Büyük Veri Analitiği
(Big Data Analytics)

M. Ali Akcayol
Gazi Üniversitesi
Bilgisayar Mühendisliği Bölümü

Bu dersin sunumları, "Mining of Massive Datasets, Jure Leskovec, Anand Rajaraman, Jeffrey David Ullman, Stanford University, 2011." kitabı kullanılarak hazırlanmıştır.

Konular

- Veri Madenciliği
  - İstatistiksel model
  - Makine öğrenmesi
  - Modellemede hesaplamalı yaklaşımlar
  - Özetleme
  - Özellik çıkarımı
- Veri Madenciliğinde İstatistiksel Limitler
  - Toplam bilgi farkındalığı
  - Bonferroni prensibi
- Temel Bilgiler
  - Veri standartlaştırma
  - Dokümanlardaki kelimelerin önemi
  - Hash fonksiyonları
  - İndeksler
  - İkincil depolama birimi

Veri Madenciliği

- **Veri madenciliğinin** en yaygın kabul edilen **tanımı**, **bilgi için model keşfetmek** şeklindedir.
- Bilgi keşfi için oluşturulan **modeller** farklı şekillerde ve **farklı amaçlar için oluşturulabilir.**
- Veriden elde edilmek istenen sonuca göre **model oluşturma süreçleri farklıdır.**
- Oluşturulan **modellerin** istenen amaca uygunluğunun **test edilerek doğrulanması gereklidir.**

3

Konular

- Veri Madenciliği
  - İstatistiksel model
  - Makine öğrenmesi
  - Modellemede hesaplamalı yaklaşımlar
  - Özetleme
  - Özellik çıkarımı
- Veri Madenciliğinde İstatistiksel Limitler
  - Toplam bilgi farkındalığı
  - Bonferroni prensibi
- Temel Bilgiler
  - Veri standartlaştırma
  - Dokümanlardaki kelimelerin önemi
  - Hash fonksiyonları
  - İndeksler
  - İkincil depolama birimi

4

İstatistiksel model

- Veri madenciliği terimini ilk defa istatistikçiler kullanmıştır.
- Veri madenciliği, **veri tarafından doğrudan desteklenmeyen bilginin çıkartılması olarak ifade edilmiştir.**
- İstatistiksel model, veriden elde edilen bir dağılımı ifade eder.
- İstatistikçiler veri madenciliğini istatistiksel model oluşturma olarak görürler.

5

Konular

- Veri Madenciliği
  - İstatistiksel model
  - Makine öğrenmesi
  - Modellemede hesaplamalı yaklaşımlar
  - Özetleme
  - Özellik çıkarımı
- Veri Madenciliğinde İstatistiksel Limitler
  - Toplam bilgi farkındalığı
  - Bonferroni prensibi
- Temel Bilgiler
  - Veri standartlaştırma
  - Dokümanlardaki kelimelerin önemi
  - Hash fonksiyonları
  - İndeksler
  - İkincil depolama birimi

6

Makine öğrenmesi

- **Makine öğrenmesinde**, **veri** bir eğitim kümesi olarak alınır ve **bir algoritmanın öğrenmesi için kullanılır.**
- Makine öğrenmesi, **Bayes ağları**, **destek vektör makinesi**, **yapay sinir ağları**, **karar ağaçları** gibi modelleri kullanır.
- Makine öğrenmesi yöntemleri **çok az bilgi kullanarak** istenen amaca yönelik **sonuçlar oluşturabilir.**

7

Konular

- Veri Madenciliği
  - İstatistiksel model
  - Makine öğrenmesi
  - Modellemede hesaplamalı yaklaşımlar
  - Özetleme
  - Özellik çıkarımı
- Veri Madenciliğinde İstatistiksel Limitler
  - Toplam bilgi farkındalığı
  - Bonferroni prensibi
- Temel Bilgiler
  - Veri standartlaştırma
  - Dokümanlardaki kelimelerin önemi
  - Hash fonksiyonları
  - İndeksler
  - İkincil depolama birimi

8

Modellemede hesaplamalı yaklaşımlar

- **Bilgisayar bilimlerinde**, veri madenciliğine bir **algoritmik problem** olarak bakılır.
- **Verilerden birtakım parametreler elde edilir.**
- Makine öğrenmesi yöntemleri çok az bilgi kullanarak istenen amaca uygun sonuçlar oluşturabilir.
- Veri, **kesin olarak veya yaklaşık olarak özetlenebilir.**
- Verideki bazı **önemli özellikler çıkartılır** diğerleri göz ardı edilir.

9

Konular

- Veri Madenciliği
  - İstatistiksel model
  - Makine öğrenmesi
  - Modellemede hesaplamalı yaklaşımlar
  - Özetleme
  - Özellik çıkarımı
- Veri Madenciliğinde İstatistiksel Limitler
  - Toplam bilgi farkındalığı
  - Bonferroni prensibi
- Temel Bilgiler
  - Veri standartlaştırma
  - Dokümanlardaki kelimelerin önemi
  - Hash fonksiyonları
  - İndeksler
  - İkincil depolama birimi

10

Özetleme

- **Web madenciliğindeki özetleme yöntemlerinde**, Web'in karmaşık yapısı her sayfa için basit verilerle özetlenebilir.
- Kullanıcıların arama yaptıkları **sorgulara göre sayfaların önemi belirlenebilir** (PageRank).
- **Özetlemenin diğer bir uygulama alanı ise öbeklemedir (clustering).**
- Veriler çok boyutlu uzayda birer nokta olarak alınır ve **birbirine yakın olanlar aynı kümeye atanır.**
- Oluşturulan **cluster**, **merkez nokta** veya **başka bir özellik hesaplanarak** elde edilen **özet veri tarafından ifade edilebilir.**

11

Konular

- Veri Madenciliği
  - İstatistiksel model
  - Makine öğrenmesi
  - Modellemede hesaplamalı yaklaşımlar
  - Özetleme
  - Özellik çıkarımı
- Veri Madenciliğinde İstatistiksel Limitler
  - Toplam bilgi farkındalığı
  - Bonferroni prensibi
- Temel Bilgiler
  - Veri standartlaştırma
  - Dokümanlardaki kelimelerin önemi
  - Hash fonksiyonları
  - İndeksler
  - İkincil depolama birimi

12

Özellik çıkarımı

- Büyük ölçekli verideki **elemanlar arasındaki ilişki, aralarındaki bağlantı kullanılarak ifade edilir**.
- **Frequent itemset**, veri içerisindeki elemanların birlikte bulunma oranlarına göre özellik çıkarımı yapar.
- Örneğin, market alışverişinde, belirli oranın üzerinde aynı alışverişte birlikte alınan ürünler.
- **Similar items**, büyük veri kümesi içerisinde birbirine benzeyen elemanları bularak özellik çıkarımı yapar.
- Örneğin, benzer ürün grubuyla ilgilinen kullanıcılar kümesi.

13

Konular

- Veri Madenciliği
  - İstatistiksel model
  - Makine öğrenmesi
  - Modellemede hesaplamalı yaklaşımlar
  - Özetleme
  - Özellik çıkarımı
- Veri Madenciliğinde İstatistiksel Limitler
  - Toplam bilgi farkındalığı
  - Bonferroni prensibi
- Temel Bilgiler
  - Veri standartlaştırma
  - Dokümanlardaki kelimelerin önemi
  - Hash fonksiyonları
  - İndeksler
  - İkincil depolama birimi

14

Toplam bilgi farkındalığı

- 2002 yılında Amerika hükümeti, kredi kartı makbuzları, otel kayıtları, seyahat verileri ve diğer çok farklı türdeki verilerin tamamında veri madenciliği yöntemlerini uygulayarak terörist aktiviteleri izlemeyi planladığını duyurmuştur (Total Information Awareness (TIA) isimli proje).
- Bu proje kongre tarafından gizlilik ve güvenlik nedenlerinden ötürü iptal edilmiştir.
- Bu kadar büyük veri içerisindeki **bazı davranışlar terörist aktivite olmamasına rağmen terörist gibi algılanabilir.**
- Gerçekten **bazı şüpheli davranışların da terörizmle ilgisi olmayabilir.**
- Terörist aktiviteyi tam olarak tanımlayıp ilgili olanların polis tarafından izlenmesi güvenlik, gizlilik ve maliyet açısından gereklidir.

15

Konular

- Veri Madenciliği
  - İstatistiksel model
  - Makine öğrenmesi
  - Modellemede hesaplamalı yaklaşımlar
  - Özetleme
  - Özellik çıkarımı
- Veri Madenciliğinde İstatistiksel Limitler
  - Toplam bilgi farkındalığı
  - Bonferroni prensibi
- Temel Bilgiler
  - Veri standartlaştırma
  - Dokümanlardaki kelimelerin önemi
  - Hash fonksiyonları
  - İndeksler
  - İkincil depolama birimi

16

Bonferroni prensibi

- **Bir veri tamamen rastgele bile olsa aranan olayın olma olasılığı vardır.**
- Verinin boyutu arttıkça aranan bu olayın olma sıklığı da artar.
- Beklenmediği kadar **çok tekrar eden (önemli görünen) bu olay gerçekte önemli olmayabilir.**
- **Bonferroni prensibi, sanki gerçekmiş gibi görünen rastgele tekrar eden bu olayları belirlemeyi sağlar.**
- Eğer bir olayın veri içerisindeki tekrarlanma sayısı, gerçek örneklerden ve beklenenden çok fazla ise sahtedir.
- Örneğin büyük bir veri içerisinde kişilerin belirlenmiş davranışlarına göre terörist sayısı çok az olmalıdır.
- **Bu sayı beklenenden çok fazla çıkarsa elde edilen sonuçlar gerçek dışıdır.**

17

Konular

- Veri Madenciliği
  - İstatistiksel model
  - Makine öğrenmesi
  - Modellemede hesaplamalı yaklaşımlar
  - Özetleme
  - Özellik çıkarımı
- Veri Madenciliğinde İstatistiksel Limitler
  - Toplam bilgi farkındalığı
  - Bonferroni prensibi
- Temel Bilgiler
  - Veri standartlaştırma
  - Dokümanlardaki kelimelerin önemi
  - Hash fonksiyonları
  - İndeksler
  - İkincil depolama birimi

18

Veri Standartlaştırma

- Verilerin standartlaştırılması bazı uygulamalarda gereklidir.
- **Öklid uzaklığına dayalı kümelemede veri standartlaştırma zorunludur.**

Örnek

- İki nitelik değerinden birisi 0-1, diğeri ise 0-1000 aralığında olsun.
- $x_i = (0.9, 720)$ ve $x_j = (0.1, 20)$ ise aralarındaki uzaklık,

$$dist(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(0.9-0.1)^2 + (720-20)^2} = 700.000457$$

olur.

- İki nitelik değerleri de 0-1 aralığında ölçeklenirse,
20 -> 0,02 ve 720 -> 0,72 olur. **Uzaklık değeri 1,063 olur.**

19

Veri Standartlaştırma

Interval-scaled attributes

- Aralık ölçeklendirme yönteminde en yaygın olarak aşağıdaki yöntemler kullanılır:
  - **range (min-max)**
  - **z-score**

20

Veri Standartlaştırma

range (min-max)

- Her nitelik için değerler minimum ve maksimum değerler arasındaki değere göre, 0-1 arasında değer alır.

$$rg(x_{if}) = \frac{x_{if} - \min(f)}{\max(f) - \min(f)}$$

- Burada, $min(f)$ f niteliğinin minimum değerini, $max(f)$ f niteliğinin maksimum değerini ve x_{if} ise i. gözlemin f. nitelik değerini ifade eder.

21

Veri Standartlaştırma

z-score

- Her nitelik için **ortalama değerden uzaklığa** ve nitelik değerlerindeki **standart sapmaya göre** yeni değeri hesaplanır.

$$\sigma_f = \sqrt{\frac{\sum_{i=1}^{n}(x_{if} - \mu_f)^2}{n-1}}$$

$$\mu_f = \frac{1}{n}\sum_{i=1}^{n} x_{if}$$

$$z(x_{if}) = \frac{x_{if} - \mu_f}{\sigma_f}$$

- Burada, σ_f f niteliğinin standart sapması, μ_f f niteliğinin ortalama değeri ve $z(x_{if})$ ise i. gözlemin f. nitelik değerinin yeni değerini ifade eder.

22

Veri Standartlaştırma

Ratio-scaled attributes

- Bazı uygulamalarda nitelik değeri üssel değişebilir.

$$f(t) = Ae^{Bt}$$

- Burada, A ve B katsayılar ve t nitelik değeridir.
- Bu tür durumlarda logaritmik değer ile standartlaştırma yapılır.

$$\log(x_{if})$$

23

Konular

- Veri Madenciliği
  - İstatistiksel model
  - Makine öğrenmesi
  - Modellemede hesaplamalı yaklaşımlar
  - Özetleme
  - Özellik çıkarımı
- Veri Madenciliğinde İstatistiksel Limitler
  - Toplam bilgi farkındalığı
  - Bonferroni prensibi
- Temel Bilgiler
  - Veri standartlaştırma
  - Dokümanlardaki kelimelerin önemi
  - Hash fonksiyonları
  - İndeksler
  - İkincil depolama birimi

24

Dokümanlardaki kelimelerin önemi

- Çoğu veri madenciliği uygulamasında, dokümanların konularına göre gruplandırılması gerekir.
- **Dokümanların konuları belirli anahtar kelimelere göre belirlenebilir.**
- Bir dokümanda **sık geçen kelimelerin o doküman için önemli olduğu tahmin edilebilir.**
- Bazen sık kullanılan kelimeler konu belirlemek için uygun olmayabilir.
- 'the', 'and' gibi kelimeler (stop words) İngilizce dokümanlarda çok sık kullanılır.
- Bir dokümanda bir kelimenin az kullanılması da konu belirlemek için tek başına yeterli değildir.

25

Dokümanlardaki kelimelerin önemi

- Kelimelerin **bir dokümanda bulunma sıklığı (term frequency)** ile diğer **tüm dokümanlarda bulunma sıklığı (inverse document frequency)** birlikte daha anlamlı sonuç vermektedir.

$$TF_{ij} = \frac{f_{ij}}{\max_k f_{kj}}$$

- Burada, f_{ij} ile **i.kelimenin j.dokümandaki frekansı** gösterilmektedir.
- $\max_k f_{kj}$ ile j.dokümanda en sık geçen kelimenin frekansı ifade edilmektedir.

$$IDF_i = \log_2(N/n_i)$$

- Burada, N tüm doküman sayısını, n_i ise i.kelimenin geçtiği doküman sayısını ifade etmektedir
- Bu iki değerin çarpımı ile bir kelimenin bir doküman için önemi hesaplanır.

$$TF_{ij} \times IDF_i$$

26

Dokümanlardaki kelimelerin önemi

Örnek

- Veritabanında 2^{20} doküman olsun.
- Bir w kelimesi 2^{10} dokümanda geçiyorsa $IDF_w = \log_2 (2^{20} / 2^{10}) = 10$ olur.
- Bir j dokümanında w kelimesi 20 kez geçiyorsa ve bu en sık geçen kelime ise $TF_{wj} = 1$ olur.
- $TF.IDF_{wj} = 10$ olur.
- Bir k dokümanında w kelimesi 1 kez geçiyorsa ve en sık geçen başka bir kelime ise 20 kez geçiyorsa $TF_{wk} = 1/20$ olur.
- $TF.IDF_{wk} = 10 \times (1 / 20) = 1/2$ olur.

27

Konular

- Veri Madenciliği
  - İstatistiksel model
  - Makine öğrenmesi
  - Modellemede hesaplamalı yaklaşımlar
  - Özetleme
  - Özellik çıkarımı
- Veri Madenciliğinde İstatistiksel Limitler
  - Toplam bilgi farkındalığı
  - Bonferroni prensibi
- Temel Bilgiler
  - Veri standartlaştırma
  - Dokümanlardaki kelimelerin önemi
  - Hash fonksiyonları
  - İndeksler
  - İkincil depolama birimi

28

Hash fonksiyonları

- **Hash fonksiyonu**, bir **h anahtarını alır** ve bir **sonuç değer üretir.**
- Bu sonuç değer, **0 ile B-1 arasında** bir tamsayı olabilir. Burada, B maksimum değer aralığını gösterir.
- **Anahtar sayısı** ile **sonuç sayısı** birbirine **eşit olabilir.**
- Bu durumda, her anahtar sadece bir sonuç üretebilir veya her sonuç için sadece bir anahtar olabilir.
- **Anahtar sayısı** ile **sonuç sayısı** birbirinden **farklı olabilir.**
- Bu durumda, her sonuç için birden fazla anahtar vardır (birden fazla anahtar aynı sonucu üretir).

$$h(x) = x \bmod B$$

29

Konular

- Veri Madenciliği
  - İstatistiksel model
  - Makine öğrenmesi
  - Modellemede hesaplamalı yaklaşımlar
  - Özetleme
  - Özellik çıkarımı
- Veri Madenciliğinde İstatistiksel Limitler
  - Toplam bilgi farkındalığı
  - Bonferroni prensibi
- Temel Bilgiler
  - Veri standartlaştırma
  - Dokümanlardaki kelimelerin önemi
  - Hash fonksiyonları
  - İndeksler
  - İkincil depolama birimi

30

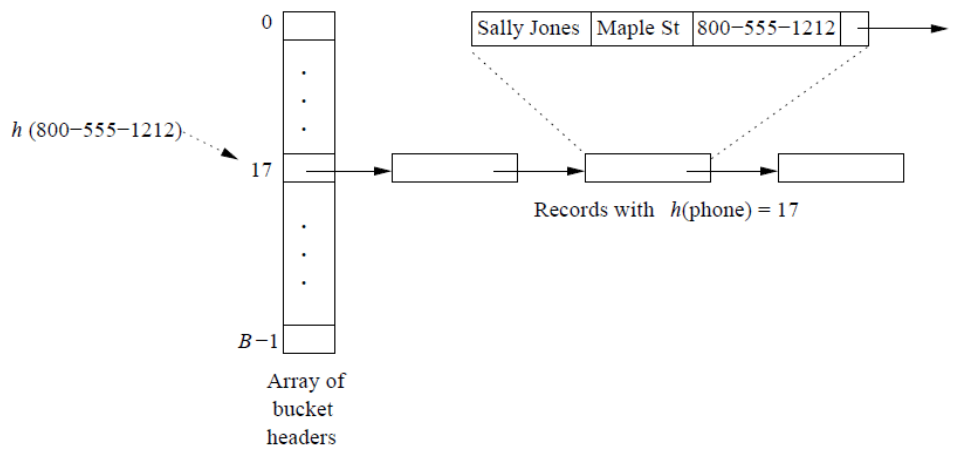
İndeksler

- **İndeks, nesnelere (kayıtlara) etkin bir şekilde ulaşmak için kullanılan veri yapısıdır**.
- İndeks, genellikle kayıt içerisindeki bir alan (field) kullanılarak oluşturulur.
- **Hash tablosu ile indeks oluşturulabilir.**
- **Field,** hash fonksiyonu için **anahtar değeri ifade eder** ve **hash fonksiyonunun sonucu kullanılarak kaydın tamamı elde edilir.**
- Sonuç değer, hafızada bir adres, diskte bir blok, vb. olabilir.

31

İndeksler

- 800-555-1212 telefon numarası hash key olarak kullanılarak kayda ulaşılıyor.



32

Konular

- Veri Madenciliği
  - İstatistiksel model
  - Makine öğrenmesi
  - Modellemede hesaplamalı yaklaşımlar
  - Özetleme
  - Özellik çıkarımı
- Veri Madenciliğinde İstatistiksel Limitler
  - Toplam bilgi farkındalığı
  - Bonferroni prensibi
- Temel Bilgiler
  - Veri standartlaştırma
  - Dokümanlardaki kelimelerin önemi
  - Hash fonksiyonları
  - İndeksler
  - İkincil depolama birimi

33

İkincil depolama birimi

- **Disk üzerindeki veriye ulaşma süresi hafızaya göre çok uzundur.**
- Disk üzerindeki bir bloğa erişim hızı hafızaya göre 10.000 kez daha yavaştır.
- Her hash anahtarıyla hesaplanan sonuç değere göre ayrı ayrı okuma yapmak performansı çok düşürür.
- **Diskler** mantıksal olarak **silindir şeklinde organize edilir** ve çok sayıdaki track üzerindeki sektör farklı okuma kafaları ile aynı anda okunur.

34