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.

Genel bilgiler

Değerlendirme

Arasınav	: 35%
Ödevler	: 25% (DersKodu-OgrenciNo-OdevNo.pdf)
Final Projesi	: 20%
Final Sınavı	: 20%

Ders kaynakları

- The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Trevor Hastie, Robert Tibshirani, Jerome Friedman, Springer, 2017.
- Mining of Massive Datasets, Jure Leskovec, Anand Rajaraman, Jeffrey David Ullman, Stanford University, 2011.
- Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners, Jared Dean, Wiley, 2014.
- Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, EMC Education Services, 2015.

e-posta : akcayol@gazi.edu.tr
web : https://bigdata.gazi.edu.tr/akcayol/

Genel bilgiler

Araştırma ödevleri

- Haftalık konu ile ilgili bir makale incelenerek detaylı rapor hazırlanacaktır.
- İncelenen makalede kullanılan **yöntemin/algoritmanın/yaklaşımın gerekçeleri** ile **elde edilen sonuçlar değerlendirilecektir.**
- İncelenen **makale son 3 yılda yayınlanmış** olacaktır.
- Makale **Q1, Q2 veya Q3 çeyrekliğinde yer alan bir dergide** yayınlanmış olacaktır (SCImago veya Scopus).
- Ödev dokümanı pdf formatında tek dosya olacak ve aşağıdaki dokümanları içerecektir:
  - İncelenen makalenin tam metni
  - Makalenin yayınlandığı yıl derginin yer aldığı çeyrekliği gösterir belge
  - Hazırlanan rapor (Kapak sayfası, İçindekiler, Özet, Materyal/Metot, Sonuçlar, Yorum)
- Dosya adı '**DersKodu-ÖğrenciNo-ÖdevNo.pdf**' formatında olacaktır.

3

Genel bilgiler

Final projeleri

- Bir yöntemin/algoritmanın bir alana uygulamasını içerecektir.
- Geliştirilecek uygulamanın algoritma kısmında hazır araç, fonksiyon veya kütüphane kullanılmayacaktır.
- Hazırlanan projenin tüm dokümanları elektronik olarak, proje raporu ise çıktı ile teslim edilecektir.

4

Genel bilgiler

Ders içeriği

1. İstatistiksel modelleme
2. Makine öğrenmesi
3. Büyük veri
4. İstatistiksel öğrenme
5. Denetimli öğrenme
6. Karar ağaçları
7. Sınıflandırıcıların değerlendirilmesi
8. Eğitim ve test kümeleri
9. Denetimsiz öğrenme
10. Kümeleme
11. Birliktelik kuralları

5

Konular

- Veri ve Bilgi
- 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

6

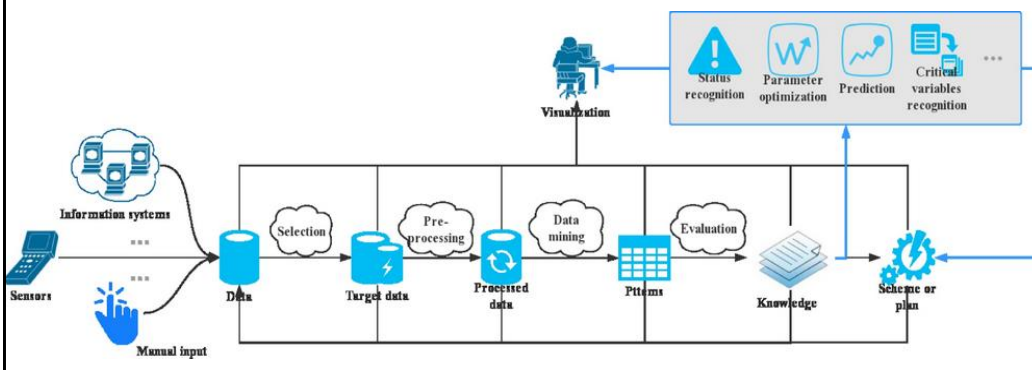
Veri ve Bilgi

- Bilgi, insanoğlu için vazgeçilmez unsurların başında gelir.
- Günümüzde **bilginin elde edilmesi, paylaşılması ve oluşturulması üzerinde teknolojik gelişmeler oldukça etkilidir.**
- Yeni teknolojilerin ortaya çıkması toplumsal yaşamın değişmesine, yeni ilişkiler ağının ortaya çıkmasına ve bilgilerin sürekli olarak yenilenmesine neden olmaktadır.
- Sözlük anlamıyla **bilgi; öğrenme, araştırma ve gözlem** yoluyla elde edilen her türlü **gerçek ve kavrayışın tümüdür.**
- Bilgi, önceden belirlenen bir dizi **sistematik kural** ve **prosedüre uygun bir biçimde işlenmiş enformasyondur.**

7

Veri ve Bilgi

- Veri ve bilgi arasındaki ilişki aşağıda görülmektedir*.



*Data and knowledge mining with big data towards smart production, Cheng, Ken Chen, Hemeng Sun, Yongping Zhang, Fei Tao, Journal of Industrial Information Integration, 9, 1-13, 2018.

8

Veri ve Bilgi

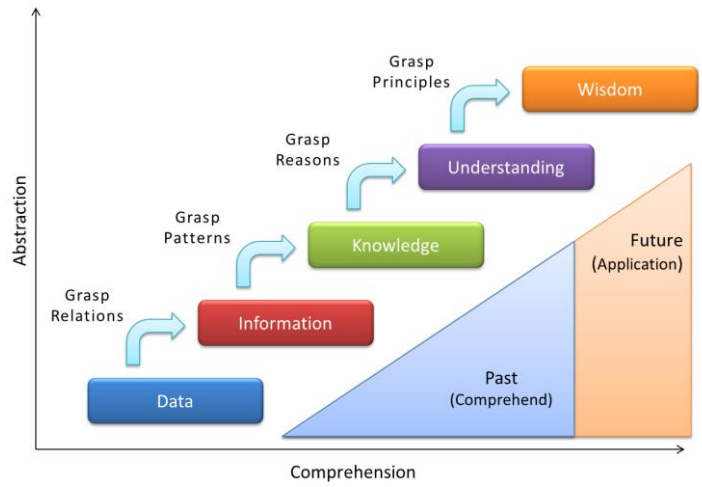
Türk Dil Kurumuna göre;

- **Veri (data):** olgu, kavram veya komutların, iletişim, yorum ve işlem için elverişli biçimde gösterimi,
- **Enformasyon (information):** haber alma, haber verme, haberleşme,
- **Bilgi (knowledge):** veriye yöneltilen anlam, insan aklının erebileceği olgu, gerçek ve ilkelerin bütünü,
- **Anlayış (understanding):** görüş ve inanış etmenlerinin etkisiyle beliren düşünme yolu, düşünüş biçimi, zihniyet, mantalite,
- **Bilgelik (wisdom):** herkesin ulaşamadığı derin, kapsamlı, bütünsel bilgi

olarak tanımlanmaktadır.

Veri ve Bilgi

- Veri ve bilgelik arasındaki ilişki aşağıda görülmektedir*.



*https://medium.com/@Iyer/strive-to-get-higher-on-the-data-information-knowledge-understanding-and-wisdom-continuum-c5ccb96438

Veri ve Bilgi

- **Veri (Data):** sayılar, rakamlar, sözcükler, metinler, resimler, olaylar vb. biçiminde temsil edilen ham gerçekliklerdir. (Örn: 54000, 01/22/2006)
- **Enformasyon (Information):** herhangi bir konu ile ilgili bir bilinmeyeni giderme konusunda yardımcı olan tanımlayıcı ifadelerdir (Örn: Nazlı'nın bankada 54.000 TL'si var, Kemal'in doğum tarihi 01/22/2006).
- **Bilgi (Knowledge):** işlenmiş enformasyondur (Örn: Nazlı'nın bankada biriken 54.000 TL'si beklediğinden fazladır).
- **Anlayış (Understanding):** sonuç veya bilgi ile ilgili neden bulma veya kavrama sürecidir (Örn: Nazlı banka işlemlerine bakınca tanımadığı birisinin 4.000 TL yatırdığını farketti. Bu nedenle bankadaki parası yüksekmiş.).
- **Bilgelik (Wisdom):** başka bir bakış açısıyla, değişen şartlar çerçevesinde ileriyi görebilme veya gözlem etkilerine göre prensipler ortaya koyma yeteneğidir (Bankaya para transferinde kişiden onay istenmelidir.).

11

Konular

- Veri ve Bilgi
- 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

12

Veri Madenciliği

- **Veri madenciliğinin** en yaygın kabul edilen **tanımı, bilgi için model keşfetmek** şeklindedir.
- Veri 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.**

13

Konular

- Veri ve Bilgi
- 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

14

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.

15

Konular

- Veri ve Bilgi
- 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

16

Makine öğrenmesi

- **Makine öğrenmesinde, veri** bir eğitim kümesi olarak alınır ve **bir modelin 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.**

17

Konular

- Veri ve Bilgi
- 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

18

Modellemede hesaplamalı yaklaşımlar

- **Bilgisayar bilimlerinde,** veri madenciliğine bir **algoritma belirleme problemi** olarak bakılır.
- **Verilerden birtakım parametreler elde edilir. Ardından optimizasyon yapılır.**
- 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.

19

Konular

- Veri ve Bilgi
- 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

20

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.**

21

Konular

- Veri ve Bilgi
- 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

22

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.

23

Konular

- Veri ve Bilgi
- 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

24

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.

25

Konular

- Veri ve Bilgi
- 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

26

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.**

27

Konular

- Veri ve Bilgi
- 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

28

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.**

29

Veri Standartlaştırma

Interval-scaled attributes

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

30

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 f niteliğine ait i. gözlem değerini ifade eder.

31

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.

32

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.

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

33

Konular

- Veri ve Bilgi
- 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

34

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.

35

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.

$$TF_{ij} \times IDF_i$$

36

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.
- $TF.IDF_{wk} = (1 / 20) \times 10 = 1/2$ olur.

37