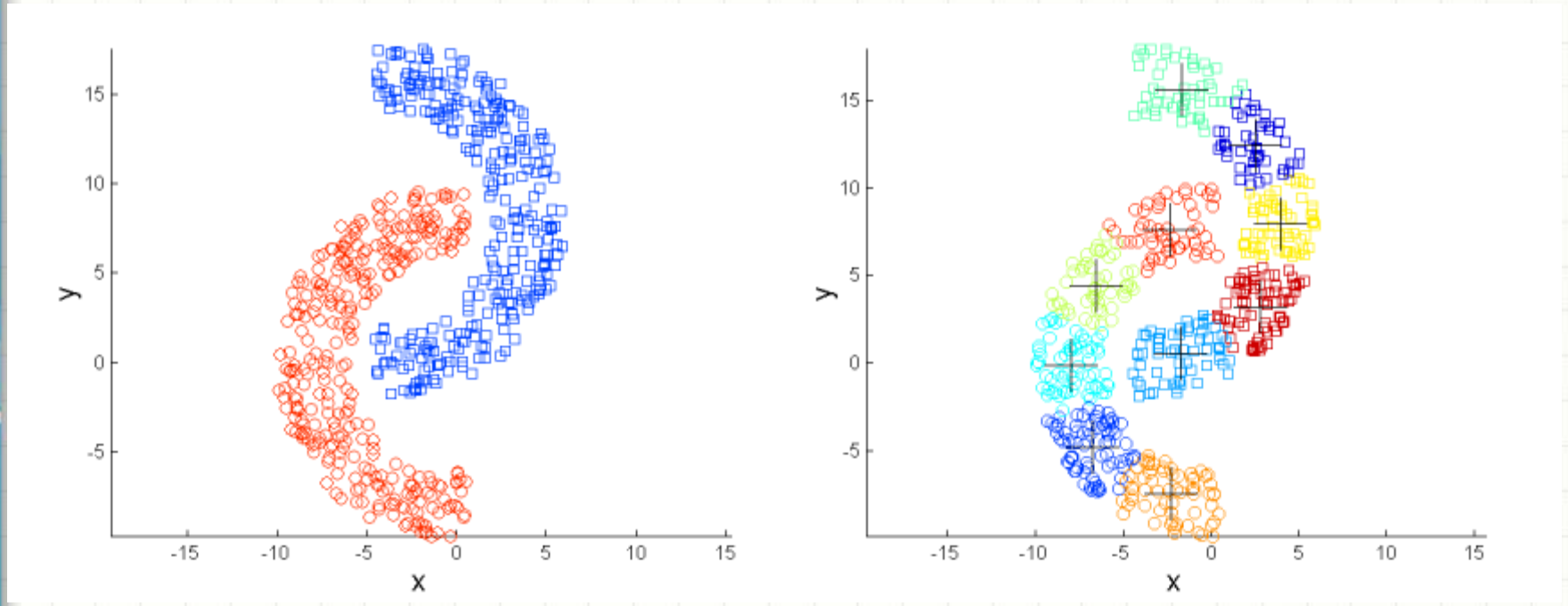
Orijinal kümeler



K-means kümeleri

- Gereğinden fazla kümeye ayrıştırmak:
  - Ancak sonunda birleştirmek gerek (nasıl?)

Fazladan kümeleme (bölme)



Orijinal kümeler

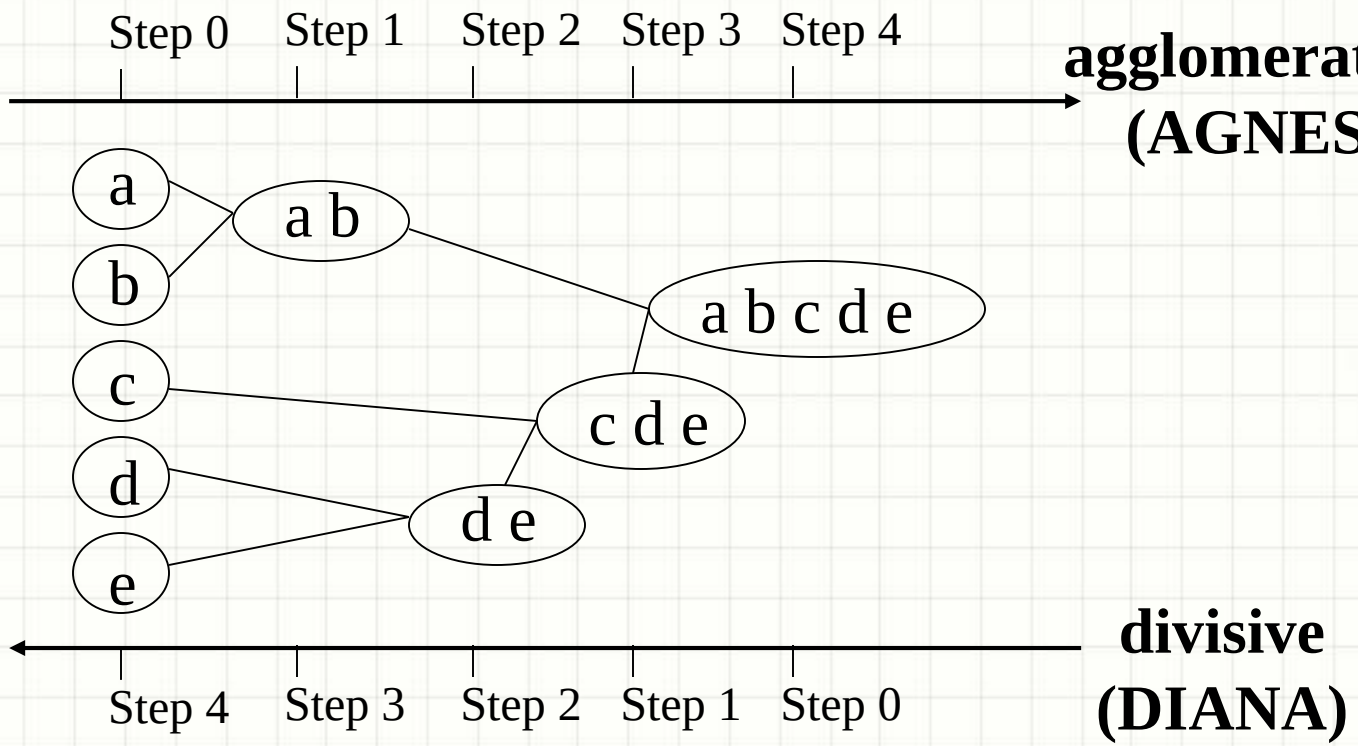
K-means kümeleri

Kümeleme Yöntemleri

- K-Means Kümeleme
- Hiyerarşik Kümeleme
- Yapay Sinir Ağları (SOM-Self Organized Feature Map)
- Genetik Algoritmalar

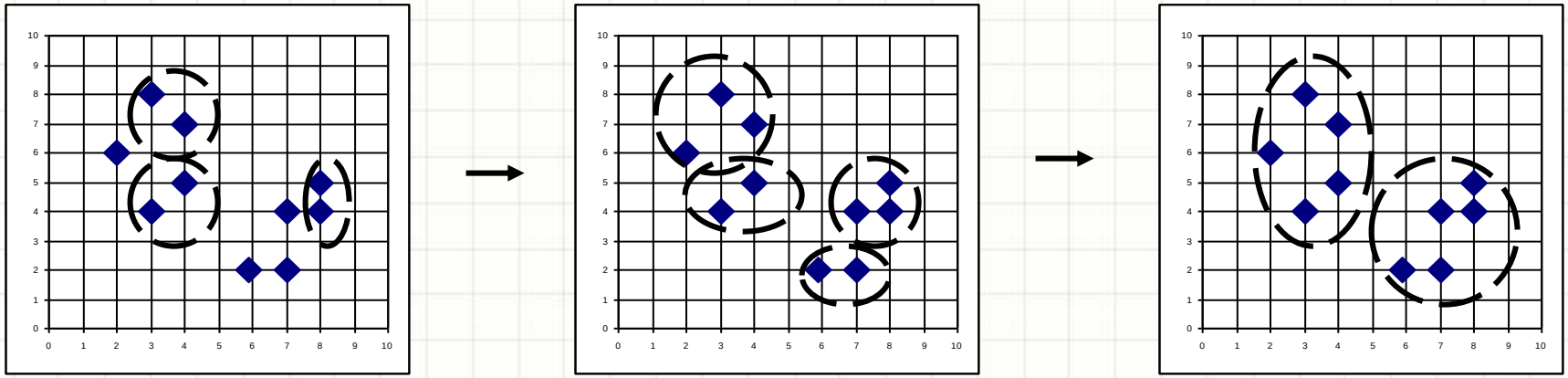
Hiyerarşik Kümeleme

- Küme sayısının bilinmesine gerek yoktur ancak bir sonlanma kriterine ihtiyaç duyar.



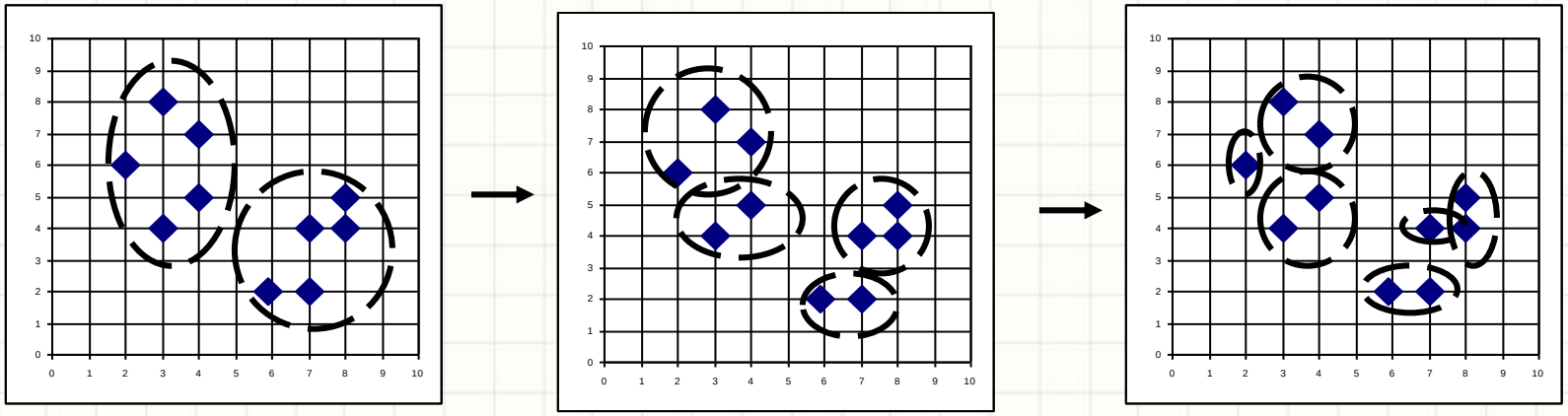
Hiyerarşik Kümeleme:
AGNES (Agglomerative Nesting)

- Başlangıçta her nesne bir küme olarak alınır.
- Aralarında **en az uzaklık** bulunan kümeler birleştirilir.
- Kümeler arasında mesafe tek bağ metodu (**single linkage** method) ile hesaplanır
- Bütün örnekler tek bir demet içinde kalana kadar birleştirme işlemi devam eder.



Hiyerarşik Kümeleme: DIANA (Divisive Analysis)

- AGNES'in yaptığı işlemlerin tersini yapar.
- Başlangıçta bütün örnekler bir demet içindeyken işlem sonunda her örnek bir demet oluşturur.



Hiyerarşik Kümeleme: Dendrogram

- Dendrogram: Kümelerin nasıl birleştiğini gösterir.