Hiyerarşik Olmayan Kümeleme

k-Ortalamalar Yöntemi

Başlangıçta belirlenen belli sayıdaki küme içi toplam ortalamayı minimize etmek amaçlanır. N boyutlu uzayda N ~~adet~~ örnekli kümelerin verildiğini varsayalım. Bu uzay $\{C_1, C_2, \dots, C_k\}$ biçiminde k adet kümeye ayrılsın O zaman

$$\sum n_k = N \qquad k = 1, 2, \dots, k$$

o.ü C_k kümesinin ortalama vektörü M_k;

$$M_k = \frac{1}{n_k} \sum_{i=1}^{n_k} x_{ik}$$

X_k C_k kümesine ait olan i. örnektir. C_k kümesi için hata-kare her bir C_k örneği ile onun merkezi (centroid) arasındaki öklid uzaklıkları toplamıdır. Bu hataya küme içi değişme adı da verilir. Küme içi değişmeler

$$e_{i}^2 = \sum_{i=1}^{n_k} (x_{ik} - M_k)^2$$

k kümesini içeren bütün kümeler uzayı için hata-kare küme içindeki değişmelerin toplamıdır. O halde söz konusu hata-kare değeri şu şekilde hesaplanır

$$E_k^2 = \sum_{k=1}^{K} e_k^2$$

Kare-hata kümeleme yönteminin amacı verilen k değeri için E_k^2 değerini minimize eden k kümelerini bulmaktır.
O halde k-ort. algoritmasında E_k^2 değerinin bir önceki iterasyona göre azalması beklenir.

Algoritma

K-ortalama algoritmasına başlamadan önce k küme sayısının belirlenmesi gerekir. k belirlendikten sonra her bir kümeye gözlem değerleri atanır böylece $C_1, C_2, \dots, C_k$ kümeleri belirlenmiş olur.

a) Herbir ~~kü~~ kümenin merkezi belirlenir. Bu merkezler ~~M_1, M_2~~ $M_1, M_2, \dots, M_k$ biçiminde

b) $e_1, e_2, \dots, e_k$ küme içi değişimler hesaplanır. Bu değişmeleri toplamı olan E_k^2 değeri bulunur

c) M_k merkez değerleri ile gözlem değerleri arasındaki uzaklıklar hesaplanır. Bir gözlem değeri hangi merkeze yakın ise o merkez ile ilgili küme içine dahil edilir.

d Yukardaki b ve c adımları, kümelerde her hangi bir değişim olmayıncaya kadar devam ettirilir.

Uygulama

x_i	Değişken1	Değişken2
x_1	4	2
x_2	6	4
x_3	5	1
x_4	10	6
x_5	11	8

k-ort ile k=2 için verileri kümeleyiniz

k=2 için tesadüfi olarak iki kümeye belirliyoruz

$C_1 = \{x_1, x_2, x_4\}$

$C_2 \{x_3, x_5\}$

Gözlem	Değişken1	Değişken2	Küme üyeliği
x_1	4	2	C_1
x_2	6	4	C_1
x_3	5	1	C_2
x_4	10	6	C_1
x_5	11	8	C_2

Adım 1

Yukarıda belirtilen kümenin merkezleri

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

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

Küme içi değerler

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

$$e_1^2 = 26,67$$

$$e_2^2 = [(5-8)^2 + (1-4.5)^2] + [(11-8)^2 + (8-4.5)]$$

$$e_2^2 = 42,5$$

$$E^2 = e_1^2 + e_2^2 = 26.67 + 42.5 \quad \boxed{E^2 = 69.17}$$

M_1 ve M_2 merkezlerinden uzaklıkların minimum olması için öklid uzaklık formulü ile mesafeler hesaplanır.

(M_1, x_1) ikilisi için $M_1 = \{6.67, 4\}$ $x_1 = \{4, 2\}$

$$d(M_1, x_1) = \sqrt{(6.67-4)^2 + (4-2)^2} = 3.33$$

$M_2 = \{8, 4.5\}$ ve $x_1 = \{4, 2\}$ olduğuna göre (M_2, x_1) uzaklığı

$$d(M_2\ x_1) = \sqrt{(8-4)^2 + (4.5-2)^2} = 4.72$$

$$d(M_1, x_1) < d(M_2, x_1)$$

Bu durumda M_1 merkezinin x_1 gözlem değerine daha yakın olduğu anlaşılır. O halde $x_1 \in C_1$ kabul edilir. Benzer şekilde

Gözlemler	M_1 den uzaklık	M_2 den uzaklık	Küme üyeliği
x_1	$d(M_1, x_1)$ 3.33	$d(M_2, x_1) = 4.72$	C_1
x_2	$d(M_1, x_2)$ 0.67	$d(M_2, x_2)$ 2.06	C_1
x_3	$d(M_1, x_3)$ 3.43	$d(M_2, x_3)$ 4.61	C_1
x_4	$d(M_1, x_4)$ 3.89	$d(M_2, x_4)$ 2.5	C_2
x_5	$d(M_1, x_5)$ 5.9	$d(M_2, x_5)$ 4.61	C_2

Bu durumda yeni üyelikler.

$C_1 = \{x_1, x_2, x_3\}$

$C_2 = \{x_4, x_5\}$

Adım 2

~~Yukarıda~~ belirtilen yeni küme merkezleri

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

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

Kume içi değişimler şu şekilde hesaplanır.

$e_1^2 = [(4-5)^2 + (2-2{,}33)^2] + [(6-5)^2 + (4-2.33)^2] + [(5-2.33)^2 + (1-2.33)^2]$

$e_1^2 = 9.33$

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

$e_2^2 = 2.50$

Toplam kare-hata

$E^2 = e_1^2 + e_2^2 = 9.33 + 2.5$

$E^2 = 11.83$

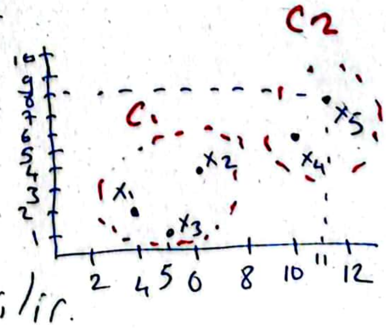
Bir önceki $E^2 = 69.17$ güncel $E^2 = 11.83$ yeni toplam kare hata daha küçüktür.

M_1 ve M_2 merkezlerinden gözlemlerin uzaklıkları hesaplanır. $(d\, M_1 < X_i) < d\, M_2 < X_i)$ kriteri ile atamalar yapılır.

Gözlemler	M_1 den uzaklık	$d(M_2, X_i)$	Küme üyeliği
X_1	$d(M_1, X_1) = 1{,}05$	8.20	C_1
X_2	$d(M_1, X_2) = 1{,}94$	5.41	C_1
X_3	$d\, M_1, X_3 = 1{,}33$	8.14	C_1
X_4	$d(M_1, X_4) = 6.20$	1.12	C_2
X_5	$d(M_1, X_5) = 8.25$	1.12	C_2

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

$C_2 = \{X_4, X_5\}$



Değişme olmadığı için itarasyona son verilir.