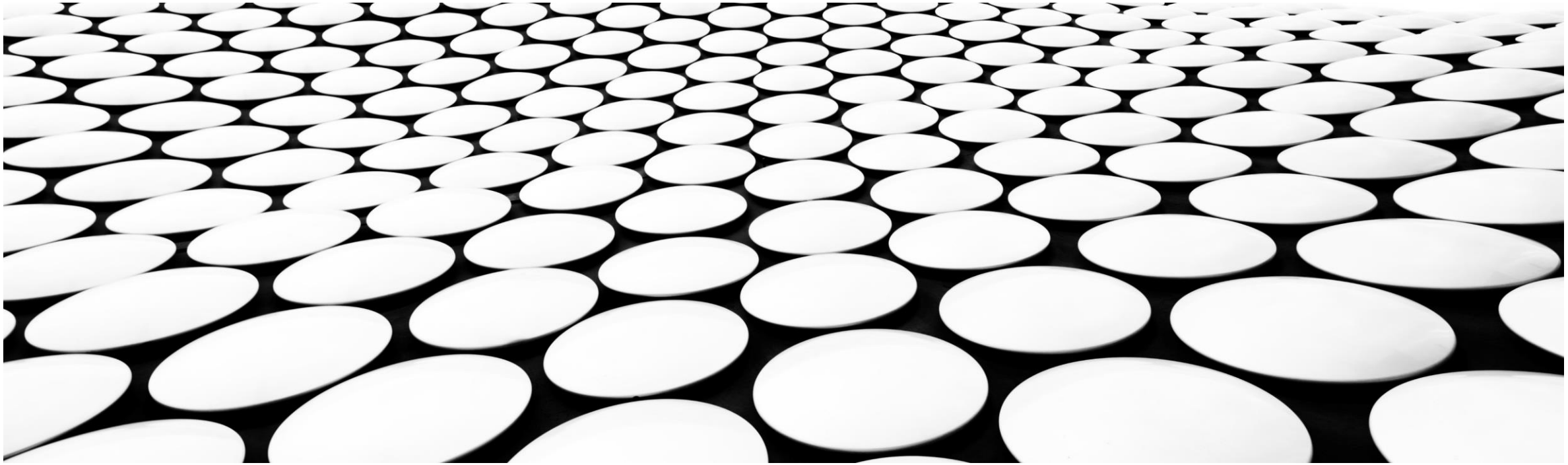
VERİ MADENCİLİĞİNDE KÜMELEME

Dr.Günay TEMÜR

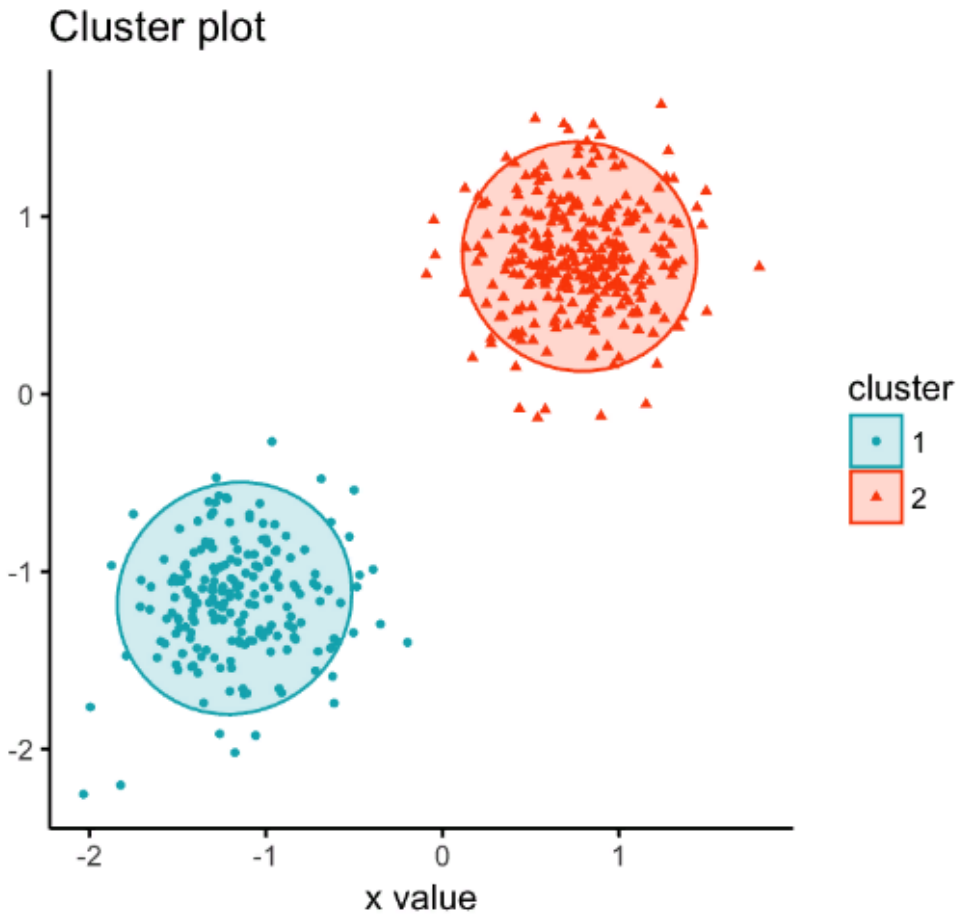


KÜMELEME UYGULAMALARI

- Örüntü tanıma
- Görüntü işleme
- Ekonomi
- Aykırılıkları belirleme
- WWW
  - Doküman kümeleme
  - Kullanıcı davranışlarını kümeleme
  - Kullanıcıları kümeleme
- Diğer veri madenciliği algoritmaları için bir önişleme adımı
- Veri azaltma – küme içindeki nesnelerin temsil edilmesi için küme merkezlerinin kullanılması

KÜMELEME (CLUSTERİNG)

- Bir veri kümesindeki verilerin, benzer özelliklerine göre belirli gruplara ayrılmasıdır. Bu grupların her birine **küme** denir.
- Kümeleme yöntemlerinin çoğu veri arasındaki uzaklıkları kullanır.
- Nesneleri kümelere (gruplara) ayırma işlemi
- Küme: birbirine benzeyen nesnelerden oluşan gruptur
  - Aynı kümedeki nesneler birbirine daha çok benzer
  - Farklı kümedeki nesneler birbirine daha az benzer



KÜMELEME ANALİZİ

- Sınıflandırmada olduğu gibi sahip olunan verileri gruplara ayırma işlemidir. Kümeleme yabancı kaynaklarda clustering ya da segmentation olarak adlandırılmaktadır. Sınıflandırma işleminde, sınıflar önceden belirli iken kümelemede sınıflar önceden belirli değildir.
- Verilerin hangi gruplara/kümelere, hatta kaç değişik gruba ayrılacağı eldeki verilerin birbirlerine olan benzerliğine göre belirlenir. Belirlenen her bir gruba da küme ismi verilir.
- Kümeleme analizi biyoloji, tıp, antropoloji, pazarlama, ekonomi ve telekomikasyon gibi birçok ve birbirinden çok farklı alanlarda kullanılmaktadır.

KÜMELEME ANALİZİ

- Kümeleme analizi ve algoritmaları 5 ana başlık altında incelenebilir.
- Bunlar:
  - Hiyerarşik Yöntemler (SLINK Algoritması, CURE Algoritması, CHAMELEON Algoritması, BIRC Algoritması)
  - Bölümlemeli Yöntemler (K-Means Algoritması, PAM Algoritması, CLARA Algoritması, CLARANS Algoritması)
  - Yoğunluğa Dayalı Yöntemler (DBSCAN Algoritması, OPTICS Algoritması, DENCLUE Algoritması)
  - Grid Temelli Algoritmalar (STING Algoritması, Dalga Kümeleme, CLINQUE Algoritması)
  - Genetik Algoritmalar olarak sayılabilir.

KÜMELEME ANALİZİ

- Kümeleme işleminde belirli gruplara ayrılan küme içindeki elemanların benzerliğinin fazla, kümeler arası benzerliğin ise en az olması amaçlanır.
- Kümeleme işlemi ile belirli özelliklere göre veriler az sayıda gruplara ayrılıp daha sonra her gruptaki verilerin özet profili çıkarılabilir.
- Benzer elemanlar gruplandırılarak veri seti küçültülebilir.
- Çok büyük miktardaki verileri analiz etmede kullanılabilecek en iyi yöntemlerden biri kümelemedir.

KÜMELEME UYGULAMALARI

- Kümeleme analizinin pratik problemler için birçok uygulaması mevcuttur.
- Bu uygulamalar kullanım amaçlarına göre iki grupta incelenebilir.
- Bunlar anlama için kümeleme ve fayda için kümelemedir.
  - Anlama için kümelemede oluşturulan sınıflar ya da nesne grupları konuyu daha iyi anlamayı, olayı bütün hatlarıyla kavramayı kolaylaştırır.
  - Fayda için kümeleme ise nesne hakkında elde edilen karakteristik bilgilerle nesnenin ait olduğu kümenin özellikleri hakkında bilgi edinmeyi sağlar

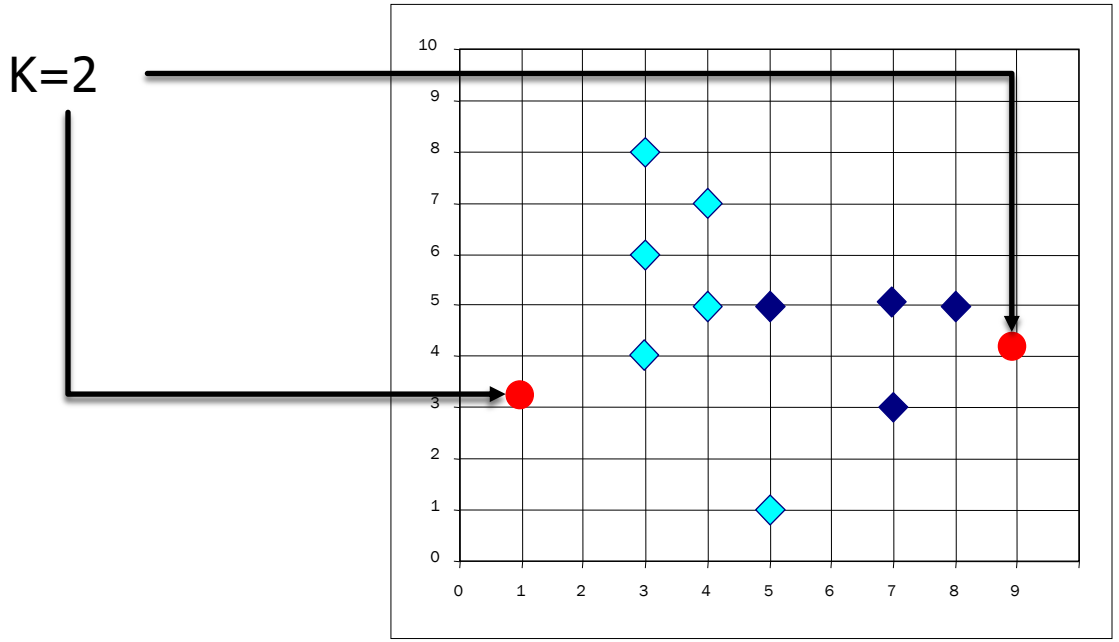
K-MEANS ALGORİTMASI

- Şu şekilde çalışır:

  1- Rastgele k tane nokta seçilir (k öbek sayısıdır)

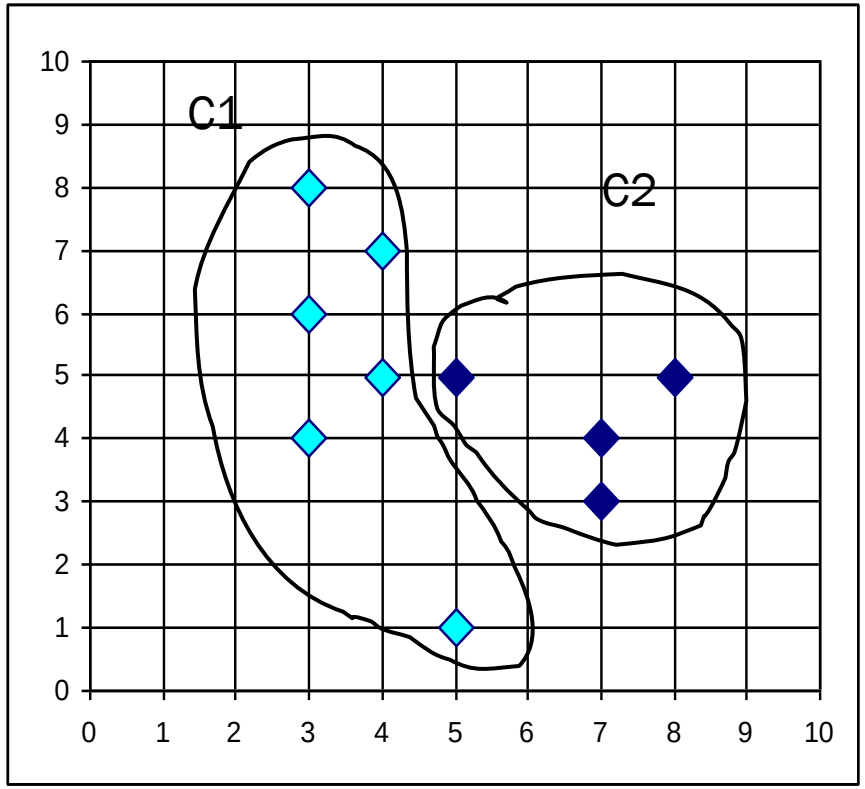
  2- Bu noktalar oluşturulacak öbeklerin merkezi varsayılır.

  3- Tüm noktaların merkezlere olan uzaklığı hesaplanır



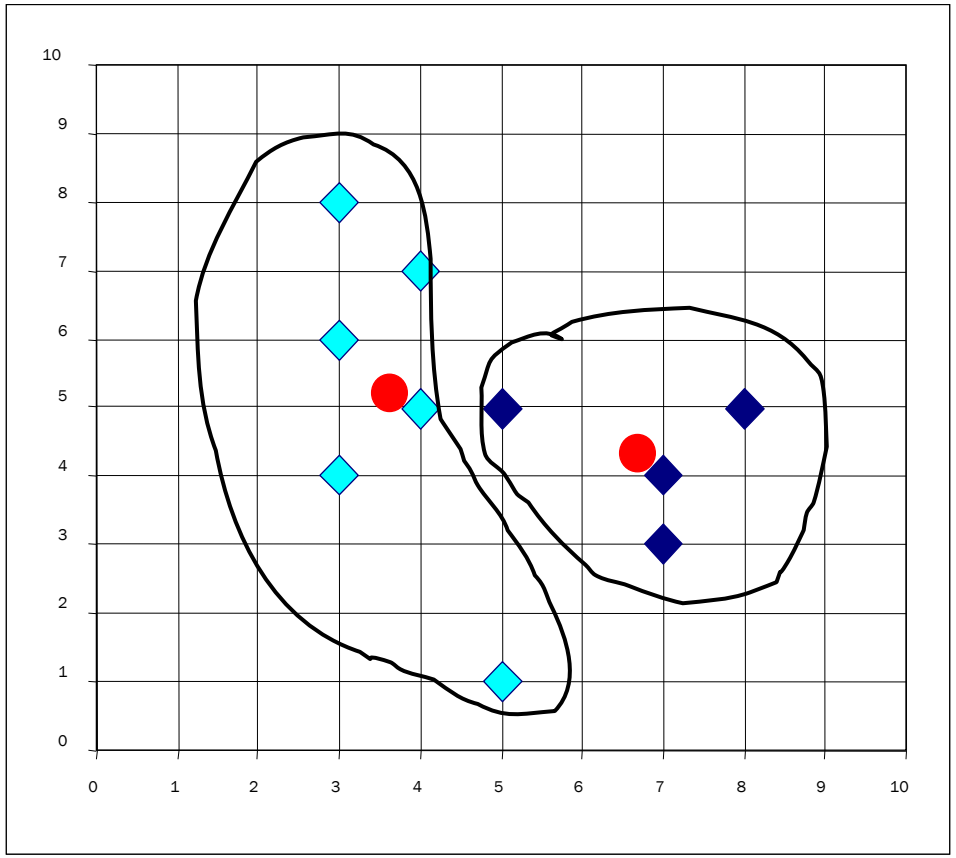
K-MEANS ALGORİTMASI

4- Her nokta kendisine yakın olan merkezin öbeğine atanır.



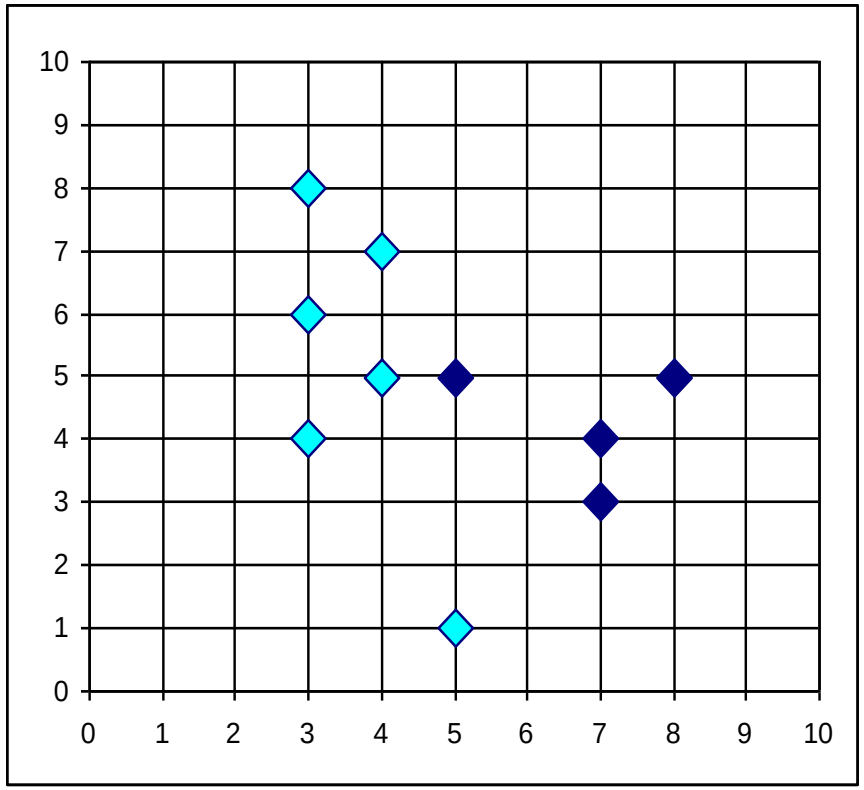
K-MEANS ALGORİTMASI

5- Oluşturulan öbeklerin merkezi yeniden hesaplanır.

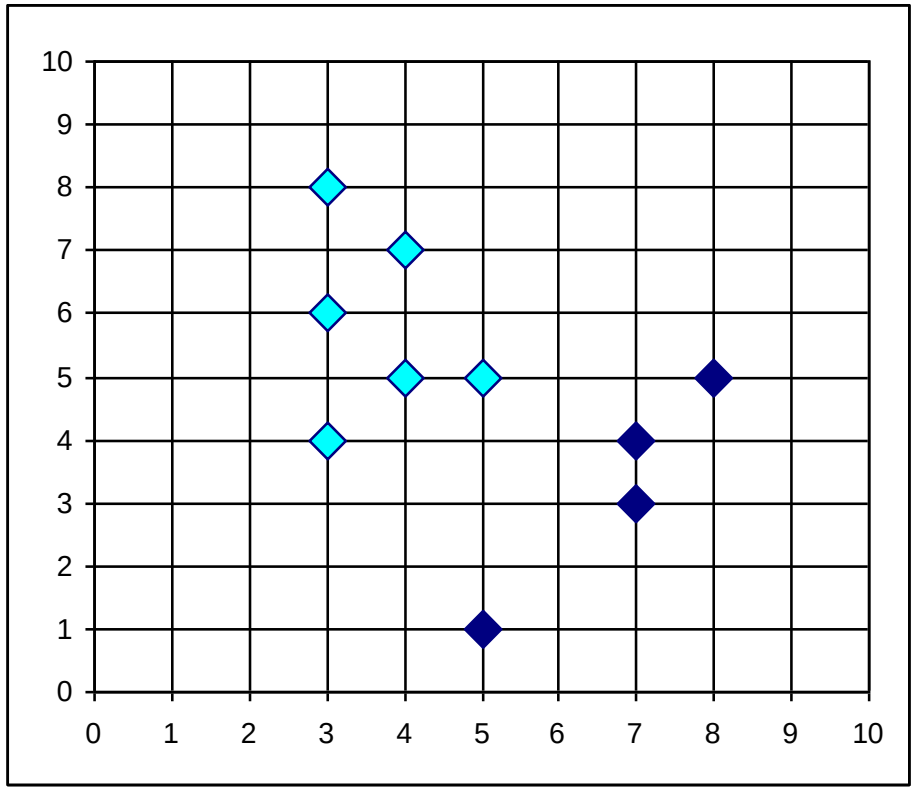


K-MEANS ALGORİTMASI

Before (Öncesi)

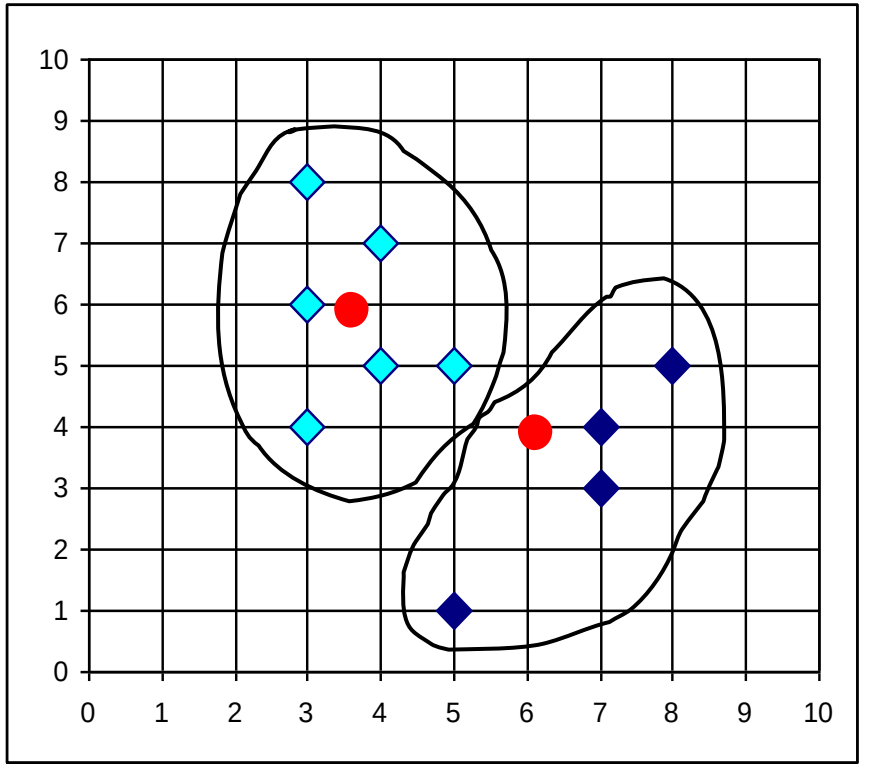


After (Sonrası)

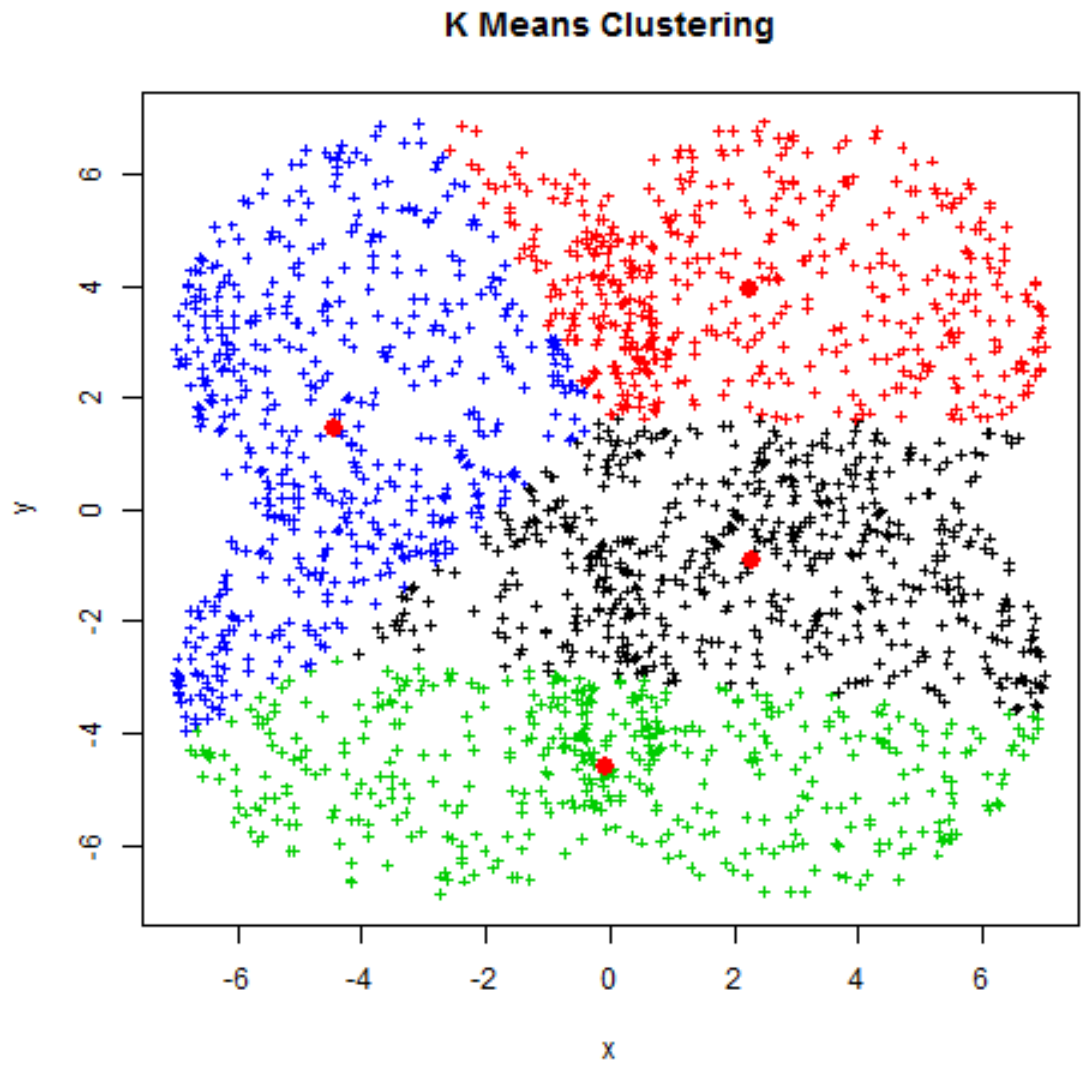


K-MEANS ALGORİTMASI

- 6- 3. Adımdan itibaren dikkate değer bir değişiklik gözlemlenirse iterasyonlara devam edilir.



ÖRNEK



VERİLEN BİR TABLO İÇİN KÜMELEME ÇALIŞMASI

- Verilen gözlem değerleri k-ortalamalar yöntemi ile kümelenmek isteniyor.

- Kümelerin sayısı başlangıçta k=2 kabul edilir. Rastgele iki küme belirlenir.

$$C_1 = \{X_1, X_2, X_4\}$$

$$C_2 = \{X_3, X_5\}$$

Gözlemler	Değişken 1	Değişken 2
X1	4	2
X2	6	4
X3	5	1
X4	10	6
X5	11	8

VERİLEN BİR TABLO İÇİN KÜMELEME ÇALIŞMASI

- Adım 1. a) Belirtilen iki kümenin merkezleri şu şekilde hesaplanır.

$$M_1 = \left\{\frac{4+6+10}{3}, \frac{2+4+6}{3}\right\} = \{6.67, 4.0\}$$

$$M_2 = \left\{\frac{5+11}{2}, \frac{1+8}{2}\right\} = \{8.0, 4.5\}$$

Gözlemler	Değişken 1	Değişken 2	Küme Üyeliği
X1	4	2	C_1
X2	6	4	C_1
X3	5	1	C_2
X4	10	6	C_1
X5	11	8	C_2

ÖRNEK

Gözlemler	Değişken 1	Değişken 2	Küme Üyeliği
X1	4	2	C_1
X2	6	4	C_1
X3	5	1	C_2
X4	10	6	C_1
X5	11	8	C_2

- b) Küme içi değişmeler şu şekilde hesaplanır.

$$e_1^2 = [(4 - 6{,}67)^2 + (2 - 4{,}0)^2] + [(6 - 6{,}67)^2 + (4 - 4{,}0)^2] + [(10 - 6{,}67)^2 + (6 - 4{,}0)^2] = 26{,}67$$

$$e_2^2 = [(5 - 8)^2 + (1 - 4{,}5)^2] + [(11 - 8)^2 + (8 - 4{,}5)^2] = 42{,}50$$

- Bu durumda toplam kare-hata şu şekilde hesaplanır.

$$E^2 = e_1^2 + e_2^2 = 26{,}67 + 42{,}50 = 69{,}17$$

ÖRNEK

Gözlemler	Değişken 1	Değişken 2	Küme Üyeliği
X1	4	2	C_1
X2	6	4	C_1
X3	5	1	C_2
X4	10	6	C_1
X5	11	8	C_2

- C) M_1 ve M_2 merkezlerinden olan uzaklıkların minimum olması istendiğinden aşağıdaki hesaplamalar yapılır. Öklid uzaklık formülü kullanılarak söz konusu mesafeler hesaplanır. Örneğin (M_1,X_1) noktaları arasındaki uzaklık M_1={6.67, 4.00} ve X_1={4, 2} olduğuna göre şu şekilde hesaplanır.

$$d(M_1, X_1) = \sqrt{(6{,}67 - 4)^2 + (4 - 2)^2} = 3{,}33$$

$$d(M_2, X_1) = \sqrt{(8 - 4)^2 + (4{,}5 - 2)^2} = 4{,}72$$

- Bu işlemler sonucunda X_1 gözlem değerinin M_1 ve M_2 merkezlerine olan uzaklıkları göz önüne alındığında $d(M_1, X_1) < d\ (M_2, X_1)$ olduğu görülür. Bu durumda M_1 merkezinin X_1 gözlem değerine daha yakın olduğu anlaşılır. O halde $X_1 \in C_1$ olarak kabul edilir. Benzer biçimde tüm gözlem değerleri için tablo oluşturulur.

ÖRNEK

Gözlemler	M_1'den uzaklık	M_2'den uzaklık	Küme Üyeliği
X1	$d(M_1, X_1) = 3{,}33$	$d(M_2, X_1) = 4{,}72$	C_1
X2	$d(M_1, X_2) = 0{,}67$	$d(M_2, X_2) = 2{,}06$	C_1
X3	$d(M_1, X_3) = 3{,}43$	$d(M_2, X_3) = 4{,}61$	C_1
X4	$d(M_1, X_4) = 3{,}89$	$d(M_2, X_4) = 2{,}50$	C_2
X5	$d(M_1, X_5) = 5{,}90$	$d(M_2, X_5) = 4{,}61$	C_2

ÖRNEK

Gözlemler	Değişken 1	Değişken 2	Küme Üyeliği
X1	4	2	C_1
X2	6	4	C_1
X3	5	1	C_2
X4	10	6	C_1
X5	11	8	C_2

Gözlemler	Değişken 1	Değişken 2	Küme Üyeliği
X1	4	2	C_1
X2	6	4	C_1
X3	5	1	C_1
X4	10	6	C_2
X5	11	8	C_2

- Bu durumda yeni kümeler şu şekilde olacaktır.

$$C_1 = X_1, X_2, X_3$$

$$C_2 = X_4, X_5$$

- Adım 2. Yukarıda belirtilen iki kümenin merkezleri şu şekilde hesaplanır.

$$M_1 = \left\{\frac{4+6+5}{3}, \frac{2+4+1}{3}\right\} = \{5, 2.33\}$$

$$M_2 = \left\{\frac{10+11}{2}, \frac{6+8}{2}\right\} = \{10.5, 7\}$$

ÖRNEK

- b) Küme içi değişmeler şu şekilde hesaplanır.

$$e_1^2 = [(4-5)^2 + (2-2.33)^2] + [(6-5)^2 + (4-2.33)^2] + [(5-5)^2 + (1-2.33)^2] = 9.33$$

$$e_2^2 = [(10-10.5)^2 + (6-7)^2] + [(11-10.5)^2 + (8-7)^2] = 2{,}50$$

- Bu durumda toplam kare-hata şu şekilde hesaplanır.

$$E^2 = e_1^2 + e_2^2 = 9.33 + 2{,}50 = 11{,}83$$

- Bu değerin bir önceki iterasyonda elde edilen $E^2 = 69{,}17$ değerinden daha küçük olduğu anlaşılmaktadır.

ÖRNEK

- M_1 ve M_2 merkezlerinden gözlem değerlerine olan uzaklıklar hesaplanır. Bunun sonucunda $d\ (M_1, X_1) < d\ (M_2, X_1)$ olduğu görülür. Bu durumda M_1 merkezinin X_1 gözlem değerine daha yakın olduğu anlaşılır. O halde $X_1 \in C_1$ olarak kabul edilir. Benzer biçimde tüm gözlem değerleri için tablo oluşturulur.

Gözlemler	M_1'den uzaklık	M_2'den uzaklık	Küme Üyeliği
X1	$d(M_1, X_1) = 1.05$	$d(M_2, X_1) = 8.20$	C_1
X2	$d(M_1, X_2) = 1.94$	$d(M_2, X_2) = 5.41$	C_1
X3	$d(M_1, X_3) = 1.33$	$d(M_2, X_3) = 8.14$	C_1
X4	$d(M_1, X_4) = 6.20$	$d(M_2, X_4) = 1.12$	C_2
X5	$d(M_1, X_5) = 8.25$	$d(M_2, X_5) = 1.12$	C_2

ÖRNEK

- Bu durumda yeni kümeler şu şekilde oluşacaktır.

$$C_1 = \{X_1, X_2, X_3\}$$

$$C_2 = \{X_4, X_5\}$$

- Kümelerde önceki adıma göre herhangi bir değişme olmadığı için iterasyona son verilir.

FORMÜLİZASYON

- Öklid
  - $d(i,j) = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2}$
- Minkowski
  - $d(i,j) = \left[\sum_{k=1}^{p}(\left|x_{ik} - x_{jk}\right|^m)\right]^{\frac{1}{m}}$
- Manhattan
  - $d(i,j) = \sum_{k=1}^{p}(\left|x_{ik} - x_{jk}\right|)$

K-EN YAKIN KOMŞU ALGORİTMASI

- Yeni bir elemanın bir gruba dahil olmasında; K-en yakın komşu algoritması, gözlem değerlerinden oluşan bir küme için aşağıdaki adımları içerir.
  - a) K parametresi belirlenir. Bu parametre verilen bir noktaya en yakın komşuların sayısıdır.
  - b) Bu algoritma verilen bir noktaya en yakın komşuları belirleyeceği için söz konusu nokta ile diğer tüm noktalar arasındaki uzaklıklar tek tek hesaplanır.
  - c) Yukarıda hesaplanan uzaklıklara göre satırlar sıralanır ve bunlar arasından en küçük olan k tanesi seçilir.
  - d) Seçilen satırların hangi kategoriye ait oldukları belirlenir ve en çok tekrarlanan kategori değeri seçilir.
  - e) Seçilen kategori, tahmin edilmesi beklenen gözlem değerinin kategorisi olarak kabul edilir.

ÖRNEK 1.

- Verilen gözlem tablosu X1 ve X2 nitelikleri ve Y sınıfından oluşmaktadır. Bu gözlem değerine bağlı olarak yeni bir gözlem değeri olan X1=8, X2=4 değerlerinin yani (8,4) gözleminin hangi sınıfa dahil olduğunu K-En yakın komşu algoritması ile bulunuz

X1	X2	Y
2	4	Kötü
3	6	İyi
3	4	İyi
4	10	Kötü
5	8	Kötü
6	3	İyi
7	9	İyi
9	7	Kötü
11	7	Kötü
10	2	Kötü

ÖRNEK 1.

- A) K'nın belirlenmesi: k=4 olarak kabul edildi.
- B) Uzaklıkların hesaplanması: (8,4) noktası ile gözlem değerlerinin her biri arasındaki uzaklıklar Öklid uzaklığına göre hesaplanır.
- Biçiminde birinci gözlem olan (2,4) noktası ile (8,4) noktası arasındaki uzaklık
- Benzer şekilde uzaklıklar hesaplandığında tablodaki sonuç ortaya çıkacaktır.

- $d(i,j) = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2}$
- $d(i,j) = \sqrt{(2-8)^2+(4-4)^2} = 6$

ÖRNEK 1.

- (8,4) noktasının gözlem değerlerine olan uzaklıkları,

X1	X2	Y
2	4	6
3	6	5,39
3	4	5
4	10	7,21
5	8	5
6	3	2,24
7	9	5,1
9	7	3,16
11	7	4,24
10	2	2,83

ÖRNEK 1.

- c) En küçük uzaklıkların belirlenmesi: Satırlar sıralanarak en küçük k=4 tanesi belirlenir. Bu dört nokta verilen (8,4) noktasına en yakın gözlem değerleridir.

X1	X2	Uzaklık
2	4	6
3	6	5,39
3	4	5
4	10	7,21
5	8	5
6	3	2,24
7	9	5,1
9	7	3,16
11	7	4,24
10	2	2,83

ÖRNEK 1.

- d) Seçilen satırların ilişkin sınıfların belirlenmesi: (8,4) noktasına en yakın olan gözlem değerlerinin Y sınıfları göz önüne alınır ve içinde hangi değerin baskın olduğu araştırılır. Bu dört gözlem içinde bir tane İYİ 3 tane KÖTÜ sınıfı vardır.
- e) Yeni gözlemin sınıfı: KÖTÜ değerlerinin sayısı İYİ değerlerinin sayısından fazla olduğu için (8,4) **noktasının sınıfı KÖTÜ olarak belirlenir.**

X1	X2	Uzaklık	Sıra	k komşunun Y Değeri
2	4	6		
3	6	5,39		
3	4	5		
4	10	7,21		
5	8	5		
6	3	2,24		İyi
7	9	5,1		
9	7	3,16		Kötü
11	7	4,24		Kötü
10	2	2,83		Kötü

ÖRNEK 2.

- Aşağıda verilen gözlem tablosunda Y sınıf niteliğini ifade etmektedir. Bu verilere dayanarak (7, 8, 5) noktasının hangi sınıf değerine sahip olduğunu belirleyelim. Gözlemlerin gerçek değerleri değil normalize edilmiş değerleri kullanılacaktır. Gözlem değerlerini (0,1) aralığına çekmek için min-max normalleştirmesi kullanılacaktır.

X1	X2	X3	Y
10	5	19	Evet
8	2	4	Hayır
18	16	6	Hayır
12	15	8	Evet
3	15	15	Evet

ÖRNEK 2.

- Min-max normalleştirmesi sonucu dönüştürülen değerler:
- $X^* = \frac{x - x_{min}}{x_{max} - x_{min}} (min - \max\ normalizasyonu)$
- Aday noktanın normalizasyon değeri (0.27, 0.43, 0.07)

X1	X2	X3	Y
0,47	0,21	1	Evet
0,33	0	0	Hayır
1	1	0,13	Hayır
0,6	0,93	0,27	Evet
0	0,93	0,73	Evet

ÖRNEK 2.

- a) K'nın belirlenmesi: k=3 kabul edildi.
- b) Uzaklıkların hesaplanması: (0.27, 0.43, 0.07) noktası ile gözlem değerlerinin her biri arasındaki uzaklıklar Öklid uzaklığına göre hesaplanır.
- $d(i,j) = \sqrt{(0.47-0.27)^2+(0.21-0.43)^2+(1-0.07)^2} = 0.98$

X1	X2	X3	Uzaklık
0,47	0,21	1	0,98
0,33	0	0	0,44
1	1	0,13	0,93
0,6	0,93	0,27	0,63
0	0,93	0,73	0,87

ÖRNEK 2.

- c) En küçük uzaklıkların belirlenmesi: Satırlar sıralanarak en küçük k=3 tanesi belirlenir.

X1	X2	X3	Uzaklık	Sıra
0,47	0,21	1	0,98	5
0,33	0	0	0,44	1
1	1	0,13	0,93	4
0,6	0,93	0,27	0,63	2
0	0,93	0,73	0,87	3

ÖRNEK 2.

- d) Seçilen satırların ilişkin sınıfların belirlenmesi: (0.27, 0.43, 0.07) noktasına en yakın olan gözlem değerlerinin Y sınıfları göz önüne alınır ve içinde hangi değerin baskın olduğu araştırılır. Bu üç gözlem içinde bir tane HAYIR 2 tane EVET sınıfı vardır.
- e) Yeni gözlemin sınıfı: EVET değerlerinin sayısı HAYIR değerlerinin sayısından fazla olduğu için **(7, 8, 5) gözleminin sınıfı EVET olarak kabul edilir**.

X1	X2	X3	Uzaklık	Sıra	K komşunun Y değeri
0,47	0,21	1	0,98	5	
0,33	0	0	0,44	1	Hayır
1	1	0,13	0,93	4	
0,6	0,93	0,27	0,63	2	Evet
0	0,93	0,73	0,87	3	Evet

AĞIRLIKLI OYLAMA

- K-en yakın komşu algoritması sınıfı bilinmeyen gözlem değeri için k gözlem içindeki en fazla tekrar eden sınıfın seçilmesi esasına dayanmaktadır. Ancak seçilen bu sınıf sadece k komşunun göz önüne alınması nedeniyle her zaman uygun olmayabilir. Bu son aşamada k komşu arasında en çok tekrarlanan sınıfı seçme yöntemi yerine ağırlıklı oylama (weighted voting) denilen bir yöntem uygulanabilir.
- Söz konusu ağırlıklı oylama yöntemi gözlem değerleri için aşağıdaki bağıntıya göre ağırlıklı uzaklıkların hesaplanmasına dayanır.

$$d(i,j)' = \frac{1}{d(i,j)^2}$$

AĞIRLIKLI OYLAMA

$$d(i,j)' = \frac{1}{d(i,j)^2}$$

d(i,j) ifadesi i ve j gözlemleri arasındaki Öklid uzaklığıdır. Her bir sınıf değeri için bu uzaklıkların toplamı hesaplanarak ağırlıklı oylama değeri elde edilir. En büyük ağırlıklı oylama değerine sahip olan sınıf değeri yeni gözlemin ait olduğu sınıf olarak kabul edilir.

AĞIRLIKLI OYLAMA SONUCU

- Ağırlıklı Oylama sonucunda da Örnek 2.'deki değerin sınıfının EVET olduğu görülür.

X1	X2	X3	Uzaklık	Sıra	K komşunun Y değeri	Ağırlıklı Oylama
0,47	0,21	1	0,98	5		
0,33	0	0	0,44	1	Hayır	$1/(0,44)^2=5,17$
1	1	0,13	0,93	4		
0,6	0,93	0,27	0,63	2	Evet	$1/(0,63)^2=2,52$
0	0,93	0,73	0,87	3	Evet	$1/(0,87)^2=3,84$

(Evet)Toplam=2,52+3,84=6,66

VERİ SETLERİ

VERİ SETLERİ

- UCI
- https://archive.ics.uci.edu/ml/index.php

- Kaggle
- https://www.kaggle.com/

BİTTİ ☺