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

- Clustering Yöntemleri
  - Noktalar, uzaylar, uzaklıklar
  - Kümeleme stratejileri
- Öklit uzayında hiyerarşik clustering
- Öklit olmayan uzaylarda hiyerarşik clustering
- K-means algoritmasında k değerinin belirlenmesi

Clustering yöntemleri

- **Clustering noktalar topluluğunun** bir uzaklık ölçütüne göre **gruplandırılmasıdır.**
- Aynı cluster içerisinde yer alan noktalar diğer cluster'lar içerisinde yer alan noktalara göre daha yakındır.

3

Konular

- Clustering Yöntemleri
  - Noktalar, uzaylar, uzaklıklar
  - Kümeleme stratejileri
- Öklit uzayında hiyerarşik clustering
- Öklit olmayan uzaylarda hiyerarşik clustering
- K-Means Algoritmasında k Değerinin Belirlenmesi

4

Noktalar, uzaylar, uzaklıklar

- Clustering için iki temel yaklaşım vardır: **hiyerarşik** ve **nokta atama**.
- **Bir veri seti noktalar topluluğudur** ve **her nokta uzaydaki bir nesnedir.**
- **Öklit uzayındaki noktalar reel sayılardan oluşan vektör ile gösterilir.**
- Vektör elemanları koordinat olarak adlandırılır.
- Günümüzdeki **clustering problemleri** çok **büyük boyuttadır.**
- Noktalar arasındaki uzaklık ölçütlerinde aşağıdaki şartlar sağlanır:
  - **Noktalar arasındaki uzaklıklar her zaman pozitif olur.**
  - **Uzaklık simetriktir.** Uzaklık hesaplanırken noktaların sırası önemli değildir.
  - **Uzaklık ölçütleri üçgen eşitsizliğine uymalıdır.** $d(x, y) + d(y, z) \geq d(x, z)$

5

Konular

- Clustering Yöntemleri
  - Noktalar, uzaylar, uzaklıklar
  - Kümeleme stratejileri
- Öklit uzayında hiyerarşik clustering
- Öklit olmayan uzaylarda hiyerarşik clustering
- K-Means Algoritmasında k Değerinin Belirlenmesi

6

Kümeleme stratejileri

- **Cluster özeti** için **Öklit uzayında** noktaların orta noktası (centroid) alınır.
- **Öklit dışındaki uzaylarda cluster özeti için farklı yöntemler kullanılır.**
- Kullanılan yöntemlere göre **clustering algoritmaları iki gruba ayrılır:**

(1) Hiyerarşik veya agglomerative

- **Her nokta kendi cluster'ına ait tek nokta** alınarak olarak başlanır.
- Yakınlık durumuna göre **noktalar birleştirilerek cluster'lar oluşturulur.**
- Algoritma **önceden belirlenen cluster sayısına ulaşıldığında** veya **noktalar arasında belirli uzaklığa ulaşıldığında sonlanır.**

(2) Nokta atama

- Başlangıçta **belirli sayıda nokta cluster belirlenir.**
- **Diğer tüm noktalar en iyi eşleştirildiği cluster'a atanır.**
- Outlier noktalar herhangi bir cluster'a atanmayabilir.

7

Konular

- Clustering Yöntemleri
  - Noktalar, uzaylar, uzaklıklar
  - Kümeleme stratejileri
- Öklit uzayında hiyerarşik clustering
- Öklit olmayan uzaylarda hiyerarşik clustering
- K-Means Algoritmasında k Değerinin Belirlenmesi

8

Öklit uzayında hiyerarşik clustering

- Hiyerarşik clustering algoritmalarında **her nokta bir cluster alınarak başlanır** ve **cluster'lar birleştirilir.**
- Öklit uzayında **cluster'ların özetleri için centroid kullanılır.**
- **Öklit olmayan uzaylarda** ise **cluster'ların özeti için** clustroid **kullanılır.**
- **Clustroid bir cluster'ı temsil eder** ve uygulamaya özgü belirlenecek bir yöntemle elde edilebilir.

9

Öklit uzayında hiyerarşik clustering

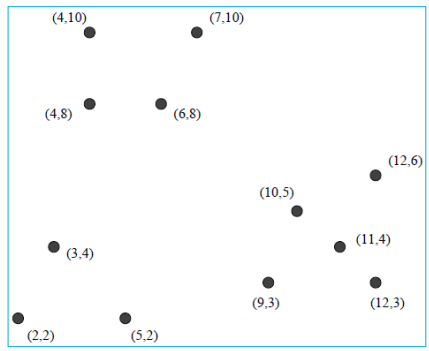
- Tüm **hiyerarşik clustering algoritmaları her noktayı bir cluster alarak başlar.**
- Küçük **iki cluster birleştirilerek** daha **büyük bir cluster oluşturulur.**
- Hiyerarşik clustering algoritmalarında aşağıdakilerin belirlenmesi gerekir:
  - **Cluster'ların nasıl gösterileceği**
  - **İki cluster'ın birleştirilmesinin nasıl yapılacağı**
  - **Cluster birleştirmenin ne zaman sonlanacağı**
- Algoritma aşağıdaki işlem adımlarını tekrarlar.

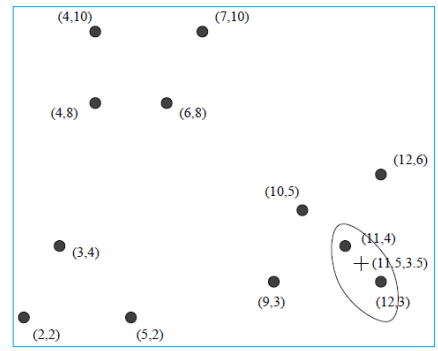
```
WHILE it is not time to stop DO
    pick the best two clusters to merge;
    combine those two clusters into one cluster;
END;
```

10

Öklit uzayında hiyerarşik clustering

- **Örnek**
  - Aşağıdaki veri kümesi iki boyutlu Öklit uzayındadır.
  - **Başlangıçta tüm noktalar** kendi cluster'ına aittir ve **orta noktadır.**
  - En yakın iki nokta çifti **(10, 5) ile (11, 4)** ve **(11, 4) ile (12, 3),** $(d = \sqrt{2})$
  - $(11, 4)$ ile $(12, 3)$ birleştirildiğinde orta nokta $(11.5, 3.5)$ olur.



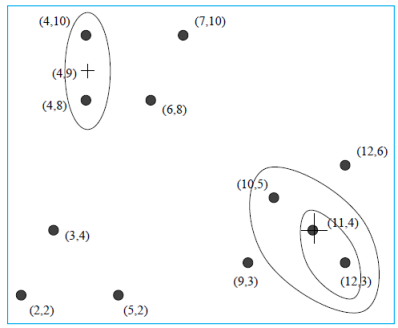


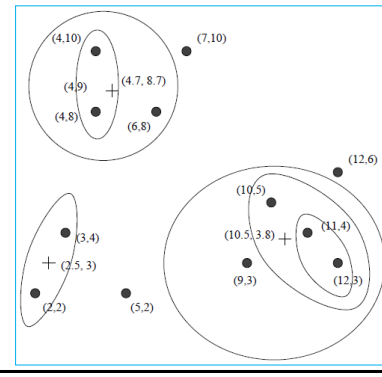
11

Öklit uzayında hiyerarşik clustering

Örnek

- Sonraki adımda **(10, 5) ile centroid arasındaki uzaklık (2.12) olur.**
- **(4, 8) ile (4, 10) arasındaki uzaklık (2.0) olur.** İkinci adımda bu iki nokta birleştirilir ve **centroid (4, 9).**
- Ardından **(10, 5) noktası ile (11.5, 3.5) cluster'ı birleştirilir.**





12

Öklit uzayında hiyerarşik clustering

Cluster birleştirme kuralları

- İki farklı cluster içindeki noktalardan **en yakın olanların uzaklığı minimum olan iki cluster birleştirilir** (En yakın komşu algoritması).
- İki farklı cluster içindeki noktalardan **en uzak olanların uzaklığı minimum olan iki cluster birleştirilir** (En uzak komşu algoritması).
- İki farklı cluster'daki **tüm nokta çiftlerinin birbirine uzaklıklarının ortalaması minimum olan iki cluster birleştirilir.**
- **Bir cluster'ın yarıçapı tüm noktaların centroid'e maksimum uzaklığını belirler.** İki cluster birleştirilirken **minimum yarıçapı oluşturacak iki cluster birleştirilir.**
- **Bir cluster'ın çapı cluster içindeki en uzak iki noktanın uzaklığını belirler.** İki cluster birleştirilirken **minimum çapı oluşturacak iki cluster birleştirilir.**

13

Öklit uzayında hiyerarşik clustering

Cluster birleştirme kısıtları

- **Cluster çapı** belirlenen **threshold değerini aştığında** birleştirme yapılmaz.
- **Cluster içindeki nokta yoğunluğu** belirlenen **threshold değeri aştığında** birleştirme yapılmaz.
- İki cluster birleştirilince **kötü bir cluster oluşacaksa** (Örn.: Cluster çapı aniden çok yükselir) **birleştirme yapılmaz.**

14

Öklit uzayında hiyerarşik clustering

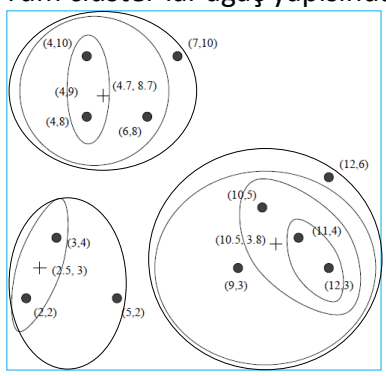
Algoritmayı sonlandırma

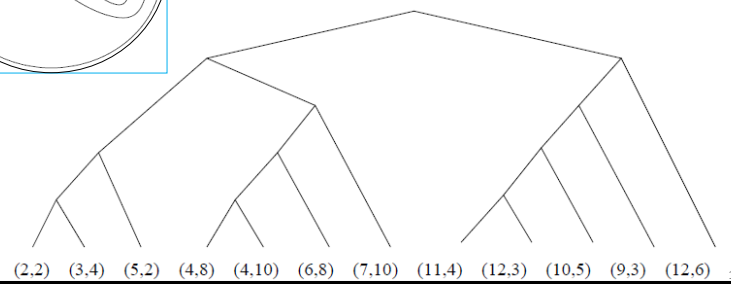
- Önceden **belirlenen sayıda cluster'a ulaşıldığında algoritma sonlandırılabilir.**
- Cluster centroid **noktasından ortalama uzaklık belirli bir threshold değeri aştığında sonlandırılabilir.** Cluster genişliği belirli bir alanda tutulmak istenebilir.
- **Tüm cluster'lar birleştirilip tek cluster elde edilince** algoritma sonlandırılır.

15

Öklit uzayında hiyerarşik clustering

- Tüm cluster'lar ağaç yapısında gösterilebilir.





16

Konular

- Clustering Yöntemleri
  - Noktalar, uzaylar, uzaklıklar
  - Kümeleme stratejileri
- Öklit uzayında hiyerarşik clustering
- Öklit olmayan uzaylarda hiyerarşik clustering
- K-Means Algoritmasında k Değerinin Belirlenmesi

17

Öklit olmayan uzaylarda hiyerarşik clustering

- Öklit olmayan uzaylarda **uzaklık ölçütünün** (Jaccard, edit, …) **belirlenmesi gereklidir.**
- Öklit olmayan uzaylarda **konum bilgisi tanımlanmaz.**
- Öklit olmayan uzaylarda **iki noktanın orta noktası da tanımlanamayabilir.**
- Öklit olmayan uzaylarda **noktalar birleştirilemez** ve **noktalardan birisi cluster'ı temsil eder (clustroid).**
- **Clustroid noktası için,**
  - Cluster içindeki **diğer noktalara uzaklığın toplamı alınabilir.**
  - Cluster içindeki **diğer noktalara maksimum uzaklığı olan nokta alınabilir.**
  - Cluster içindeki **diğer noktalara uzaklıkların karelerinin toplamı alınabilir.**
  - **Tüm noktalara en yakın nokta clustroid alınabilir.**

18

Öklit olmayan uzaylarda hiyerarşik clustering

Örnek

- Öklit olmayan uzaylarda uzaklık ölçütünün belirlenmesi gereklidir.
- Bir cluster **abcd**, **aecdb**, **abecb**, **ecdab** noktalarına sahip olsun.
- Aralarındaki uzaklık **edit distance** ile aşağıdaki gibi hesaplanır.

	ecdab	abecb	aecdb
abcd	5	3	3
aecdb	2	2	
abecb	4		

- Cluster'ın centroid noktası için üç kriter aşağıdaki gibi hesaplanır.

Point	Sum	Max	Sum-Sq
abcd	11	5	43
aecdb	7	3	17
abecb	9	4	29
ecdab	11	5	45

- Üç kritere göre de **aecdb** noktası centroid alınır.

19

Öklit olmayan uzaylarda hiyerarşik clustering

Cluster birleştirme

- Clustroid noktaları **birbirine en yakın olan cluster'lar birleştirilebilir.**
- Cluster'lardaki **tüm noktaların arasındaki uzaklıkların minimum olduğu iki cluster birleştirilebilir.**
- Cluster'lardaki **noktaların uzaklıklarının ortalamalarının minimum olduğu iki cluster birleştirilebilir.**

Birleştirmenin sonlandırılması

- Cluster içerisindeki **nokta yoğunluğuna göre** birleştirme sonlandırılabilir.
- Cluster **yarıçapı** veya **çapına göre** birleştirme sonlandırılabilir.

20

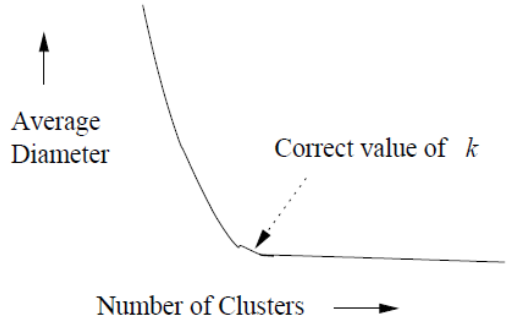
Konular

- Clustering Yöntemleri
  - Noktalar, uzaylar, uzaklıklar
  - Kümeleme stratejileri
- Öklit uzayında hiyerarşik clustering
- Öklit olmayan uzaylarda hiyerarşik clustering
- K-Means Algoritmasında k Değerinin Belirlenmesi

21

K-Means Algoritmasında k Değerinin Belirlenmesi

- Doğru *k* değeri bilinemeyebilir, ancak **farklı *k* değerleri için clustering kalitesi ölçülebilir.**
- Seçilen **cluster sayısı**, **doğru** cluster sayısına **eşit** veya **büyükse cluster yarıçapı** veya **çapı, nokta ekledikçe yavaş bir şekilde artar.**
- Seçilen cluster sayısı, doğru cluster sayısından **küçük** ise **yarıçap** veya **çap aniden yükselir.**



22

K-Means Algoritmasında k Değerinin Belirlenmesi

- Doğru k değerine ilişkin bir bilgi yoksa, ***k* değeri** $1, 2, 4, 8, \ldots$ şeklinde **artırılarak denenir** ve **en uygun *k* değeri belirlenir.**
- **Yarıçap** veya **çap** değeri **hangi aralıkta aniden düşerse** o aralıkta **binary search ile doğru *k* değeri belirlenebilir.**
- k değerinin x ile y arasında olacağı belirlenmiş olsun.
- $z = (x + y) / 2$ değerine bakılır.
- x ile z arasındaki değişim ile z ile y arasındaki değişime bakılır.
- Hangi aralıkta değişim yüksekse o aralıkla devam edilir.

23

Ödev

- Görüntü verileri için kullanılan clustering algoritmalarını içeren bir makale ödevi hazırlayınız.

24