KÜMELEME

Birbirlerine benzeyen veri parçalarını ayırma gruplama işlemidir. Temel yaklaşım Öklid Manhattan Minkowski uzaklık bağıntıları kümeleme işleminde alt işlem olarak kullanılır.

K en yakın komşu algoritması ve en uzak komşu bilinen yöntemlerdir. Hiyerarşik

K means yöntemi Hiyerarşik olmayan

Uzaklık ölçütleri

Kümeleme yöntemlerinin çoğu gözlem değerleri arasındaki uzaklıkların hesaplanması esasına dayanır. Çeşitli değişkenlerden oluşan X matrisi verilsin

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \\ x_{41} & x_{42} & x_{43} \\ x_{51} & x_{52} & x_{53} \end{bmatrix}$$

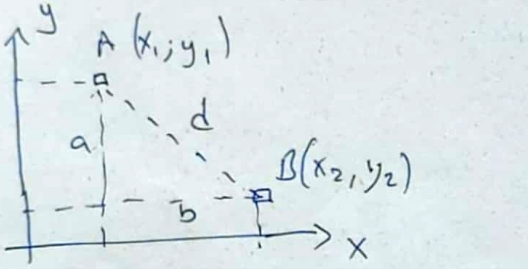
x_{11} x_{12} x_{13} Birinci gözlem
x_{21} x_{22} x_{23} İkinci " Bu iki gözlem arası uzaklık $d(2,1)$ biçiminde gösterilebilir.

Yukarıdaki X matrisinin her bir satırının diğerine olan uzaklığı $d(i,j)$ biçiminde ifade edilecek olursa D uzaklıklar matrisi

$$D = \begin{vmatrix} 0 & & & & \\ d(2,1) & 0 & & \text{Simetrik} & \\ d(3,1) & d(3,2) & 0 & & \\ d(4,1) & d(4,2) & d(4,3) & 0 & \\ d(5,1) & d(5,2) & d(5,3) & d(5,4) & 0 \end{vmatrix}$$

Yukarıdaki matrisin üst kısmı alt kısmının simetriğidir. $d(i,j) = d(j,i)$ Birden fazla uzaklık bağıntısı vardır. Yaygın olan üçtanesi

a) Öklid



$$d(A,B) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \qquad d(i,j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$

b) Manhattan

$$d(i,j) = \sum_{k=1}^{p} (|x_{ik} - x_{jk}|) \qquad i,j = 1,2,\dots n \quad k = 1,2,\dots p$$

c) Minkowski uzaklığı

$$d(i,j) = \sum_{k=1}^{p} \left[|x_{ik} - x_{jk}|^m \right]^{1/m} \qquad i,j = 1,2,\dots,n \quad k = 1,2,\dots,p$$

$m = 2$ için öklid

```
clc
clear all
close all
a=normrnd(20,30,300,2);
b=normrnd(100,30,300,2);
plot(a(:,1),a(:,2),'o'),hold on
plot(b(:,1),b(:,2),'*'),
hold off
data=[a;b];
figure
plot(data(:,1),data(:,2),'o'),hold on
%küme merkezleri seçiliyor...Adım 2
m1=[-20 -30];
m2=[200 200];
kume1=[];
kume2=[];
plot(m1(:,1), m1(:,2),'*m')
plot(m2(:,1), m2(:,2),'*m')
pause
%dataların küme merkezlerine uzaklıkları ve kümeler oluş...
%Adım 3
for iterasyon=1:20
    iterasyon
    for i=1:length(data)
        if dist(m1,data(i,:)')<dist(m2,data(i,:)')
            kume1=[kume1;data(i,:)];
        else
            kume2=[kume2;data(i,:)];
        end
    end
    %adım 4
    m1=mean(kume1);
    m2=mean(kume2);
    plot(kume1(:,1),kume1(:,2),'og')
    plot(kume2(:,1),kume2(:,2),'ok')

    plot(m1(:,1), m1(:,2),'*r')
    plot(m2(:,1), m2(:,2),'*r')
    pause
end
```

ÖRNEK

A, B, ve C üç değişken ve 5 gözlemden oluşan veri için uzaklıkları hesaplayalım

Gözlem	A	B	C
1	2	3	1
2	4	1	3
3	5	7	3
4	4	8	2
5	3	9	5

Öklid

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + (x_{i3} - x_{j3})^2}$$

İkinci gözlem ile birinci gözlem arası uzaklık

$$d(2,1) = \sqrt{(x_{21} - x_{11})^2 + (x_{22} - x_{12})^2 (x_{23} - x_{13})^2} = \sqrt{(4-2)^2 + (1-3)^2 + (3-1)^2} = 3.46$$

Üçüncü gözlem ile birinci gözlem arası uzaklık

$$d(3,1) = \sqrt{(x_{31} - x_{11})^2 + (x_{32} - x_{12})^2 + (x_{33} - x_{13})^2} =$$

$$\sqrt{(5-2)^2 + (7-3)^2 + (3-1)^2} = 5.39$$

Öklid

	1	2	3	4	5
1	0				
2	3.46	0			
3	5.39	6.08	0		
4	5.48	7.07	1.73	0	
5	7.28	8.31	3.46	3.32	0