Büyük Veri İçin İstatistiksel Öğrenme (Statistical Learning for Big Data)

M. Ali Akcayol
Gazi Üniversitesi
Bilgisayar Mühendisliği Bölümü

Bu dersin sunumları, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Trevor Hastie, Robert Tibshirani, Jerome Friedman, Springer, 2017." ve "Mining of Massive Datasets, Jure Leskovec, Anand Rajaraman, Jeffrey David Ullman, Stanford University, 2011." kitapları kullanılarak hazırlanmıştır.

Konular

- Denetimsiz Öğrenmenin Temelleri
- Kümeleme
- Uzaklık Ölçütleri
- K-means Algoritması
- Kümelerin Gösterimi
- Hiyerarşik Kümeleme
- Kümeleme Değerlendirmesi

Denetimsiz Öğrenmenin Temelleri

- **Denetimli öğrenme giriş verileri ile çıkış niteliği arasındaki ilişkiyi ortaya çıkartır.**
- Elde edilen model ile yeni verilerle ileriye dönük tahmin yapılması amaçlanmaktadır.
- **Denetimsiz öğrenmede eğitim sürecinde hedef nitelik bulunmamaktadır.**
- **Denetimsiz öğrenmede veriler arasında bazı yapısal ilişkilerin veya örüntülerin ortaya çıkartılması amaçlanmaktadır.**

3

Konular

- Denetimsiz Öğrenmenin Temelleri
- Kümeleme
- Uzaklık Ölçütleri
- K-means Algoritması
- Kümelerin Gösterimi
- Hiyerarşik Kümeleme
- Kümeleme Değerlendirmesi

4

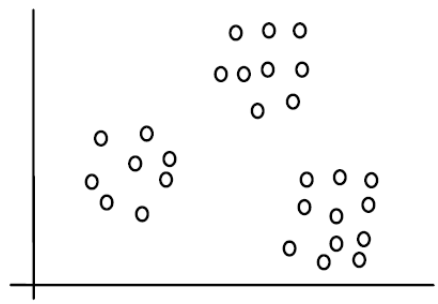
Kümeleme

- **Kümeleme (clustering), veri içerisinde benzer grupların (küme) bulunmasını sağlayan teknikleri kullanır.**
- Kümelemede veri içerisindeki benzer örneklerin yakınlıklarına göre kümeler oluşturulur.
- Birbirine belirlenmiş bir seviyeden daha uzak olanlar ayrı kümelere atanır.
- Kümeleme, denetimsiz öğrenme (unsupervised learning) olarak adlandırılır.
- **Apriori algoritması ile yapılan birliktelik kural madenciliği unsupervised learning olarak nitelendirilir.**

5

Kümeleme

- Aşağıdaki veri kümesinde **uzaklıklara göre üç küme** görülmektedir.
- Bu şekilde yapılan kümelemeye **partitional clustering** denilir.
- Farklı özellikler gözönüne alınırsa küme sayısı daha fazla veya daha az olabilir.



- Sağlık, psikoloji, tarım, sosyoloji, biyoloji, arkeoloji, pazarlama, sigortacılık, kütüphane gibi çok farklı alanlarda kullanılmaktadır.

6

Kümeleme

Örnek

- Her gün haber ajansları tarafından Dünya genelinde çok sayıda haber metni oluşturulur.
- Bu haberlerin ilgili oldukları konulara göre sınıflandırılması gerekir.
- Bu kadar çok sayıdaki haber metninin manuel sınıflandırılması mümkün değildir.
- Sınıflandırılmadan tüm kullanıcılara sunulması da kullanıcıların ilgili olduklarını seçmeleri zor olacağından uygun değildir.
- **Dokümanların** konulara göre **kümelenmesi için clustering algoritmaları kullanılabilir.**
- Bu şekilde sınıflandırmaya **hiyerarşik kümeleme** denilmektedir.
- **Kümeleme algoritmalarının temelinde uzaklık ölçümü yer alır.**

7

Konular

- Denetimsiz Öğrenmenin Temelleri
- Kümeleme
- Uzaklık Ölçütleri
- K-means Algoritması
- Kümelerin Gösterimi
- Hiyerarşik Kümeleme
- Kümeleme Değerlendirmesi

8

Uzaklık Ölçütleri

- Kümeleme problemlerinde **problemin yapısına** ve niteliklerin değerlerine **bağlı olarak farklı uzaklık ölçütleri kullanılabilir.**
- Yaygın kullanılan uzaklık ölçütleri:
  - Öklid uzaklığı
  - Manhattan uzaklığı
  - Minkowski uzaklığı

9

Uzaklık Ölçütleri

Öklid uzaklığı

- **N boyutlu öklit uzayında bir nokta** reel sayılardan oluşan **n elemanlı bir vektördür.**
- Bu uzaydaki **L_2-norm** uzaklık aşağıdaki gibidir:

$$d([x_1, x_2, \ldots, x_n],\ [y_1, y_2, \ldots, y_n]) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

- **L_r-norm** uzaklık aşağıdaki gibidir:

$$d([x_1, x_2, \ldots, x_n],\ [y_1, y_2, \ldots, y_n]) = (\sum_{i=1}^{n}|x_i - y_i|^r)^{1/r}$$

10

Uzaklık Ölçütleri

Manhattan Uzaklığı

- **L_1-norm** uzaklık **Manhattan uzaklığı** olarak adlandırılır.

$$d([x_1, x_2, \ldots x_n], [y_1, y_2, \ldots y_n]) = \left(\sum_{i=1}^{n} |x_i - y_i| \right)$$

- **L_∞-norm** ise *r* sonsuza giderken limiti gösterir.

$$d([x_1, x_2, \ldots x_n], [y_1, y_2, \ldots y_n]) = \lim_{r \to \infty} \left(\sum_{i=1}^{n} |x_i - y_i|^r \right)^{1/r} = \max_{i=1}^{n} (|x_i - y_i|)$$

- **Örnek:** x=(2, 7) ve y=(6, 4) noktaları için aşağıdaki uzaklıklar hesaplanır.

$$L_1 - norm = |2-6| + |7-4| = 7$$

$$L_2 - norm = \sqrt{(2-6)^2 + (7-4)^2} = 5$$

$$L_\infty - norm = \max(|2-6| + |7-4|) = 4$$

11

Uzaklık Ölçütleri

Minkowski Uzaklığı

- **L_r-norm** uzaklık **Minkowski uzaklığı** olarak adlandırılır.
- Minkowski uzaklığı aşağıdaki şekilde hesaplanır:

$$d([x_1, x_2, \ldots x_n], [y_1, y_2, \ldots y_n]) = \left(\sum_{i=1}^{n} |x_i - y_i|^r \right)^{1/r}$$

- Burada, ***r* = 2** için **Öklid uzaklık** bağıntısı ve ***r* = 1** için **Manhattan uzaklık** bağıntısı elde edilir.

12

Örnek

- Aşağıdaki tabloda 5 gözlem değeri için 3 niteliğin değerleri görülmektedir.
- Gözlem değerleri arasındaki hesaplanan uzaklıklar, farklı ölçüm yöntemlerinde farklı olmaktadır.

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

13

Örnek

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 uzaklıkları

Gözlem	1	2	3	4	5
1	0,00				
2	3,46	0,00			
3	5,39	6,08	0,00		
4	5,48	7,07	1,73	0,00	
5	7,28	8,31	3,46	3,32	0,00

Manhattan uzaklıkları

Gözlem	1	2	3	4	5
1	0,00				
2	6,00	0,00			
3	9,00	7,00	6,00		
4	8,00	8,00	3,00	0,00	
5	11,00	11,00	6,00	5,00	0,00

Minkowski uzaklıkları
$r = 3$

Gözlem	1	2	3	4	5
1	0,00				
2	2,88	0,00			
3	4,63	6,01	0,00		
4	5,12	7,01	1,44	0,00	
5	6,55	8,05	2,88	3,07	0,00

14

Konular

- Denetimsiz Öğrenmenin Temelleri
- Kümeleme
- Uzaklık Ölçütleri
- K-means Algoritması
- Kümelerin Gösterimi
- Hiyerarşik Kümeleme
- Kümeleme Değerlendirmesi

15

K-means algoritması

- **Kümeleme algoritmalarının kalitesinin ölçümü için iki kriter vardır:**
  - Inter-cluster uzaklık (maksimize edilir.)
  - Intra-cluster uzaklık (minimize edilir.)
- **Kümelerin arasında mesafe olabildiği kadar fazla olmalıdır.**
- Kümelerin içindeki **elemanlar arasındaki uzaklık olabildiği kadar az olmalıdır.**
- Uzaklık ölçüm yöntemi her problem için ayrı tanımlanabilir ve uygun olanın seçilmesi gereklidir.
- K-means algoritması **partitional clustering** yapmaktadır.

16

K-means algoritması

- K-means algoritması başlangıçta k değeri kadar küme oluşturur.
- **Her küme bir merkez noktasına (centroid) sahiptir.**
- Kümeye ait elemanların tümü, kümenin orta noktasına diğer kümelerin orta noktalarına göre daha yakındır.
- **Algoritma başlangıçta rastgele k adet veri noktasını küme merkezleri olarak seçer.**
- Her merkez noktaya yakın noktalar bu kümeye ait olarak alınırlar.
- **Tüm kümelerin merkez noktaları tekrar hesaplanır.**
- Yeni merkez noktalara göre yeniden küme elemanları belirlenir.
- **Kümelerarası eleman değişimi olmayıncaya veya merkez noktalarda değişim olmayıncaya kadar işlemler devam eder.**

17

K-means algoritması

- Veri kümesi $\mathrm{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ olsun. Her bir $\mathbf{x}$ noktası ise,

  $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ir})$ şeklinde tanımlanan bir reel sayılar vektörüdür.

  $X \subseteq R^r$ ve r nitelik sayısıdır.
- Algoritma kümelerdeki hataların karelerinin toplamını (sum of squared error) minimize etmeye çalışır.

$$SSE = \sum_{j=1}^{k} \sum_{\mathbf{x} \in C_j} dist(\mathbf{x}, \mathbf{m}_j)^2$$

- Burada, $\boldsymbol{k}$ küme sayısını, C_j j.kümeyi, $\mathbf{x}$ C_j kümesine ait nitelikler kümesini, $\boldsymbol{m_j}$ j.kümenin orta noktasını gösterir.
- $dist(\mathbf{x}, \mathbf{m}_j)$ kümenin orta noktasına $\mathbf{x}$ noktalarının uzaklığıdır.

18

K-means algoritması

- Kümelerin orta noktası ise aşağıdaki gibi hesaplanır.

$$\mathbf{m}_j = \frac{1}{|C_j|}\sum_{\mathbf{x}_i \in C_j} \mathbf{x}_i$$

- Burada, bir kümeye ait olan tüm $\mathbf{x}$ noktalarının **nitelik değerlerinin ortalamaları hesaplanır.**
- $|C_j|$ kümeye ait **nokta sayısını ifade eder.**
- Kümeye ait noktaların merkez noktaya uzaklıkları ise aşağıdaki gibi hesaplanır. Burada, $\mathbf{m}_j$ j. kümenin orta noktasıdır.

$$dist(\mathbf{x}_i, \mathbf{m}_j) = \| \mathbf{x}_i - \mathbf{m}_j \|$$
$$= \sqrt{(x_{i1} - m_{j1})^2 + (x_{i2} - m_{j2})^2 + \ldots + (x_{ir} - m_{jr})^2}$$

19

K-means algoritması

Algoritma

$$dist(\mathbf{x}_i, \mathbf{m}_j) = \| \mathbf{x}_i - \mathbf{m}_j \|$$

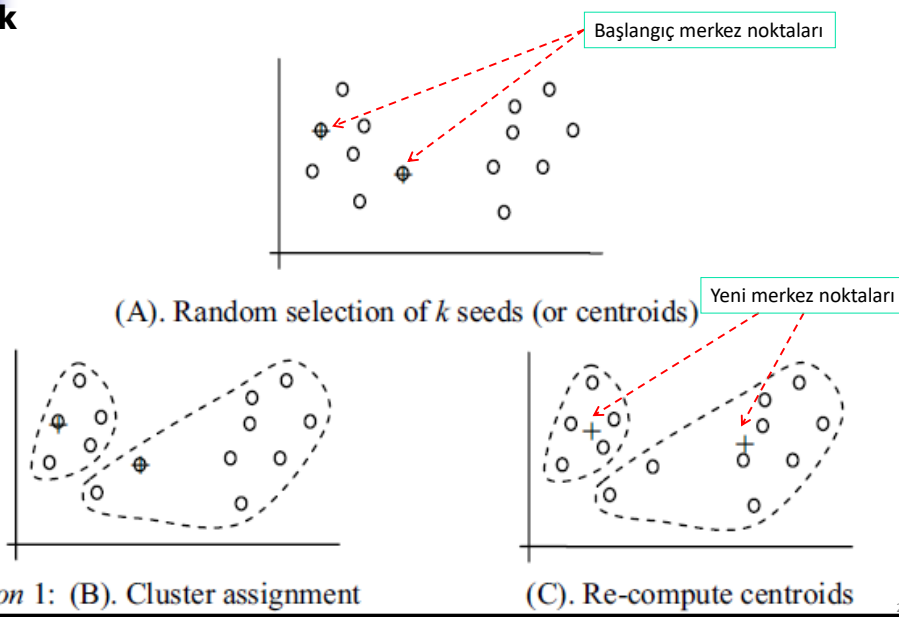
Algorithm *k*-means(*k*, *D*)
1 choose *k* data points as the initial centroids (cluster centers)
2 **repeat**
3 **for** each data point $\mathbf{x} \in D$ do
4 compute the distance from $\mathbf{x}$ to each centroid;
5 assign $\mathbf{x}$ to the closest centroid // a centroid represents a cluster
6 **endfor**
7 re-compute the centroid using the current cluster memberships
8 **until** the stopping criterion is met

$$\mathbf{m}_j = \frac{1}{|C_j|}\sum_{\mathbf{x}_i \in C_j} \mathbf{x}_i$$

20

K-means algoritması

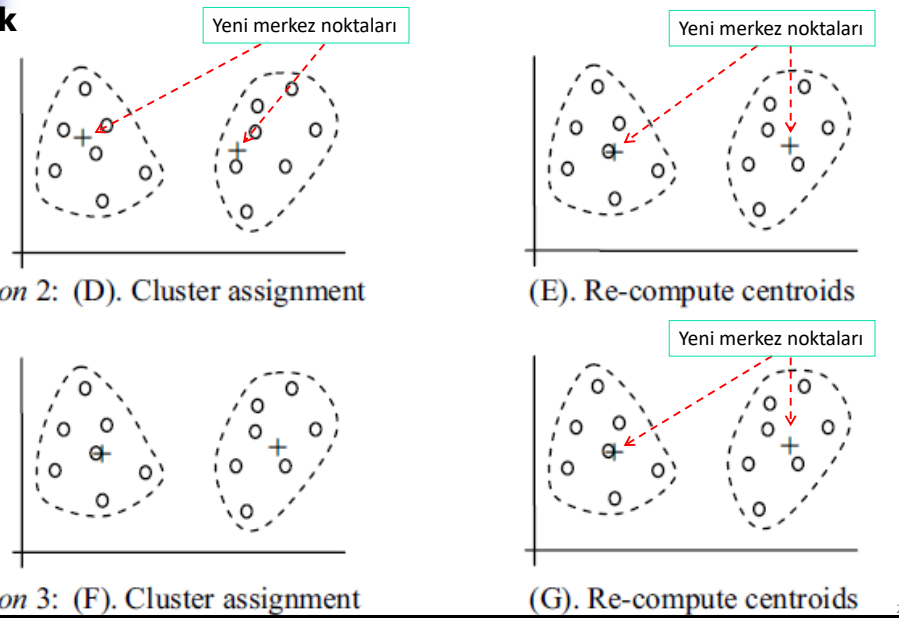
Örnek



21

K-means algoritması

Örnek

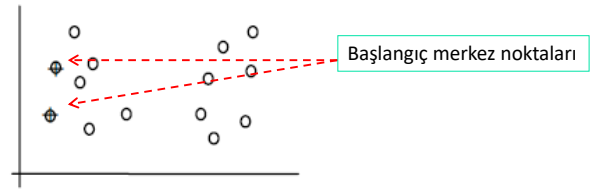


22

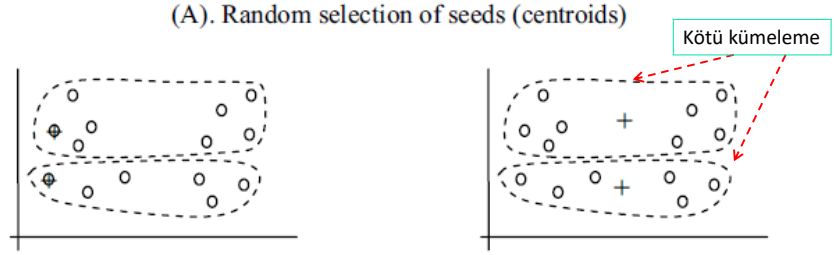
K-means algoritması

K-means algoritmasının zayıf yönleri

- K-means algoritması başlangıç merkez noktalarına bağlı kümeler oluş



(A). Random selection of seeds (centroids)



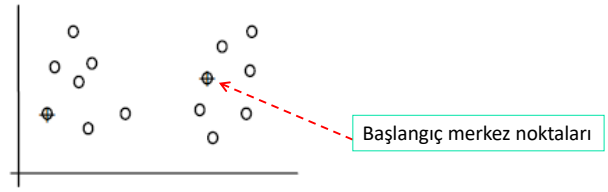
(B). Iteration 1 (C). Iteration 2

23

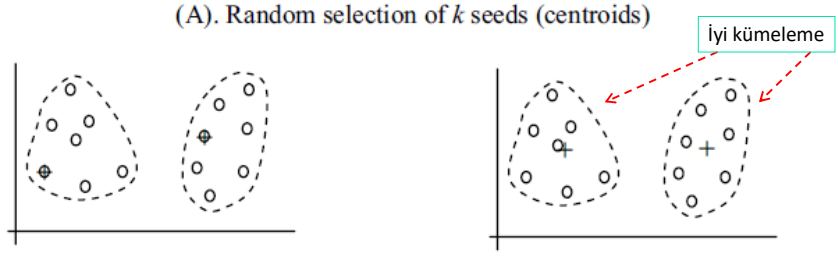
K-means algoritması

K-means algoritmasının zayıf yönleri

- K-means algoritması başlangıç merkez noktalarına bağlı kümeler oluşturur.



(A). Random selection of k seeds (centroids)



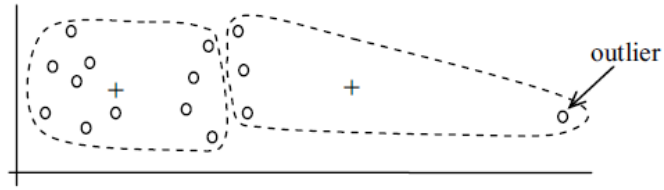
(B). Iteration 1 (C). Iteration 2

24

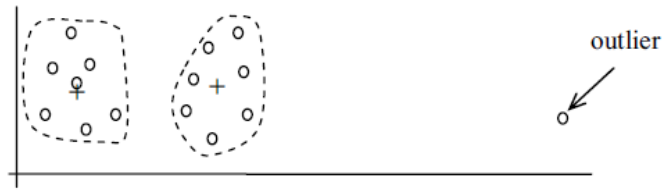
K-means algoritması

K-means algoritmasının zayıf yönleri

- Outlier dataya karşı hassastır.



(A): Undesirable clusters



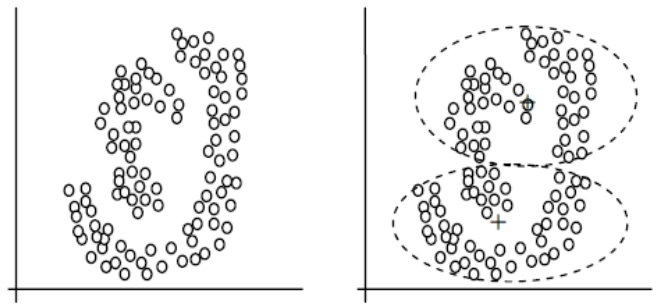
(B): Ideal clusters

25

K-means algoritması

K-means algoritmasının zayıf yönleri

- Bazı durumlarda doğal olarak kümeler oluşmuş durumdadır. Uzaklığa dayalı kümeleme doğal yapıya uygun olmayabilir.
- Bu durumlarda **komşulukları göz önüne alan algoritmalar** kullanılır.



(A): Two natural clusters (B): *k*-means clusters

26

Konular

- Denetimsiz Öğrenmenin Temelleri
- Kümeleme
- Uzaklık Ölçütleri
- K-means Algoritması
- Kümelerin Gösterimi
- Hiyerarşik Kümeleme
- Kümeleme Değerlendirmesi

27

Kümelerin Gösterimi

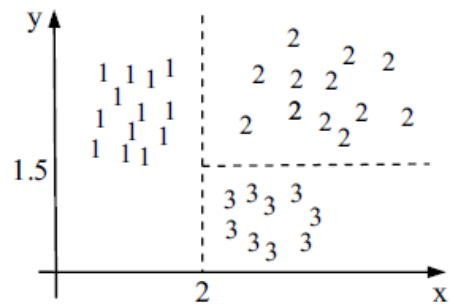
- Bazen kümelerin farklı şekillerde gösterimi gerekebilir.
- **Bazı uygulamalarda sadece kümelerin merkez noktalarının ve yarıçaplarının gösterimi yeterlidir.**
- Dairesel küme yapısına sahip durumlarda faydalıdır ve kümenin yarıçapı kapsadığı alanı gösterir.
- Dairesel olmayan kümeler için merkez ve yarıçap ile gösterim uygun değildir.

28

Kümelerin Gösterimi

- Bazı uygulamalarda sınıflandırma modelleri ile kümeler gösterilebilir.
- **Kümelerin gösterimi karar ağaçları ile yapılabilir.**

$x \leq 2 \rightarrow$ cluster 1
$x > 2, y > 1.5 \rightarrow$ cluster 2
$x > 2, y \leq 1.5 \rightarrow$ cluster 3



29

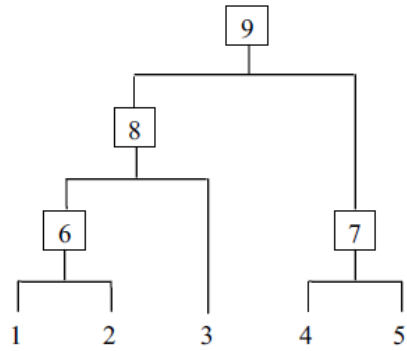
Konular

- Denetimsiz Öğrenmenin Temelleri
- Kümeleme
- Uzaklık Ölçütleri
- K-means Algoritması
- Kümelerin Gösterimi
- Hiyerarşik Kümeleme
- Kümeleme Değerlendirmesi

30

Hiyerarşik Kümeleme

- Hiyerarşik kümeleme diğer bir kümeleme yaklaşımıdır ve ağaç şeklinde gösterilir **(dendrogram).**
- Elemanlar birbirine benzerlik durumuna göre hiyerarşik kümelenir.
- En alt seviyede tek elemanlar bulunur.



31

Hiyerarşik Kümeleme

- Hiyerarşik kümeleme için 2 farklı yöntem kullanılmaktadır.
  - **Agglomerative (bottom up) clustering**

    Öncelikle **en yakın ikili elemanlar ile kümeler oluşturulur.**

    Daha sonra yakın olan kümeler birleştirilerek yeni kümeler oluşturulur.
  - **Divisive (top down) clustering**

    Öncelikle tüm elemanlar tek küme alınır.

    Küme iki parçaya ayrılarak iki küme elde edilir.

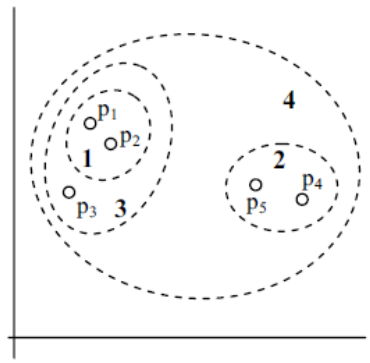
    Elde edilen **kümeler recursive olarak tek elemanlara ulaşıncaya kadar parçalanır.**

    k-means algoritması veya diğer algoritmalar kullanılabilir.

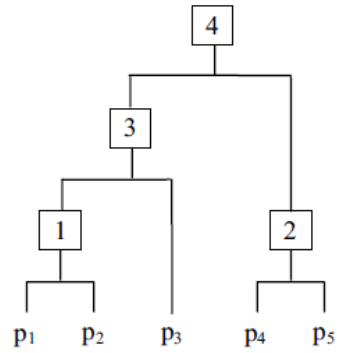
32

Hiyerarşik Kümeleme

Örnek



(A). Nested clusters



(B) Dendrogram

33

Konular

- Denetimsiz Öğrenmenin Temelleri
- Kümeleme
- Uzaklık Ölçütleri
- K-means Algoritması
- Kümelerin Gösterimi
- Hiyerarşik Kümeleme
- Kümeleme Değerlendirmesi

34

Kümeleme Değerlendirmesi

- Kümeleme sonuçlarının değerlendirilmesi için **uygulama alanına göre farklı yöntemler kullanılmaktadır.**
- Bunlardan yaygın kullanılanlar;
  - User inspection
  - Ground truth
  - Entropy
  - Purity
  - Indirect evaluation

35

Kümeleme Değerlendirmesi

User inspection

- Bir grup **uzman tarafından yapılan skorlama** ile değerlendirme yapılır.
- Değerlendirme kişisel olduğundan **tüm skorların ortalaması alınır.**
- **Değerlendirme süreci uzun süre alabilir.**
- **Metin sınıflandırma gibi uygulamalarda faydalı olabilir.**
- Ancak, milyonlarca veriye sahip bir VTYS üzerinde **kümelemenin değerlendirilmesi uzun zaman alır** ve doğru değerlendirme yapılamayabilir.

36

Kümeleme Değerlendirmesi

Ground truth

- **Verilerin küme sayısı belirli ise elde edilen sonuç ona göre değerlendirilir.**
- Her küme içerisinde doğru atanmış elemanlara göre de değerlendirme yapılabilir.

37

Kümeleme Değerlendirmesi

Entropy

- Her küme için entropi hesaplanır. Kümedeki farklı etiketlerin olasılıkları alınır.

$$entropy(D_i) = -\sum_{j=1}^{k} \Pr_i(c_j) \log_2 \Pr_i(c_j)$$

- Burada, D_i i. küme, $Pr_i(c_j)$ j. sınıf etiketinin olasılığıdır.
- Tüm kümeler için entropi hesaplanır.

$$entropy_{total}(D) = \sum_{i=1}^{k} \frac{|D_i|}{|D|} \times entropy(D_i)$$

- $|D_i|$ i. kümedeki eleman sayısıdır. $|D|$ toplam eleman sayısıdır.

38

Kümeleme Değerlendirmesi

Purity

- Her küme için purity hesaplanır.

$$purity(D_i) = \max_j(\Pr_i(c_j))$$

- Burada, D_i i. küme, $Pr_i(c_j)$ j. küme etiketinin olasılığıdır.
- Tüm kümeler için purity hesaplanır.

$$purity_{total}(D) = \sum_{i=1}^{k} \frac{|D_i|}{|D|} \times purity(D_i)$$

- $|D_i|$ i. kümedeki eleman sayısıdır. $|D|$ toplam eleman sayısıdır.

39

Kümeleme Değerlendirmesi

Örnek

$$E = -(\frac{250}{280}\log\frac{250}{280} + \frac{20}{280}\log\frac{20}{280} + \frac{10}{280}\log\frac{10}{280}) = 0{,}589$$

- D kümesi 900 dokümana sahiptir. Tüm dokümanlar Science, Sports ve Politics olarak 3 konuya ayrılmaktadır.
- Her konu 300 dokümana sahiptir.

False Pozitif

$$E = \frac{250}{280} = 0{,}893$$

Cluster	Science	Sports	Politics	Entropy	Purity
Tahmin Science 1	250	20	10	0.589	0.893
Tahmin Sports 2	20	180	80	1.198	0.643
Tahmin Politics 3	30	100	210	1.257	0.617
Total	300	300	300	1.031	0.711

False Negatif

True Pozitif

$$E = -(\frac{280}{900}*0{,}589 + \frac{280}{900}*1{,}198 + \frac{340}{900}*1{,}257) = 1{,}031$$

$$P = -(\frac{280}{900}*0{,}893 + \frac{280}{900}*0{,}643 + \frac{340}{900}*0{,}617) = 0{,}711$$

- Precision, recall ve f-skor değerleri de hesaplanabilir.

40

Kümeleme Değerlendirmesi

Indirect evaluation

- Bazı uygulamalarda oluşturulan kümeler yerine başka parametreler kullanılarak değerlendirme yapılabilir.
- Bir kitap tavsiye sisteminde **müşteriler profil bilgilerine ve geçmişte ilgilendikleri ürünlere göre kümelenebilir.**
- Ancak, **değerlendirme tavsiye edilen kitapların seçilme oranına göre yapılır.**

41