Azərbaycan TEXNİKİ Universiteti

Mühazirə 10: Klasterləşdirmə metodları

tex.f.d.,dos. Yadigar İmamverdiyev

2021, payız semestri

01.12.2021

Azərbaycan TEXNİKİ Universiteti

Mühazirənin planı

- Giriş
- Klasterləşdirmə metodlarının təsnifatı
- **K-means** metodu
- İyerarxik klasterləşdirmə
- Klasterləşdirmənin keyfiyyətinin qiymətləndirilməsi

- Bu mühazirə təqdimatı hazırlanarkən İnternetdə əlyetər çox sayda təqdimatdan istifadə edilmişdir. Təəssüf ki, onların hər birinin müəllifini qeyd etmək mümkün olmadı.
- Bu müəlliflərin hər birinə dərindən təşəkkür edirəm.

Azərbaycan TEXNİKİ Universiteti

Klasterləşdirmənin ümumi tərifi

*Klasterləşdirmə alqoritmləri öz xarakteristikalarına görə **yaxın obyektləri** bir klasterdə, **uzaq obyektləri** isə müxtəlif klasterlərdə yerləşdirməklə obyektlərin verilmiş çoxluğunu **qruplara** (klasterlərə) bölür.*

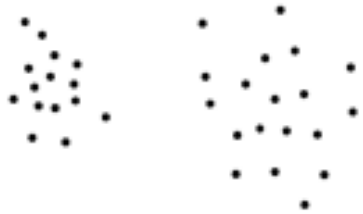
Azərbaycan TEXNİKİ Universiteti

Klasterləşdirmənin məqsədləri

- Verilənlərin sonrakı emalını asanlaşdırmaq, obyektlər çoxluğunu oxşar obyektlər qruplarına ayırmaq və hər bir qrup ilə ayrılıqda işləmək (klassifikasiya, reqresiya, proqnozlaşdırma məsələləri);
- Hər bir klasterdən bir nümayəndə saxlamaqla verilənlərin həcmini azaltmaq (verilənlərin sıxılmasl məsələsi);
- Klasterlərin heç birinə yaxın düşməyən qeyri-tipik obyektləri ayırmaq (*birsinifli klassifikasiya məsələsi*)
- Obyektlər çoxluğunun iyerarxiyasını qurmaq (**taksonomiya məsələsi**).

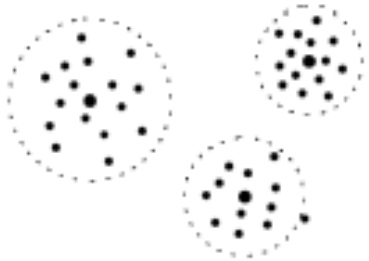
Azərbaycan TEXNİKİ Universiteti

Klaster strukturlarının tipləri



Klasterlərarası məsafələr, bir qayda olaraq, klasterdaxili məsafələrdən böyükdür.

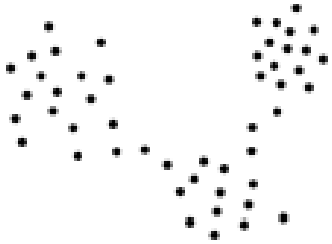
Lentşəkilli klasterlər



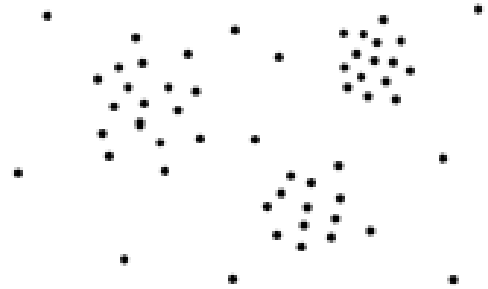
Mərkəzi olan klasterlər

Azərbaycan TEXNİKİ Universiteti

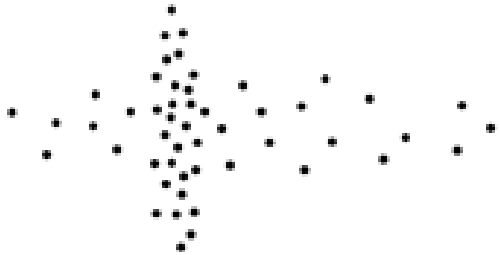
Klaster strukturlarının tipləri



Klasterlər körpülərlə birləşə bilərlər.



Klasterlər çox seyrək yerləşmiş obyektlərin seyrək fonu üzərində yerləşə bilərlər.



Klasterlər kəsişə bilərlər.

Azərbaycan TEXNİKİ Universiteti

Klaster strukturlarının tipləri

Klasterlər oxşarlığa görə deyil, müntəzəmliyin digər növünə görə yarana bilər.

Klasterlər, ümumiyyətləyə, olmaya bilərlər.

- Hər bir klasterləşdirmə metodunun öz məhdudiyyətləri var və klasterlərin yalnız müəyyən tiplərini ayırırlar.
- “Klaster strukturunun tipi” anlayışı da metoddan asılıdır və formal tərifi yoxdur.

Azərbaycan TEXNİKİ Universiteti

Məsələnin formal qoyuluşu

- Test nümunələrinin $X = \{x_1, x_2, \dots, x_n\}$ toplusu və nümunələr arasında ***məsafə funksiyası*** var.
- *X*-i kəsişməyən altçoxluqlara (klasterlərə) elə bölmək tələb edilir ki, hər bir altçoxluq oxşar obyektlərdən ibarət olsun, müxtəlif altçoxluqların obyektləri isə əhəmiyyətli dərəcədə fərqlənsinlər.

Azərbaycan TEXNİKİ Universiteti

Klasterləşdirmə metodlarının təsnifatı

- Bölünmə (Partitioning) əsasında klasterləşdirmə
- İyerarxik klasterləşdirmə
- Sıxlığa əsaslanan klasterləşdirmə (EM, DBSCAN)

Azərbaycan TEXNİKİ Universiteti

k-ortalar (k means) metodu

k-ortalar metodu (k-means)

Fərz edək ki, obyektlər çoxluğu artıq müəyyən qaydada *K* qrupa (klasterə) bölünüb, C_k ilə *k*-ci klasteri işarə edək. Tutaq ki, *C*(*k*) – *k*-cı obyektin aid olduğu qrupun nömrəsinə bərabərdir.

k-cı klasterə aid olan nöqtələr sisteminin ağırlıq mərkəzini m_k ilə işarə edək:

$$m_k = \frac{1}{N_k} \sum_{x_i \in C_k} x_i, k = 1,\ldots, K$$

Məqsəd: $\min_{C, m_k} \sum_{i=1}^{n} \rho(x_i, m_k)$

Azərbaycan TEXNİKİ Universiteti

k-ortalar (k means) metodu

① Klasterlərin sayını (K) qərarlaşdırıb onların başlanğıc ağırlıq mərkəzlərini təsadüfi seçirik və obyektləri yaxın olduqları mərkəzə uyğun klasterlərə yerləşdiririk.

② Klasterlərin m_k ağırlıq mərkəzlərini yenidən hesablayırıq.

③ x_i nöqtəsindən bütün m_k-lara $\rho(x_i, m_k)$ məsafələrini hesablayırıq.

Əgər $\rho(x_i, m_j) = \min_{k=1,\dots,K} \rho(x_i, m_k)$ olarsa, onda $x_i \in C_j$.

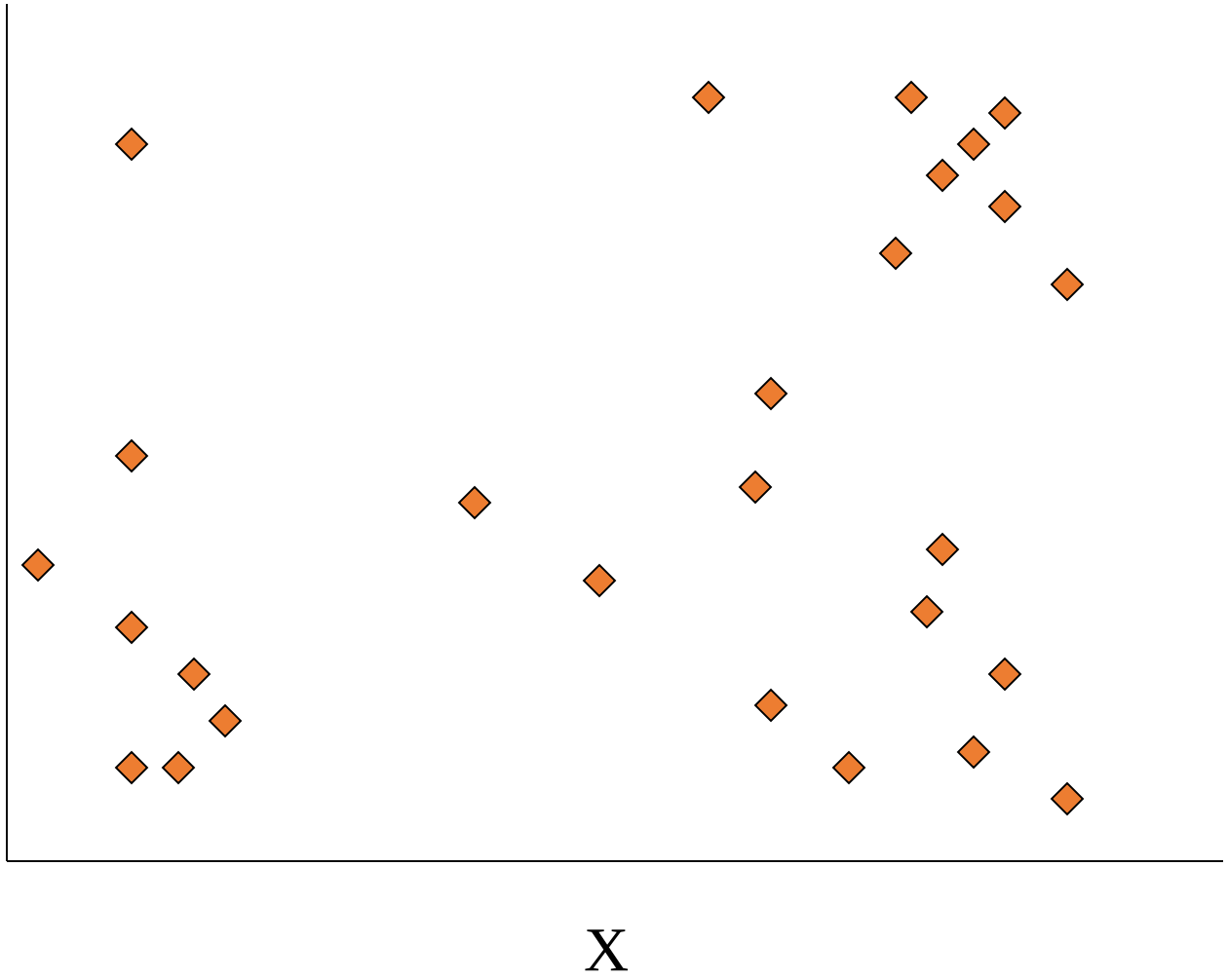
④ Addım 3-ü bütün x_i, i=1,..., n obyektləri üçün təkrarlayırıq.

⑤ Əgər heç olmasa, bir C_k klasteri dəyişirsə, onda *Addım 2*-yə keçirilir, əks halda klasterləşdirmə prosesi sona çatır.

Azərbaycan TEXNİKİ Universiteti

k-ortalar alqoritminə misal (1)

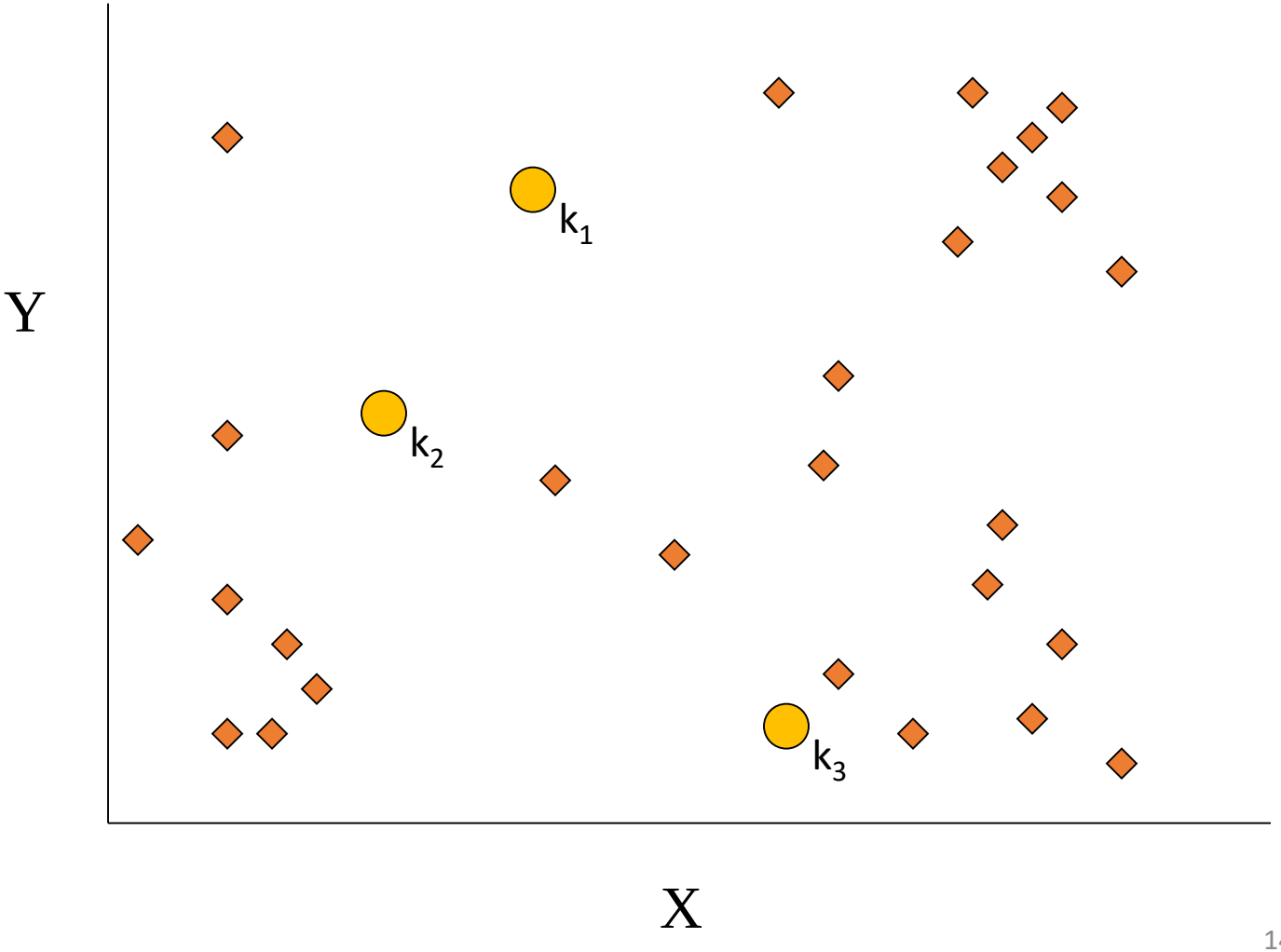
Neçə klasterin olmasını qərarlaşdırırıq. Tutaq ki, k=3.



Azərbaycan TEXNİKİ Universiteti

k-ortalar alqoritminə misal (2)

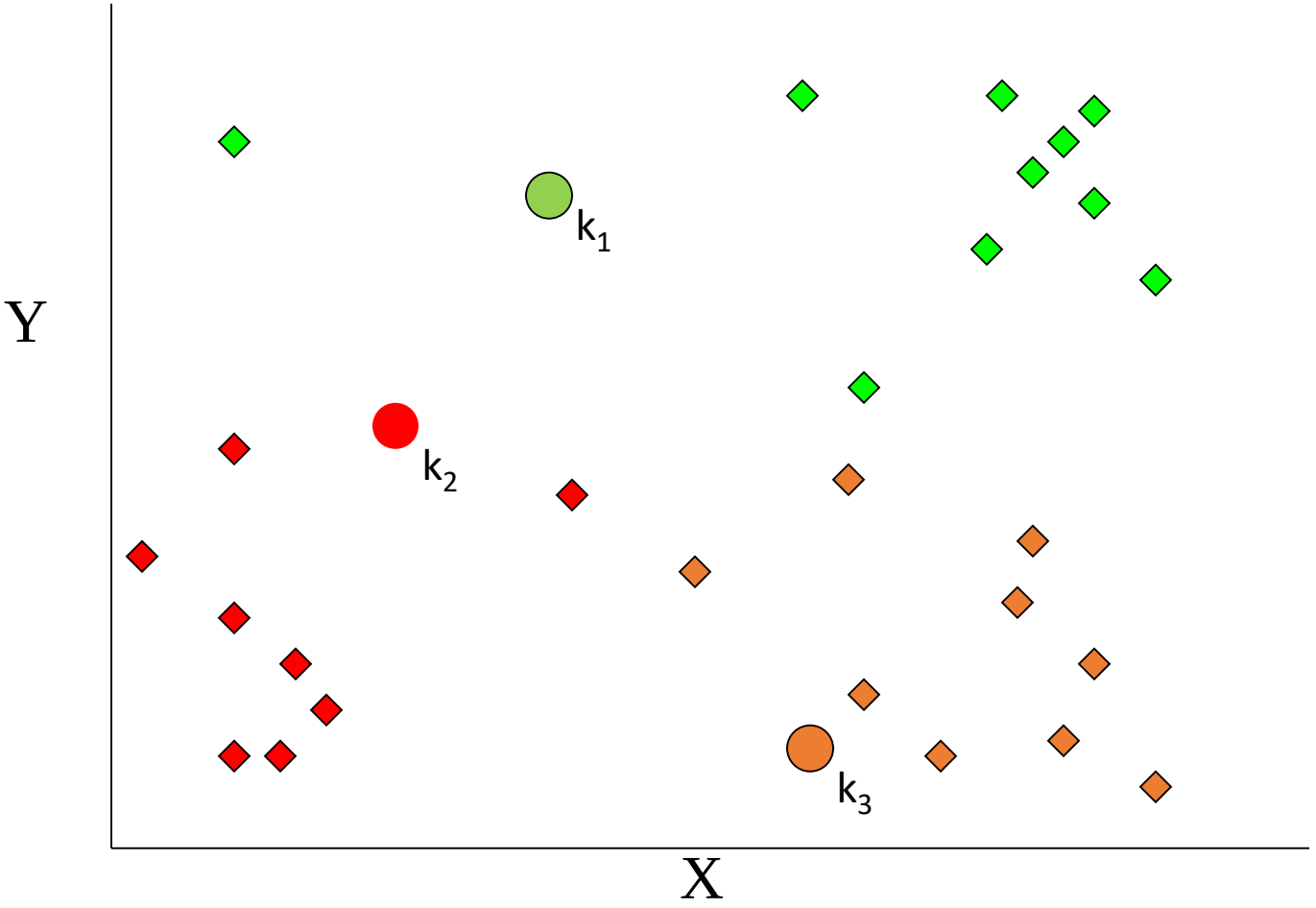
3 klasterin ağırlıq mərkəzini təsadüfi seçirik.



Azərbaycan TEXNİKİ Universiteti

k-ortalar alqoritminə misal(3)

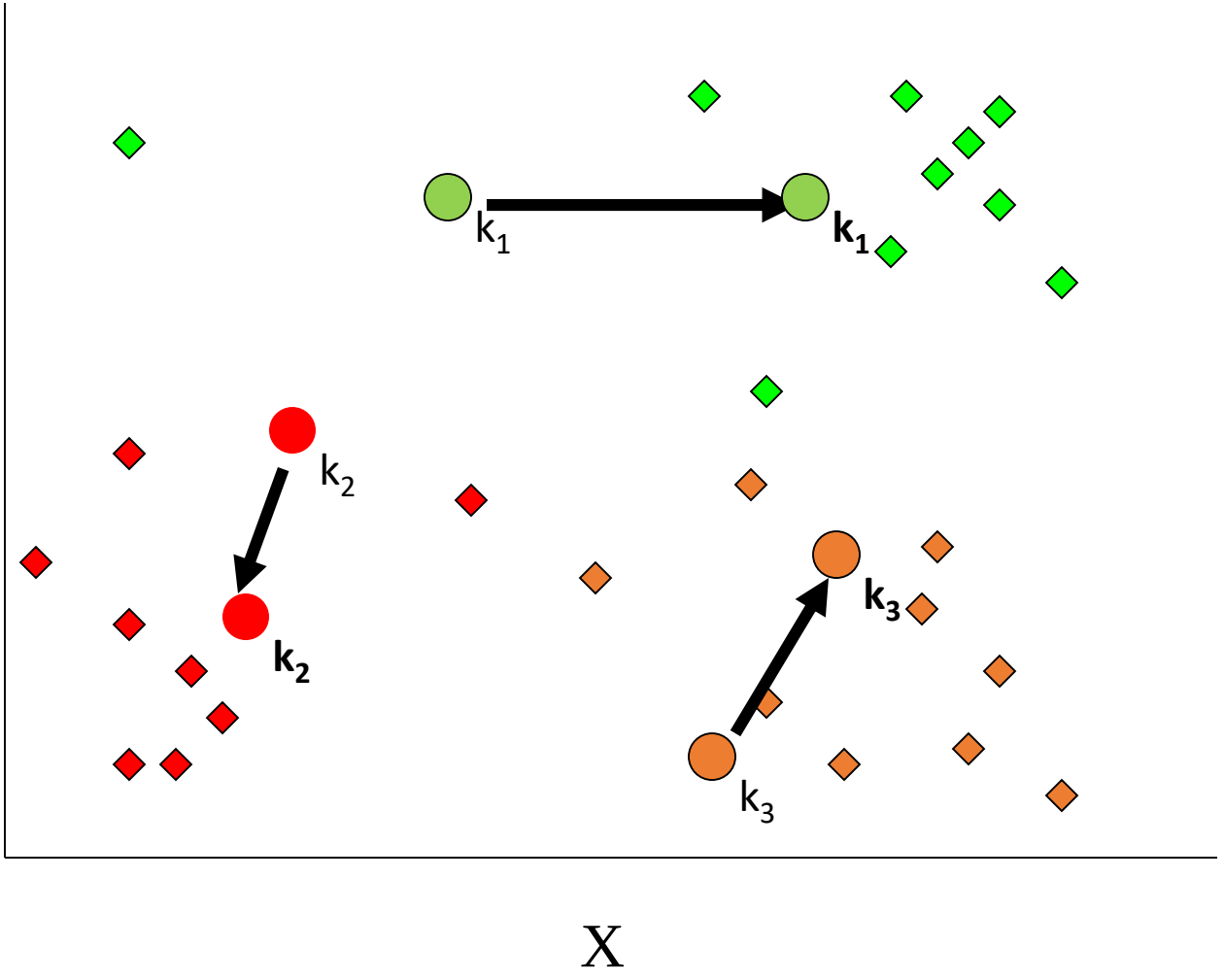
Seçilmiş məsafə metrikası ilə məsafələri hesablayırıq.
Hər bir nöqtəni yaxın olduğu mərkəzə uyğun klasterə aid edirik.



Azərbaycan TEXNİKİ Universiteti

k-ortalar alqoritminə misal (4)

Hər bir klasterin mərkəzini yenidən hesablayırıq və mərkəzin yerini dəyişirik.

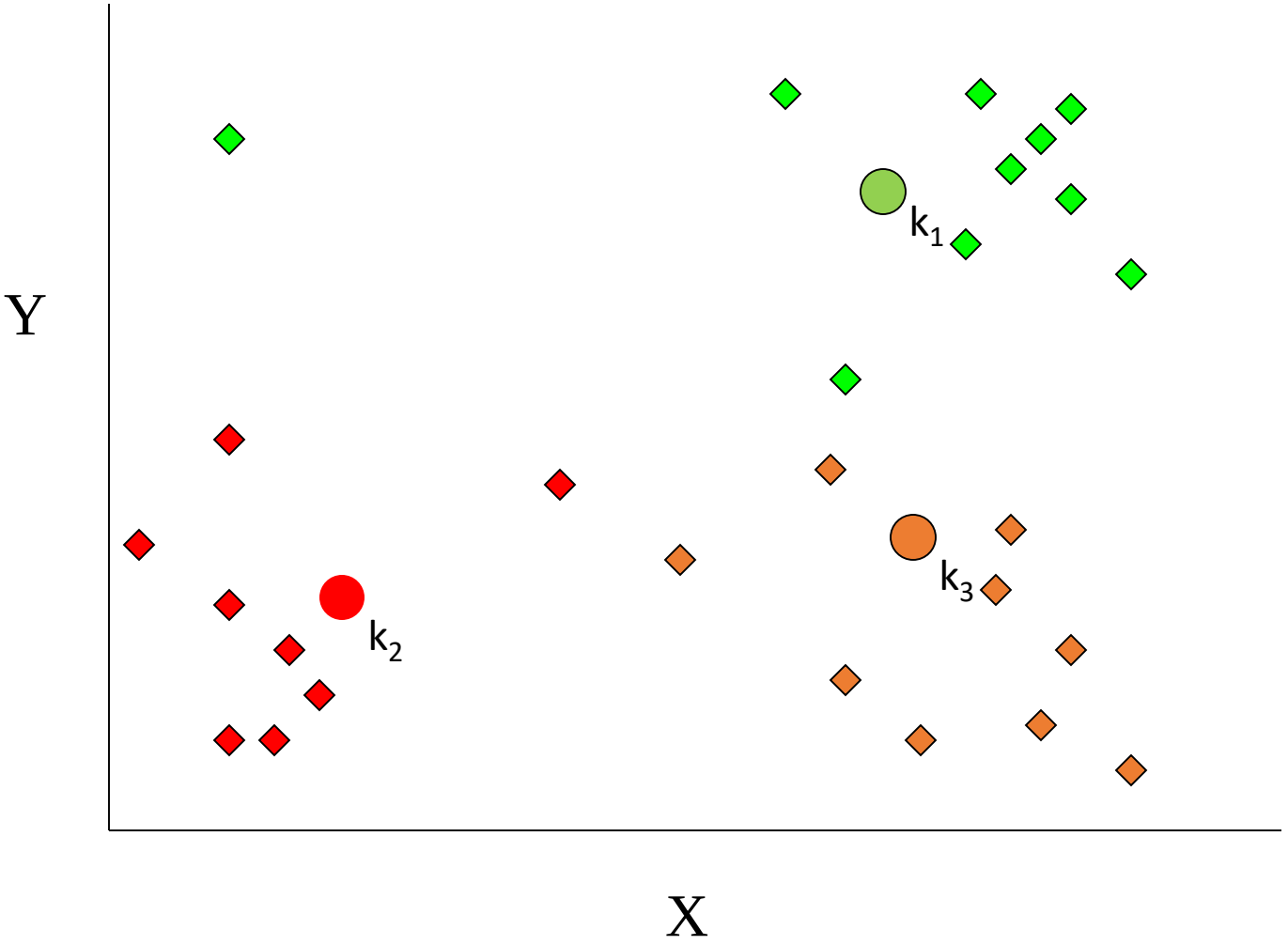


Azərbaycan TEXNİKİ Universiteti

k-ortalar alqoritminə misal (5)

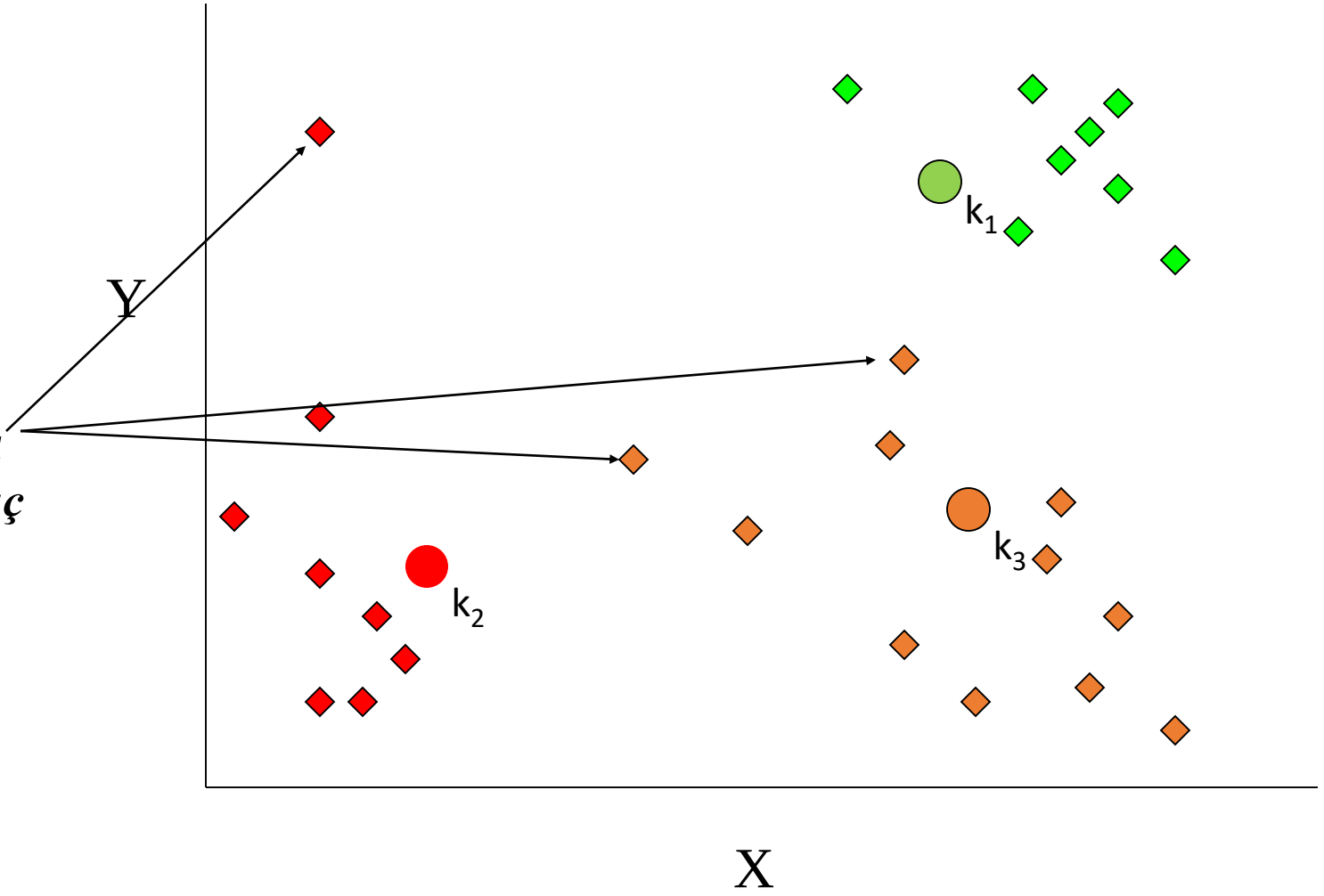
Nöqtələrdən mərkəzlərə məsafələr yenidən hesablanır və mərkəzlərə ən yaxın nöqtələr yenidən paylanır

Sual: Hansı nöqtələr yenidən paylandı?



Azərbaycan TEXNİKİ Universiteti

k-ortalar alqoritminə misal (6)

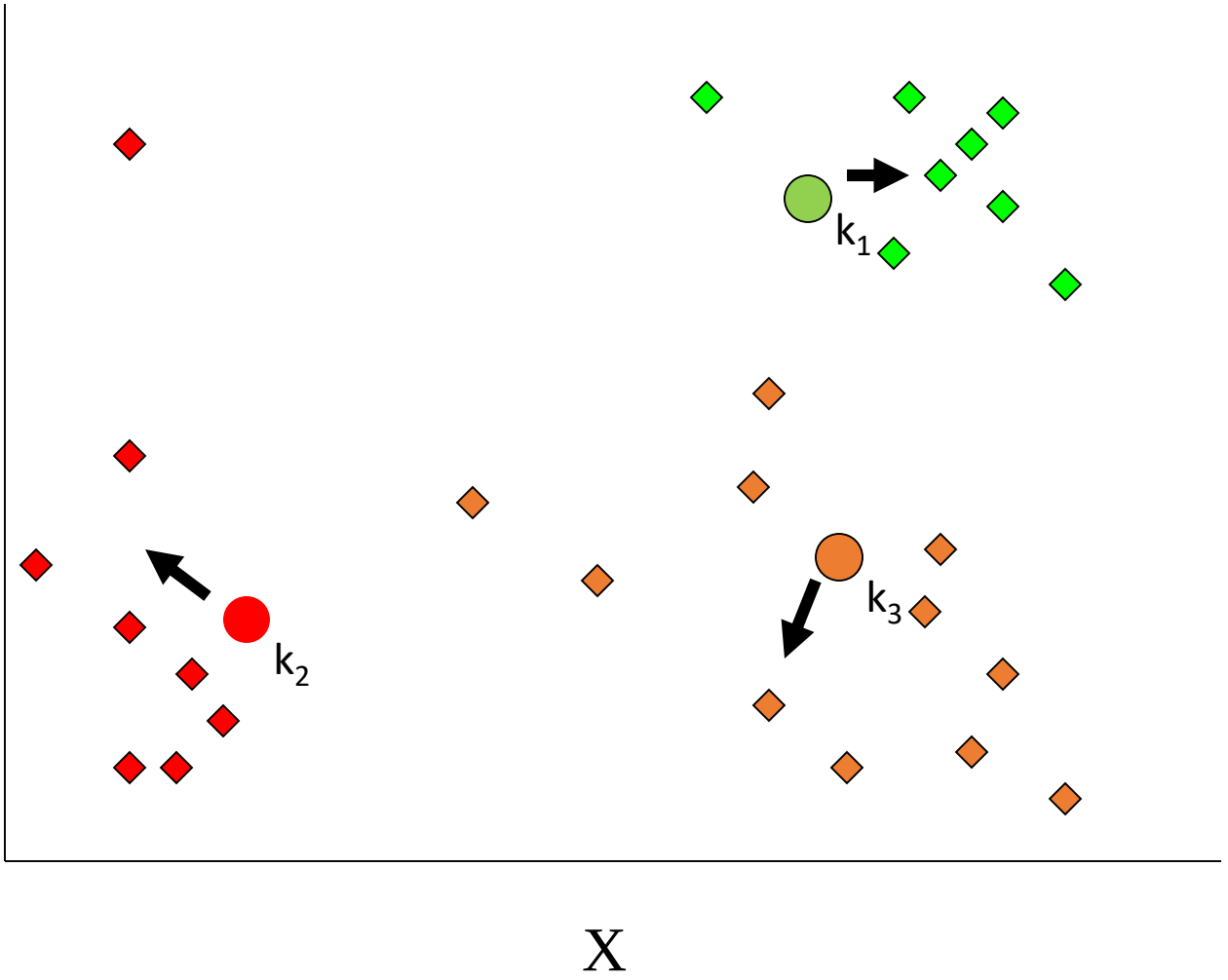


Cavab: animasiya etdirilən üç nöqtə

Azərbaycan TEXNİKİ Universiteti

k-ortalar alqoritminə misal (7)

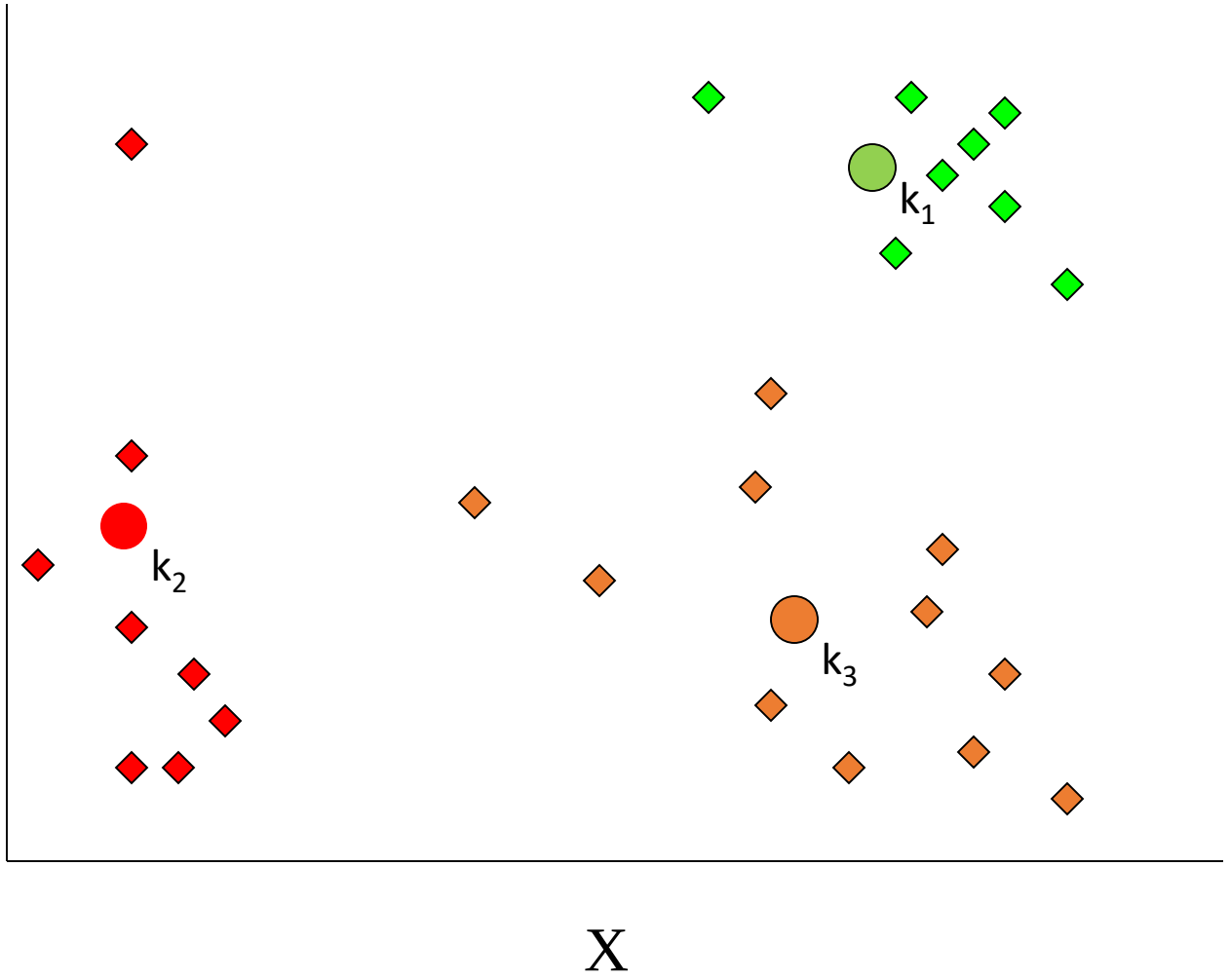
Klasterlərin ağırlıq mərkəzləri yenidən hesablanır



Azərbaycan TEXNİKİ Universiteti

k-ortalar alqoritminə misal (8)

Klaster mərkəzləri klasterlərin ağırlıq mərkəzlərinə köçürülür



Azərbaycan TEXNİKİ Universiteti

k-ortalar alqoritminin üstünlükləri

- Sadədir, başa düşüləndir
- Mürəkkəbliyi aşağıdır
  - *Mürəkkəblik O*(*nkt*), burada *t* - iterasiyaların sayıdır

Azərbaycan TEXNİKİ Universiteti

k-ortalar alqoritminin nöqsanları

- Kvadratik meyllərin cəminin qlobal minimumunun tapılmasına zəmanət vermir, lokal minimumlardan biri tapılır
- Nəticə klasterlərin başlanğıc mərkəzlərinin seçilməsindən asılıdır, onların optimal seçimi hələ də məlum deyil.
- Klasterlərin sayını əvvəlcədən bilmək lazımdır.

Azərbaycan TEXNİKİ Universiteti

K-ortalar alqoritminin nöqsanlarının aradan qaldırılması

① Bir neçə təsadüfi klasterləşdirmə edilir: keyfiyyət funksionalına görə ən yaxşısı seçilir.

② Klasterlərin *k* sayı tədricən artırılır.

Azərbaycan TEXNİKİ Universiteti

Medoidlər metodu (k-medoids)

- **K-medoids** – ədədi ortanın əvəzinə hər bir klasterin medianı istifadə edilir.
  - 1, 3, 5, 7, 9-un ədədi ortası: **5**
  - 1, 3, 5, 7, 1009 -un ədədi ortası: **205**
  - 1, 3, 5, 7, 1009 -un medianı: **5**
  - Medianın üstünlüyü: ekstremal qiymətlərin təsiri yoxdur

Azərbaycan TEXNİKİ Universiteti

Medoidlər metodu (k-medoids)

- Hər bir klasterdə ağırlıq mərkəzlərinin əvəzinə medoid hesablanır.
- Medoid – verilənlər çoxluğunun və ya klasterin elə obyektidir ki, digər obyektlərdən ona qədər olan orta məsafə minimaldır.
- Obyektlərin klasterdən klasterə yerdəyişməsi alqoritmi *k*-ortalar alqoritminə analojidir.
- Medoid – obyektdir, buna görə bütün obyektlər arasındakı məsafələr matrisini bilmək kifayətdir (üstünlük!)

İyerarxik klasterləşdirmə

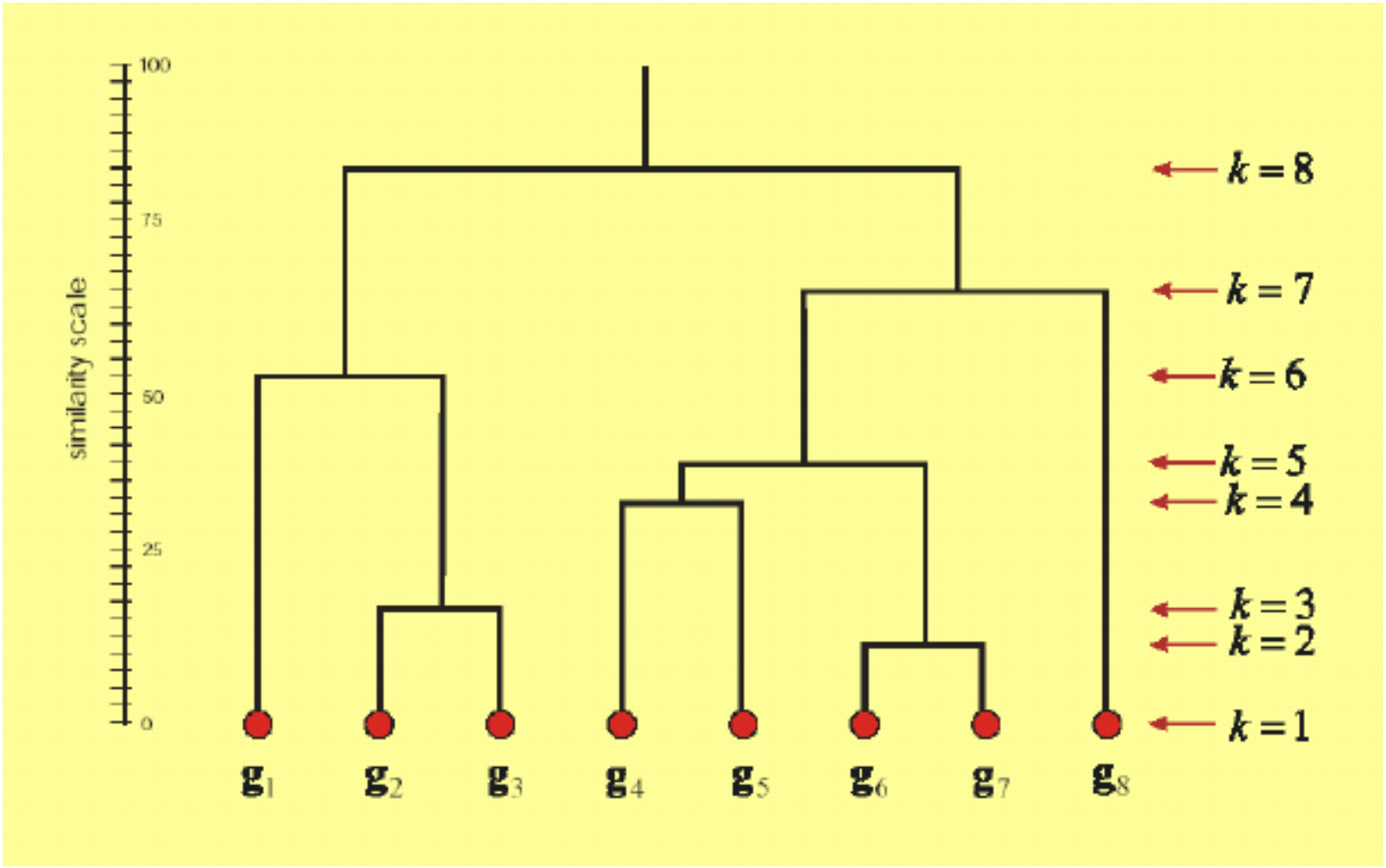
Azərbaycan TEXNİKİ Universiteti

İyerarxik klasterləşdirmə (taksonomiya)

- **Əsas ideya**: aşağı səviyyələrdəki klasterlər yuxarıdakı klasterləri bölməklə alınır.
- Klassifikasiyanın təpəsində bütün obyektləri əhatə edən bir klaster durur.
- Ən aşağı səviyyədə hər birində bir obyekt olan *n* klaster olur.
- Belə iyerarxik sturukturları kök ağacları (**dendroqramlar**) şəklində göstərmək əlverişlidir.

Azərbaycan TEXNİKİ Universiteti

İyerarxik Klasterləşdirmə



Azərbaycan TEXNİKİ Universiteti

İyerarxik klasterləşdirmə alqoritmlərinin üstünlüyü

- İyerarxik klasterləşdirmə alqoritmlərinin girişinə klasterlərin sayını vermək lazım deyil.
- Dendroqramın əsasında istifadəçi özü müəyyən səviyyədə onu kəsərək müəyyən sayda klasterlər ala bilər.

Azərbaycan TEXNİKİ Universiteti

İyerarxik klasterləşdirmə alqoritmlərinin iki sinfi

- İyerarxik klasterləşdirmə alqoritmlərinin iki növü vardır:
  - **Aqqlomerativ (“aşağıdan yuxarıya”)**:
    - Bir elementdən ibarət çoxluqlardan başlayırlar
    - Onlar bütün obyektlər bir klasterdə olana kimi birləşdirilir.
    - Çox geniş yayılmış metoddur.
  - **Diviziv (“yuxarıdan aşağıya”)**:
    - Obyektlər çoxluğu tək elementlərdən ibarət klasterlər alınana kimi rekursiv bölünür.

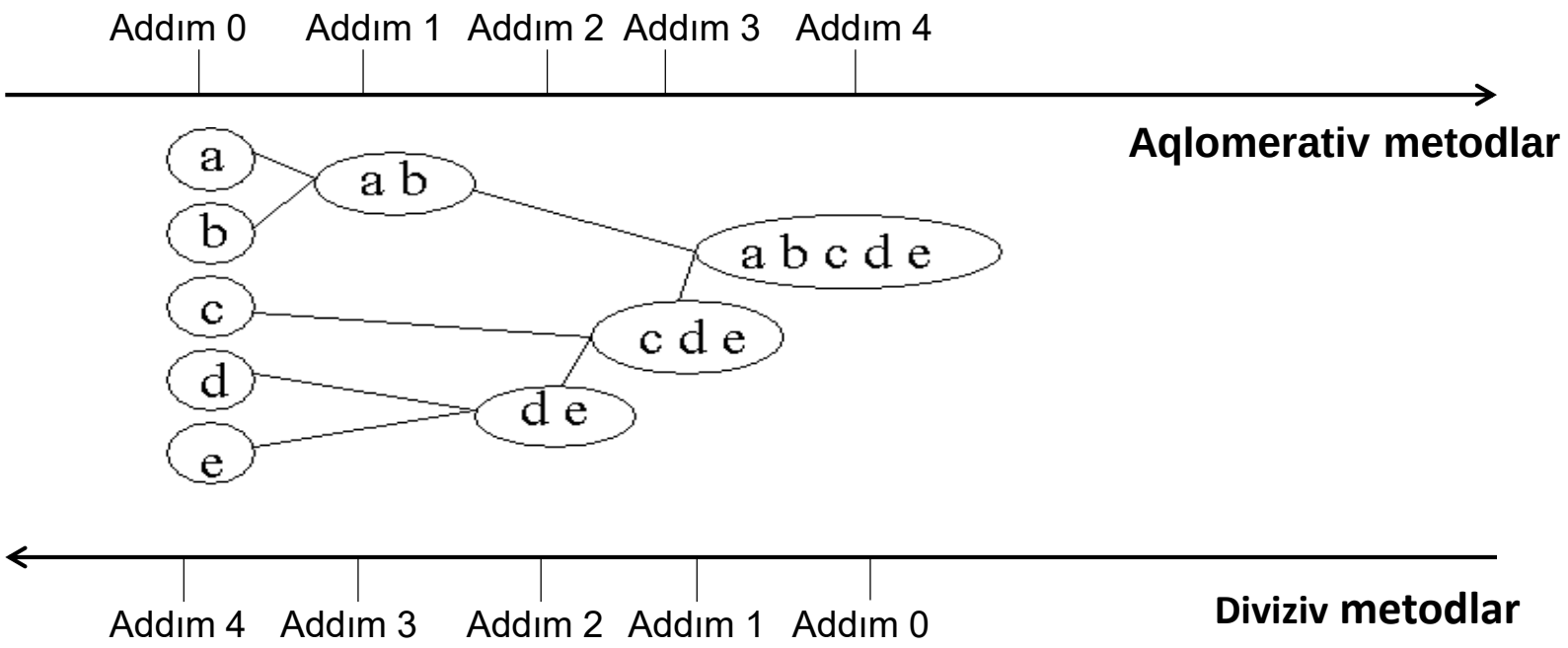
Azərbaycan TEXNİKİ Universiteti

Aqlomerativ metodlar

- Başlanğıc elementlərin ardıcıl birləşdirilməsi və müvafiq olaraq klasterlərin sayının azalması
- Başlanğıcda bütün obyektlər ayrıca klasterlər olur.
- Birinci addımda ən oxşar obyektlər klasterdə birləşdirilir.
- Sonrakı addımlarda birləşdirmə o vaxta kimi davam etdirilir ki, bütün obyektlər bir klaster təşkil etsin.

Azərbaycan TEXNİKİ Universiteti

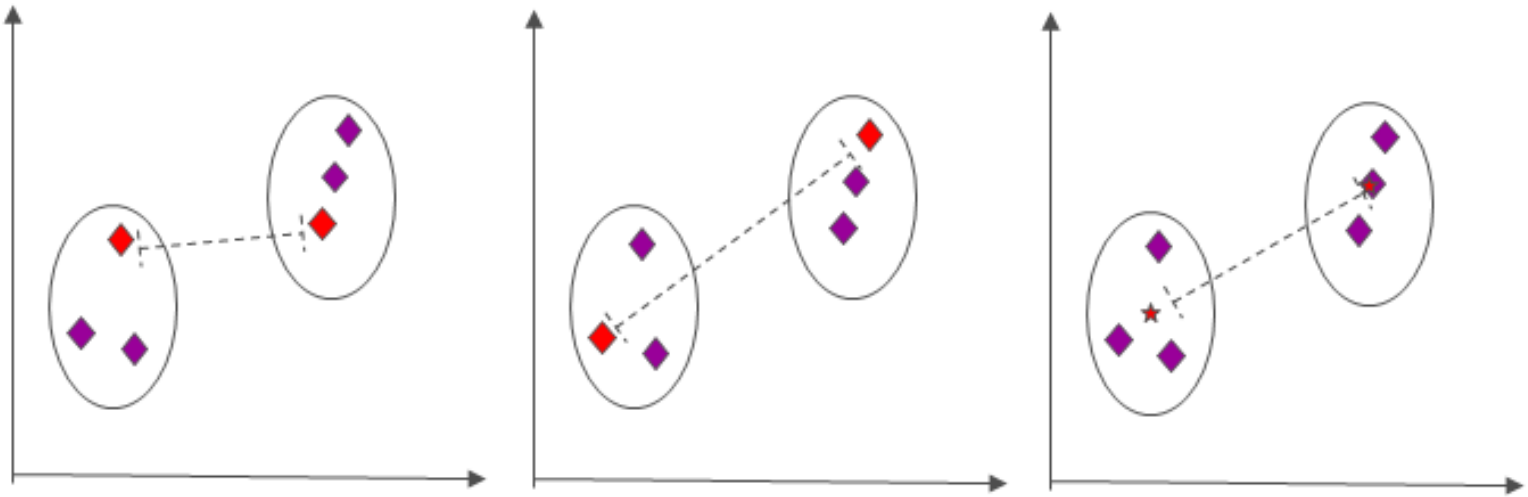
Aqlomerativ metodlar



Azərbaycan TEXNİKİ Universiteti

Klasterlər arasındakı məsafə (1)

- Aqlomerativ metodlarda klasterləri birləşdirmək üçün klasterlər arasında məsafə funksiyası müəyyən edilməlidir.
- Üç yanaşma geniş istifadə edilir.



Ən yaxın qonşu məsafəsi

Uzaq qonşu məsafəsi

Ortalama qrup məsafəsi

Klasterlər arasındakı məsafə (2)

① Ən yaxın qonşu məsafəsi:

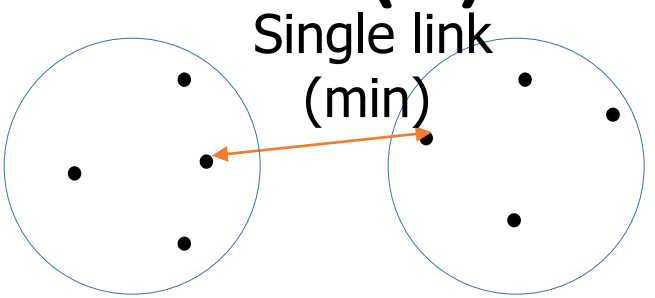
$$D(C_i, C_j) = \min\{\rho(x, z) \mid x \in C_i, z \in C_j\}$$

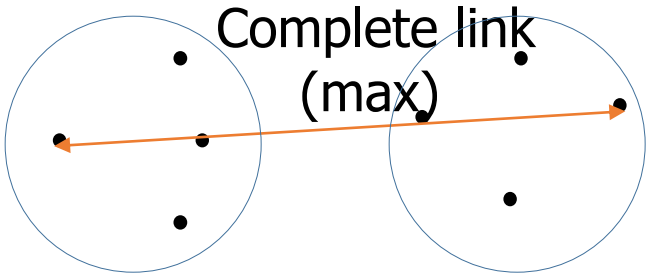
② Uzaq qonşu məsafəsi:

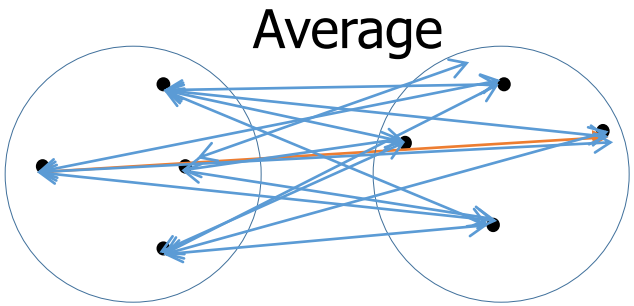
$$D(C_i, C_j) = \max\{\rho(x, z) \mid x \in C_i, z \in C_j\}$$

③ Ortalama qrup məsafəsi:

$$D(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{x \in C_i} \sum_{z \in C_j} \rho(x, z)$$







Azərbaycan TEXNİKİ Universiteti

Aqlomerativ metodlar

- *n* verilənlər obyekti verilib.
- **İyerarxik aqlomerativ klasterləşdirmə** alqoritmi aşağıdakı addımlarla həyata keçirilir:

Addım 1. *n* verilənlər obyekti üçün məsafələr matrisi hesablanır

Addım 2. Hər bir obyekt klaster kimi götürülür

Addım 3. Klasterlərin sayı 1 olana kimi təkrarlanır

Addım 3.1. İki ən yaxın klaster birləşdirilir

Addım 3.2. Məsafə matrisi yenilənir

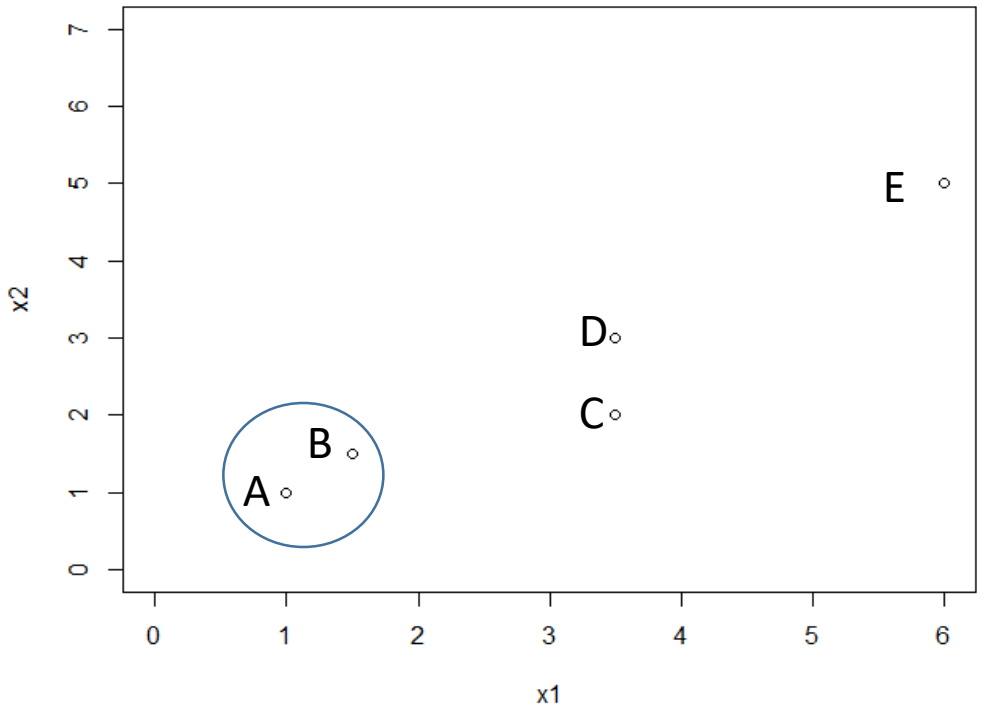
İyerarxik aqlomerativ klasterləşdirmə – Misal

Misal. 5 obyekt verilib:

	x1	x2
A	1.0	1.0
B	1.5	1.5
C	3.5	2.0
D	3.5	3.0
E	6.0	5.0

Məsafə Matrisi

	A	*B*	*C*	*D*	*E*
A	0	0.71	2.69	3.20	6.4
B	0.71	0	2.06	2.6	5.7
C	2.69	3.2	0	1	3.9
D	3.20	2.6	1	0	3.2
E	6.4	5.7	3.9	3.2	0



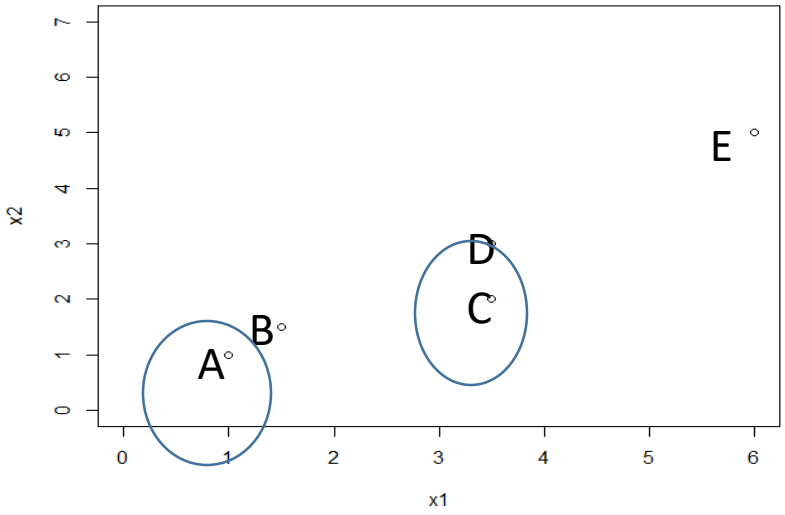
İyerarxik aqlomerativ klasterləşdirmə – Misal

Single Link(age) funksiyasından istifadə edərək, məsafə matrisini yeniləyirik.

	A	B	C	D	E
A	0	0.71	2.69	3.20	6.4
B	0.71	0	2.06	2.6	5.7
C	2.69	3.2	0	1	3.9
D	3.20	2.6	1	0	3.2
E	6.4	5.7	3.9	3.2	0

$$d((A,B),C) = \min(d(A,C),d(B,C)) = \min(2.69,2.06) = 2.06$$
$$d((A,B),D) = \min(d(A,D),d(B,D)) = \min(3.20,2.6) = 2.6$$
$$d((A,B),E) = \min(d(A,E),d(B,E)) = \min(6.4,5.7) = 5.7$$

	(A,B)	C	D	E
(A,B)	0	2.06	2.6	5.7
C	2.06	0	1	3.9
D	2.6	1	0	3.2
E	5.7	3.9	3.2	0



İyerarxik Aqlomerativ Klasterləşdirmə – Misal

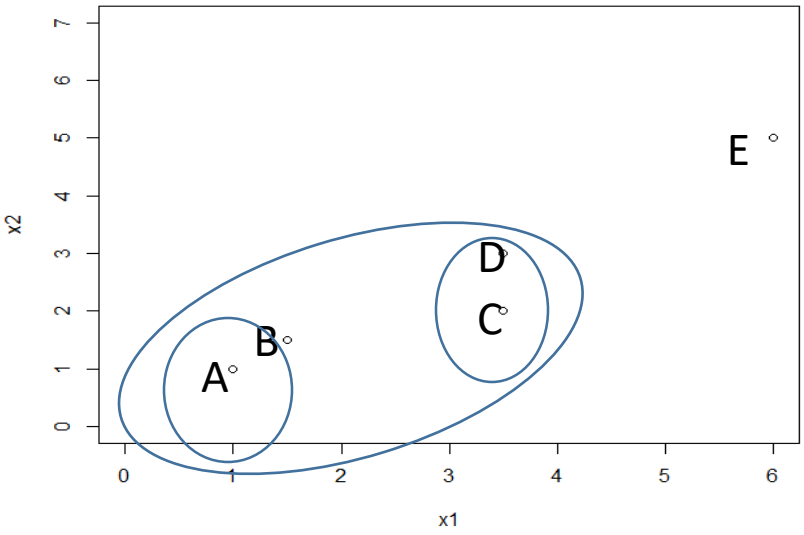
Single Linkage funksiyasından istifadə edərək, məsafə matrisini yeniləyirik.

	(A,B)	C	D	E
(A,B)	0	2.06	2.6	5.7
C	2.06	0	1	3.9
D	2.6	1	0	3.2
E	5.7	3.9	3.2	0

$$d((C,D),(A,B)) = \min(d(C,(A,B)), d(D,(A,B))) = \min(2.06, 2.6) = 2.06$$

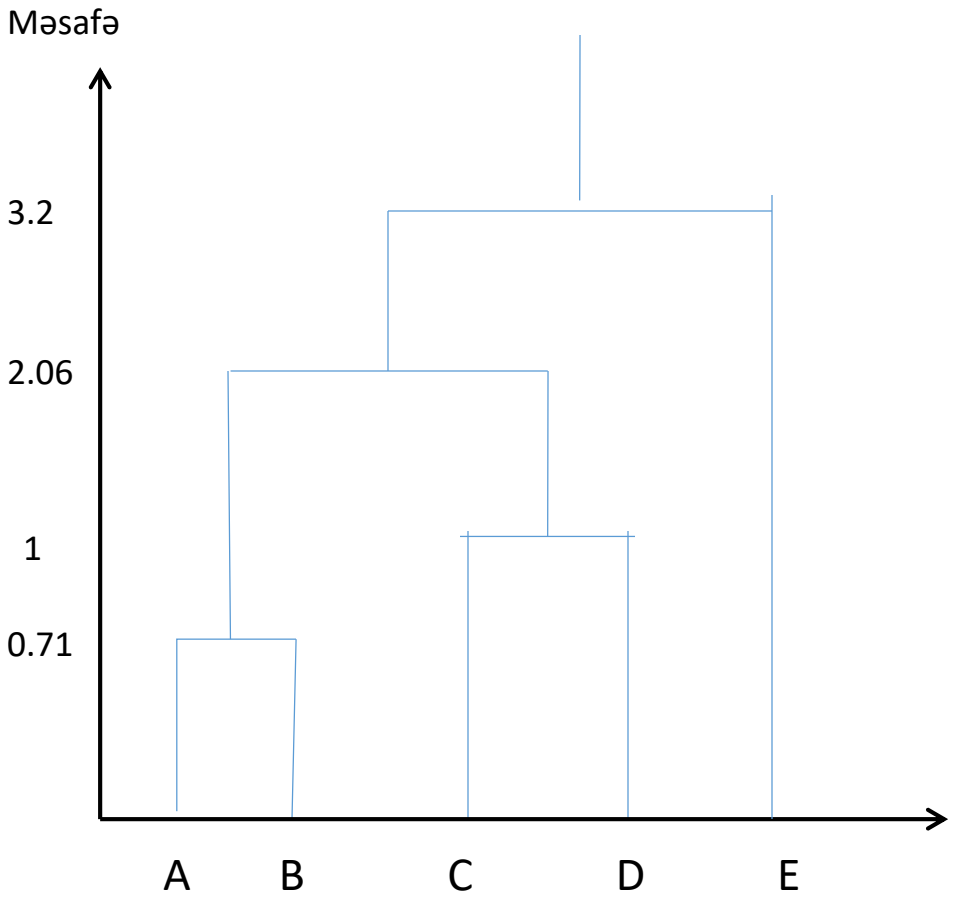
$$d((C,D),E) = \min(d(C,E), d(D,E)) = \min(3.9, 3.2) = 3.2$$

	(A,B)	(C,D)	E
(A,B)	0	2.06	5.7
(C,D)	2.06	0	3.2
E	5.7	3.2	0



İyerarxik aqlomerativ klasterləşdirmə – Misal

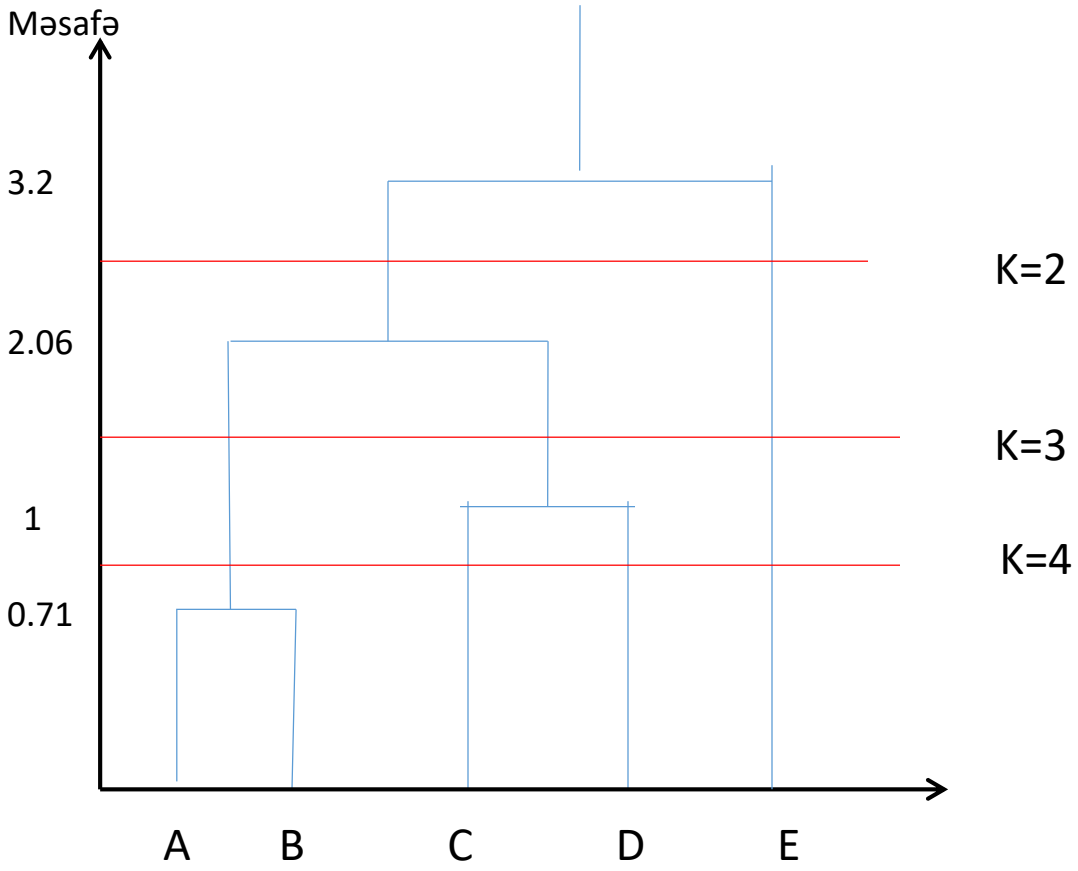
Dendroqram



1. Başlanğıcda 5 klaster var.
2. A və B klasterlərini 0.71 məsafəsində (A, B) klasterində birləşdiririk.
3. C və D klasterlərini 1 məsafəsində (C, D) klasterində birləşdiririk.
4. (A,B) və (C, D) klasterlərini 2.06 məsafəsində ((A, B), (C, D)) klasterində birləşdiririk.
5. ((A, B), (C, D)) və E klasterlərini 3.2 məsafəsində birləşdiririk.
6. Axırıncı klasterə bütün obyektlər daxildir, bununla da hesablamalar dayanır.

İyerarxik aqlomerativ klasterləşdirmə – Misal

- Klasterlərin sayını necə qərarlaşdırırıq? Ağacı haradasa kəsirik.



Azərbaycan TEXNİKİ Universiteti

Diviziv (“yuxarıdan aşağıya”) metodlar

- Ağac kökdən yarpaqlara doğru qurulur.

Yanaşmalardan biri:

① Birinci addımda obyektlər çoxluğuna hər hansı alqoritm(ağırlıq mərkəzləri, medoidləri) tətbiq edərək bu çoxluğu iki klasterə bölürlər

② Sonra alınmış klasterlərdən hər birini bölürlər və s.

Azərbaycan TEXNİKİ Universiteti

Klasterləşdirmənin keyfiyyətinin qiymətləndirilməsi

Azərbaycan TEXNİKİ Universiteti

Klasterləşdirmənin keyfiyyətinin qiymətləndirilməsi

- Vahid (hamılıqla qəbul edilən, bütün hallarda tətbiq edilə bilən) qiymətləndirmə metodu yoxdur.
- Qiymətləndirmə nəzərdə tutur ki, toplu (və ya toplunun bir hissəsi) insan tərəfindən işarələnib.
  - *Klasterlər* – klasterləşdirmənin nəticəsidir.
  - *Siniflər* – insan tərəfindən işarələnmənin nəticəsidir.

Azərbaycan TEXNİKİ Universiteti

Səhvlər matrisi (Confusion matrix)

- n = obyektlərin sayı
- m_i = i klasterində obyektlərin sayı
- c_j = j sinfində obyektlərin sayı
- n_{ij}= i klasterində j sinfindən olan obyektlərin sayı
- $p_{ij} = n_{ij}/m_i$= i klasterində olan obyektin j sinfinə mənsub edilməsi ehtimalı

	Class 1	Class 2	Class 3	
Cluster 1	n_{11}	n_{12}	n_{13}	m_1
Cluster 2	n_{21}	n_{22}	n_{23}	m_2
Cluster 3	n_{31}	n_{32}	n_{33}	m_3
	c_1	c_2	c_3	n

	Class 1	Class 2	Class 3	
Cluster 1	p_{11}	p_{12}	p_{13}	m_1
Cluster 2	p_{21}	p_{22}	p_{23}	m_2
Cluster 3	p_{31}	p_{32}	p_{33}	m_3
	c_1	c_2	c_3	n

Azərbaycan TEXNİKİ Universiteti

Metrikalar

	Class 1	Class 2	Class 3	
Cluster 1	p_{11}	p_{12}	p_{13}	m_1
Cluster 2	p_{21}	p_{22}	p_{23}	m_2
Cluster 3	p_{31}	p_{32}	p_{33}	m_3
	c_1	c_2	c_3	n

- Entropiya:
  - i klasterinin entropiyası: $e_i = -\sum_{j=1}^{L} p_{ij} \log p_{ij}$
    - Müntəzəm paylandıqda ən yüksəkdir, bir sinif olduqda sıfra bərabərdir
  - Bütün klasterlərin entropiyası: $e = \sum_{i=1}^{K} \frac{m_i}{n} e_i$
- Saflıq (ing. Purity):
  - i klasterinin saflığı: $p_i = \max_j p_{ij}$
  - Bütün klasterlərin saflığı: $p(C) = \sum_{i=1}^{K} \frac{m_i}{n} p_i$

Azərbaycan Texniki Universiteti

Metrikalar

	Class 1	Class 2	Class 3	
Cluster 1	p_{11}	p_{12}	p_{13}	m_1
Cluster 2	p_{21}	p_{22}	p_{23}	m_2
Cluster 3	p_{31}	p_{32}	p_{33}	m_3
	c_1	c_2	c_3	n

- Precision:
  - i klasterinin j sinfinə nəzərən: $Prec(i,j) = p_{ij}$
- Recall:
  - i klasterinin j sinfinə nəzərən: $Rec(i,j) = \frac{n_{ij}}{c_j}$
- F-measure:
  - Precision və Recall-un Harmonik ortasıdır:

$$F(i,j) = \frac{2 * Prec(i,j) * Rec(i,j)}{Prec(i,j) + Rec(i,j)}$$

Azərbaycan TEXNİKİ Universiteti

Metrikalar

	Class 1	Class 2	Class 3	
Cluster 1	n_{11}	n_{12}	n_{13}	m_1
Cluster 2	n_{21}	n_{22}	n_{23}	m_2
Cluster 3	n_{31}	n_{32}	n_{33}	m_3
	c_1	c_2	c_3	n

Klasterlər və klasterləşdirmə üçün Precision/Recall

- i klasterinə elə k_i sinfi təyin olunur ki, $k_i = \arg\max_j n_{ij}$
- Precision:
  - i klasteri: $Prec(i) = \frac{n_{ik_i}}{m_i}$
  - Bütün klasterlər üçün: $Prec(C) = \sum_i \frac{m_i}{n} Prec(i)$
- Recall:
  - i klasteri: $Rec(i) = \frac{n_{ik_i}}{c_{k_i}}$
  - Bütün klasterlər üçün: $Rec(C) = \sum_i \frac{m_i}{n} Rec(i)$
- F-measure:
  - Precision və Recall-un Harmonik ortası

Azərbaycan TEXNİKİ Universiteti

Metrikalar barəsində bəzi qeydlər

Entropiya – sinfin klasterlər üzrə «yayılmasıdır». Nə qədər kiçikdirsə, bir o qədər yaxşıdır, ideal halda *Entropy*=0.

Saflığın qiyməti nə qədər böyükdürsə, bir o qədər yaxşıdır. İdeal halda, *Purity*=1.

F-ölçü klasterləşdirmənin ümumi keyfiyyətini göstərir, lakin klasterlərin özlərinin quruluşunu göstərmir.

Azərbaycan TEXNİKİ Universiteti

Yaxşı və pis klasterləşdirmə: Misal

	Class 1	Class 2	Class 3	
Cluster 1	2	3	85	90
Cluster 2	90	12	8	110
Cluster 3	8	85	7	100
	100	100	100	300

Purity: (0.94, 0.81, 0.85)
– Bütün klasterlər üçün: 0.86
Precision: (0.94, 0.81, 0.85)
– Bütün klasterlər üçün: 0.86
Recall: (0.85, 0.9, 0.85)
- Bütün klasterlər üçün: 0.87

	Class 1	Class 2	Class 3	
Cluster 1	20	35	35	90
Cluster 2	30	42	38	110
Cluster 3	38	35	27	100
	100	100	100	300

Purity: (0.38, 0.38, 0.38)
– Bütün klasterlər üçün: 0.38
Precision: (0.38, 0.38, 0.38)
– Bütün klasterlər üçün: 0.38
Recall: (0.35, 0.42, 0.38)
– Bütün klasterlər üçün: 0.39