

VERİ MADENCİLİĞİ

Veri Önışleme

Prof. Dr. Şule Gündüz Öğüdücü

1

Konular

- Veri
- Veri Önışleme
- Veriyi Tanıma
- Veri temizleme
- Veri birleřtirme
- Veri dönüşümü
- Veri azaltma
- Benzerlik ve farklılık

2

Veri Nedir?

- nesneler ve nesnelerin niteliklerinden oluşan küme
 - kayıt (record), varlık (entity), örnek (sample, instance) nesne için kullanılabilir.
- Nitelik (attribute) bir nesnenin (object) bir özelliğidir
 - bir insanın yaşı, ortamın sıcaklığı..
 - boyut (dimension), özellik (feature, characteristic) olarak da kullanılır
- Nitelikler ve bu niteliklere ait değerler bir nesneyi oluşturur.

Tid	nitelikler				Dolan dırıcı
	Geri Ödeme	Medeni Durum	Gelir		
1	Evet	Bekar	125K	-1	
2	Hayır	Evli	100K	-1	
3	Hayır	Bekar	70K	-1	
4	Evet	Evli	120K	-1	
5	Hayır	Boşanmış	95K	1	
6	Hayır	Evli	60K	-1	
7	Evet	Boşanmış	220K	-1	
8	Hayır	Bekar	85K	1	
9	Hayır	Evli	75K	-1	
10	Hayır	Bekar	90K	1	

3

Değer Kümeleri

- Nitelik için saptanmış sayılar veya semboller
- Nitelik & Değer Kümeleri
 - aynı nitelik farklı değer kümelerinden değer alabilir
 - ağırlık: kg, lb
 - farklı nitelikler aynı değer kümesinden değer alabilirler
 - ID, yaş: her ikisi de sayısal

4

Nitelik Türleri

- Belli aralıkta yeralan değişkenler (interval)
 - sıcaklık, tarih
- İkili değişkenler (binary)
 - cinsiyet
- Ayrık ve sıralı değişkenler (nominal, ordinal, ratio scaled)
 - göz rengi, posta kodu

5

Konular

- Veri
- Veri Önışleme
- Veriyi Tanıma
- Veri temizleme
- Veri birleřtirme
- Veri dönüşümü
- Veri azaltma
- Benzerlik ve farklılık

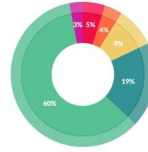
6

Problem

- Gerçek uygulamalarda toplanan veri kirlidir
 - eksik: bazı nitelik değerleri bazı nesneler için girilmemiş, veri madenciliği uygulaması için gerekli bir nitelik kaydedilmemiş
 - meslek = "
 - gürültülü: hatalar var
 - maaş= "-10"
 - tutarsız: nitelik değerleri veya nitelik isimleri uyumsuz
 - yaş= "35", d.tarihi: "03/10/2004"
 - önceki oylama değerleri: "1,2,3", yeni oylama değerleri: "A,B,C"
 - bir kaynaktan nitelik değeri 'ad', diğerinde 'isim'

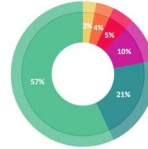
7

Veri Madenciliği Aşamaları



What data scientists spend the most time doing

- Building training sets: 2%
- Cleaning and organizing data: 60%
- Collecting data sets: 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 3%



What's the least enjoyable part of data science?

- Building training sets: 10%
- Cleaning and organizing data: 57%
- Collecting data sets: 21%
- Mining data for patterns: 10%
- Refining algorithms: 4%
- Other: 3%

Kaynak : <https://whatsthebigdata.com/2016/05/01/data-scientists-spend-most-of-their-time-cleaning-data/>

8

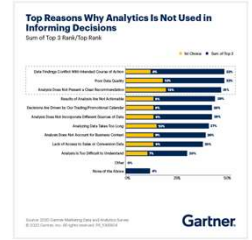
Verinin Gürültülü Olma Nedenleri

- Eksik veri kayıtlarının nedenleri
 - Veri toplandığı sırada bir nitelik değerinin elde edilememesi, bilinmemesi
 - Veri toplandığı sırada bazı niteliklerin gerekliliğinin görülememesi
 - İnsan, yazılım ya da donanım problemleri
- Gürültülü (hatalı) veri kayıtlarının nedenleri
 - Hatalı veri toplama gereçleri
 - İnsan, yazılım ya da donanım problemleri
 - Veri iletimi sırasında problemler
- Tutarsız veri kayıtlarının nedenleri
 - Verinin farklı veri kaynaklarında tutulması
 - İşlevsel bağımlılık kurallarına uyulmaması

9

Veri Hazırlama İhtiyaçları

- Veri güvenilirmez
 - Veri madenciliği sonuçlarına güvenilebilir mi?
 - Kullanılabilir veri madenciliği sonuçları kaliteli veri ile elde edilebilir
- Veri kaliteli ise veri madenciliği uygulamaları ile yararlı bilgi bulma şansı daha fazla



Kaynak : Gartner Marketing Data and Analytics Survey 2020

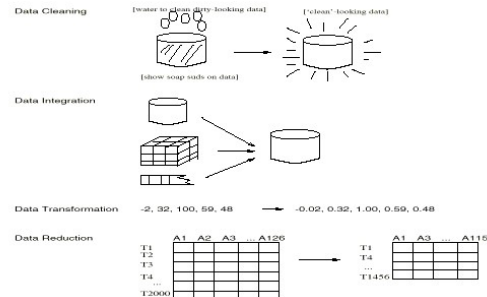
10

Veri Önışleme

- Veri temizleme
 - Eksik nitelik değerlerini tamamlama, hatalı veriyi düzeltme, aykırılıkları saptama ve temizleme, tutarsızlıkları giderme
- Veri birleştirme
 - Farklı veri kaynağındaki verileri birleştirme
- Veri dönüşümü
 - Normalizasyon ve biriktirme
- Veri azaltma
 - Aynı veri madenciliği sonuçları elde edilecek şekilde veri miktarını azaltma

11

Veri Önışleme



12

Konular

- Veri
- Veri Önleme
- Veriyi Tanıma
- Veri temizleme
- Veri birleştirme
- Veri dönüşümü
- Veri azaltma
- Benzerlik ve farklılık

13

Veriyi Tanımlayıcı Özellikler

- Amaç: Veriyi daha iyi anlamak
 - Merkezi eğilim (central tendency), varyasyon, dağılım
- Verinin dağılım özellikleri
 - Ortanca, en büyük, en küçük, sıklık derecesi, aykırılık, varyans
- Sayısal nitelikler -> sıralanabilir değerler
 - verinin dağılımı
 - kutu grafiği çizimi ve sıklık derecesi incelemesi

Popular Python libraries for Data Science



14

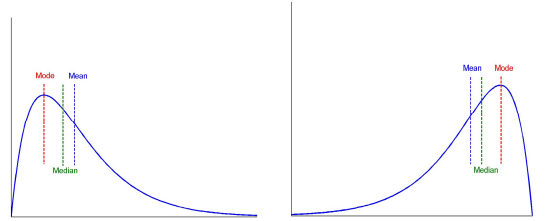
Merkezi Eğilimi Ölçme

- Ortalama: $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ $\mu = \frac{\sum x}{N}$ $\bar{x} = \frac{\sum w_i x_i}{\sum w_i}$
 - ağırlıklı ortalama
 - kırılmış ortalama: Uç değerleri kullanmadan hesaplama
- Ortanca (median): Verinin tümü kullanılarak hesaplanır
 - veri sayısı tek ise ortadaki değer, çift sayı ise ortadaki iki değerlerin ortalaması
- Mod $median = L_1 + \left(\frac{n/2 - (\sum f)_1}{f_{median}} \right) c$
 - Veri içinde en sıklıkla görülen değer
 - Unimodal, bimodal, trimodal
 - deneysel formül: $mean - mode = 3 \times (mean - median)$

15

Simetrik – Asimetrik Dağılım

- Simetrik ve asimetrik verinin ortalama, ortanca ve mod değerleri



Verinin Dağılımını Ölçme

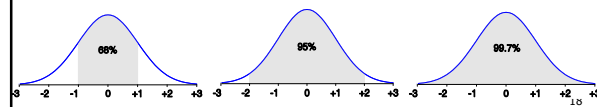
- Çeyrek, aykırılıklar, kutu grafiği çizimi
 - Çeyrek (quartile) : nitelik değerleri küçükten büyüğe doğru sıralanır.
 - Q1: ilk %25, Q3: ilk %75
 - Dörtlü aralık (Inter-quartile Range): IQR= Q3-Q1
 - Beş Sayılı Özet (Five Number Summary): min, Q1, median, Q3, max
 - Kutu Grafiği Çizimi:
 - Q1 ve Q3 aralığında bir kutu
 - kutu içinde ortanca noktayı gösteren bir çizgi
 - kutudan min ve max değerlere birer uzantı
 - Aykırılıklar: 1,5xIQR değerinden küçük/büyük olan değerler
- Varyans ve standart sapma

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 \right] \quad \sigma^2 = \frac{1}{N} \sum_{i=1}^n (x_i - \mu)^2 = \frac{1}{N} \sum_{i=1}^n x_i^2 - \mu^2$$

17

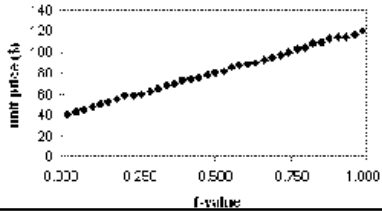
Normal Dağılımın Özellikleri

- Normal dağılım
 - $\mu - \sigma$ - $\mu + \sigma$ arasında verinin %68'i bulunur (μ : ortalama, σ : standart sapma)
 - $\mu - 2\sigma$ - $\mu + 2\sigma$ arasında verinin %95'i bulunur
 - $\mu - 3\sigma$ - $\mu + 3\sigma$ arasında verinin %99.7'i bulunur



Quantile Plot

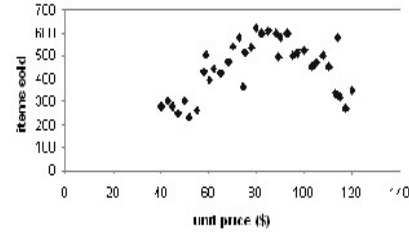
- Verinin bütününü bir nitelik değerine göre görüntüleme
- Veri bir nitelik değerine göre küçükten büyüğe doğru sıralanır
 - x_i değeri için % 100 n miktardaki veri x_i değerinden küçük ya da eşittir



19

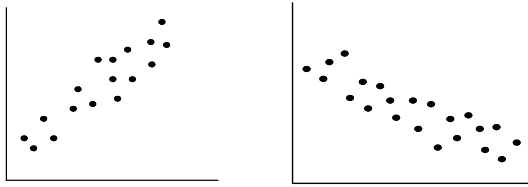
Scatter Plot

- İki sayısal nitelik değeri arasındaki ilişkiyi görmek



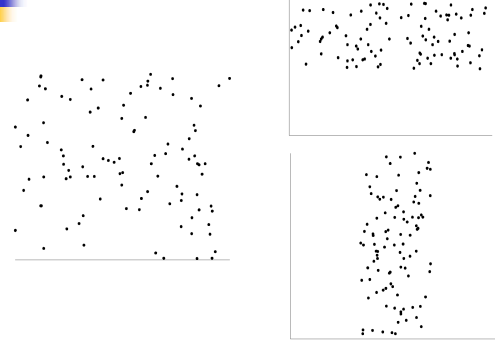
20

Pozitif – Negatif Korelasyon



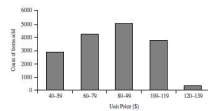
21

Korelasyon Bulunmayan Veri



Histogram

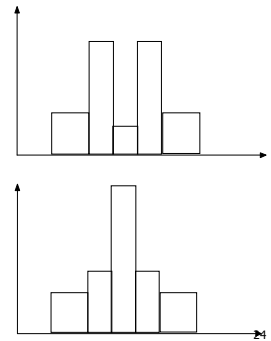
- Gruplandırılmış bir verinin dağılımı sütun grafiği ile gösterilir.
- Veriler küçükten büyüğe doğru sıralanır ve aralıklara bölünür.
- Her aralıkta kaç adet verinin o değeri aldığı sütun grafiği ile gösterilir.



23

Histogram – Kutu Grafiği

- Histogram kutu grafiğinden daha çok bilgi verebilir
- Yandaki iki histogram ile verilen nitelikler için aynı kutu grafiği çizimi elde edilebilir
 - Beş Sayılı Özet aynı
- Ancak dağılımları farklı



Konular

- Veri
- Veri Önışleme
- Veriyi Tanıma
- Veri temizleme
- Veri birleřtirme
- Veri dönüşümü
- Veri azaltma
- Benzerlik ve farklılık

25

Veri Temizleme

- Gerçek uygulamalarda veri eksik, gürültülü veya tutarsız olabilir.
- Veri temizleme işlemleri
 - Eksik nitelik değerlerini tamamlama
 - Aykırılıkların bulunması ve gürültülü verinin düzeltilmesi
 - Tutarsızlıkların giderilmesi



26

Eksik Veri

- Veri için bazı niteliklerin değerleri her zaman bilinemeyebilir.
- Eksik veri
 - diğer veri kayıtlarıyla tutarsızlığı nedeniyle silinmesi
 - Bazı nitelik değerleri hatalı olması dolayısıyla silinmesi
 - yanlış anlama sonucu kaydedilmeme
 - veri giriři sırasında bazı nitelikleri önemsiz görme

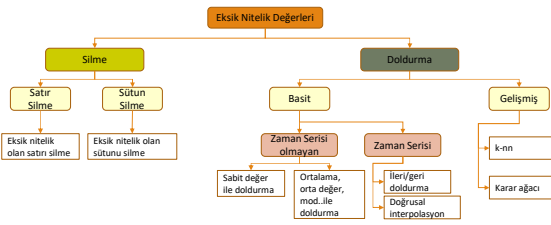
27

Eksik Veriler nasıl Tamamlanır?

- Eksik nitelik değerleri olan veri kayıtlarını kullanma
- Eksik nitelik değerlerini elle doldur
- Eksik nitelik değerleri için global bir değışken kullan (Null, bilinmiyor,...)
- Eksik nitelik değerlerini o niteliğin ortalama değeri ile doldur
- Aynı sınıfa ait kayıtların nitelik değerlerinin ortalaması ile doldur
- Olasılığı en fazla olan nitelik değeriyle doldur

28

Eksik Nitelik Değerleri



29

Gürültülü Veri

- Ölçülen bir değerdeki hata
- Yanlış nitelik değerleri
 - hatalı veri toplama gereęleri
 - veri giriři problemleri
 - veri iletimi problemleri
 - teknolojik kısıtlar
 - nitelik isimlerinde tutarsızlık
- Diğer nedenler
 - Tekrarlı kayıtlar
 - Eksik veri
 - Tutarsız veri

30

Gürültülü Veri nasıl Düzeltilir?

- Gürültüyü yok etme
 - Bölmeleme
 - veri sıralanır, eşit genişlik veya eşit derinlik ile bölünür
 - Demetleme
 - aykırılıkları belirler
 - Eğri uydurma
 - veriyi bir fonksiyona uydurarak gürültüyü düzeltir
 - Uzman görüşü
 - Şüpheli kayıtlar belirlenir ve uzman görüşüne sunulur

31

Bölmeleme

- Veri sıralanır: 4, 8, 15, 21, 21, 24, 25, 28, 34
 - Eşit genişlik: Bölme sayısı belirlenir. Eşit aralıklarla bölünür
 - Eşit derinlik: Her bölmede eşit sayıda örnek kalacak şekilde bölünür.
- her bölme ortalamayla ya da bölmenin en alt ve üst sınırlarıyla temsil edilir

Bölme derinliği:3
1. Bölme: 4, 8, 15
2. Bölme: 21, 21, 24
3. Bölme: 25, 28, 34

Ortalamayla düzeltme:

1. Bölme: 9, 9, 9
2. Bölme: 22, 22, 22
3. Bölme: 29, 29, 29

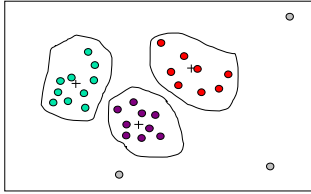
Alt-üst sınırla düzeltme:

1. Bölme: 4, 4, 15
2. Bölme: 21, 21, 24
3. Bölme: 25, 25, 34

32

Demetleme

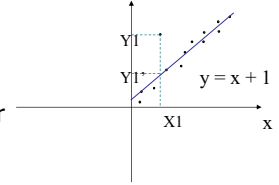
- Benzer veriler aynı demette olacak şekilde gruplanır
- Bu demetlerin dışında kalan veriler aykırılık olarak belirlenir ve silinir



33

Eğri Uydurma

- Veri bir fonksiyona uydurulur. Doğrusal eğri uydurmada, bir değişkenin değeri diğer bir değişken kullanılarak bulunabilir.



34

Konular

- Veri
- Veri Önizleme
- Veriyi Tanıma
- Veri temizleme
- Veri birleştirme
- Veri dönüşümü
- Veri azaltma
- Benzerlik ve farklılık

35

Veri Birleştirme

- Farklı kaynaklardan verilerin tutarlı olarak birleştirilmesi
- Şema birleştirilmesi
 - Aynı varlıkların saptanması:
A.cust_id=B.cust_num
 - meta veri kullanılır
- Nitelik değerlerinin tutarsızlığının saptanması
 - Aynı nitelik için farklı kaynaklarda farklı değerler olması
 - Farklı metrikler kullanılması



36

Gereksiz Veri

- Farklı veri kaynaklarından veriler birleştirilince gereksiz (fazla) veri oluşabilir
 - aynı nitelik farklı kaynaklarda farklı isimle
 - bir niteliğin değeri başka bir nitelik kullanılarak hesaplanabilir
 - korelasyon hesaplaması: sayısal nitelikler
 - =0: nitelikler bağımsız, >0: pozitif korelasyon, <0: negatif korelasyon

$$r_{A,B} = \frac{\sum (A - \bar{A})(B - \bar{B})}{(n-1)\sigma_A\sigma_B} \quad \bar{A} = \frac{\sum A}{n} \quad \sigma_A = \sqrt{\frac{\sum (A - \bar{A})^2}{n-1}}$$

- korelasyon hesaplaması: aynı nitelikler (chi-square test)

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

37

Örnek: Korelasyon Hesaplaması

$$Expected_{(male,fiction)} = \frac{Count(male) \times Count(fiction)}{n} = \frac{300 \times 450}{1500} = 90$$

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507,93$$

	male	female	Total
fiction	250 (90)	200 (360)	450
Non-fiction	50 (210)	1000 (840)	1050
Total	300	1200	1500

Null Hypothesis: Cinsiyet ve kitap türü birbirinden bağımsız

Degree of freedom= (2-1)x(2-1)=1
0,001 istatistiksel anlamlılık için 10,83

507,83 > 10,83 => Cinsiyet ve kitap türü bağımlı değişkenler

	p value															
	0.25	0.2	0.15	0.1	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005				
25%	20%	15%	10%	5%	2.5%	2%	1%	0.05%	0.025%	0.0125%	0.00625%	0.003125%				
1	1.92	1.64	1.28	0.84	0.50	0.32	0.20	0.10	0.05	0.025	0.0125	0.00625	0.003125	0.0015625	0.00078125	0.000390625
2	2.77	2.33	1.75	1.06	0.64	0.40	0.25	0.13	0.06	0.03	0.015	0.0075	0.00375	0.001875	0.0009375	0.00046875
3	3.84	3.00	2.34	1.36	0.89	0.55	0.35	0.18	0.09	0.045	0.0225	0.01125	0.005625	0.0028125	0.00140625	0.000703125
4	5.02	3.94	3.00	1.75	1.06	0.64	0.40	0.25	0.13	0.06	0.03	0.015	0.0075	0.00375	0.001875	0.0009375
5	6.63	5.41	4.10	2.34	1.36	0.89	0.55	0.35	0.18	0.09	0.045	0.0225	0.01125	0.005625	0.0028125	0.00140625
6	7.88	6.36	4.75	2.71	1.60	1.00	0.63	0.38	0.19	0.09	0.045	0.0225	0.01125	0.005625	0.0028125	0.00140625
7	9.49	7.88	5.82	3.00	1.75	1.06	0.64	0.40	0.25	0.13	0.06	0.03	0.015	0.0075	0.00375	0.001875
8	10.59	8.56	6.34	3.36	1.96	1.21	0.75	0.45	0.22	0.11	0.05	0.025	0.0125	0.00625	0.003125	0.0015625
9	11.99	9.49	7.00	3.71	2.16	1.36	0.84	0.50	0.25	0.13	0.06	0.03	0.015	0.0075	0.00375	0.001875
10	13.44	10.49	7.78	4.00	2.34	1.50	0.90	0.55	0.28	0.14	0.07	0.035	0.0175	0.00875	0.004375	0.0021875

38

Konular

- Veri
- Veri Önleme
- Veriyi Tanıma
- Veri temizleme
- Veri birleştirme
- Veri dönüşümü
- Veri azaltma
- Benzerlik ve farklılık

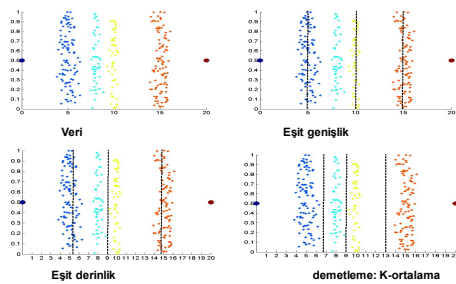
39

Veri Dönüşümü

- Veri, veri madenciliği uygulamaları için uygun olmayabilir
 - Seçilen algoritmaya uygun olmayabilir
 - Veri belirleyici değil
- Çözüm
 - Veri düzeltme
 - Bölmeleme
 - Demetleme
 - Eğri Uydurma
 - Biriktirme
 - Genelleme
 - Normalizasyon
 - Nitelik oluşturma

40

Bölmeleme ve Demetleme Yöntemleri ile Veri Düzeltme



41

Normalizasyon

- min-max normalizasyon

$$v' = \frac{v - \min_a}{\max_a - \min_a} (\text{new_max}_a - \text{new_min}_a) + \text{new_min}_a$$

- z-score normalizasyon

$$v' = \frac{v - \text{mean}_A}{\text{stand_dev}_A}$$

- ondalık normalizasyon

$$v' = \frac{v}{10^j} \quad j: \text{Max}(|v'|) < 1 \text{ olacak şekildeki en küçük tam sayı}$$

42

Nitelik Oluşturma

- Yeni nitelikler yarat
 - orjinal niteliklerden daha önemli bilgi içersin
 - alan=boy x en
 - veri madenciliği algoritmalarının başarımı daha iyi olsun

43

Konular

- Veri
- Veri Önileme
- Veriyi Tanıma
- Veri temizleme
- Veri birleştirme
- Veri dönüşümü
- Veri azaltma
- Benzerlik ve farklılık

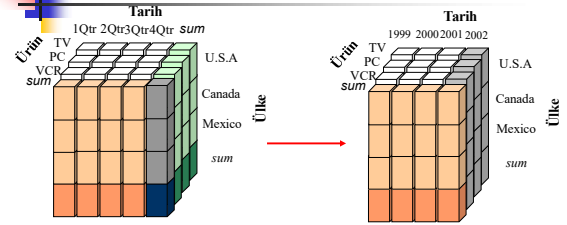
44

Veri Azaltma

- Veri miktarı çok fazla olduğu zaman veri madenciliği algoritmalarının çalışması ve sonuç üretmesi çok uzun sürebilir
 - veriyi azaltma başarımı artırır
 - sonucun (nerdeyse) hiç değişmemesi gerekir
- Veri azaltma
 - nitelik birleştirme
 - nitelik azaltma
 - veri sıkıştırma
 - veri ayrıştırma ve kavram oluşturma
 - veri küçültme
 - eğri uydurma
 - demetleme
 - histogram
 - örnekleme

45

Nitelik Birleştirme



- Sorgulama için gerekli olan boyutlar kullanılıyor.

46

Nitelik Seçme - Nitelik Azaltma

- Nitelik Seçme
 - Nitelikler kümesinin bir alt kümesi seçilerek veri madenciliği işlemi yapılır
- Nitelik azaltma
 - d boyutlu veri kümesi $k < d$ olacak şekilde k boyuta taşınır

47

Nitelik Seçme

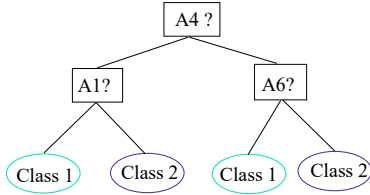
- Nitelik seçme
 - Veri madenciliği uygulaması için gerekli olan niteliklerin seçilmesi
 - Nitelikler altkümesi kullanılarak elde edilen sınıfların dağılımları gerçek dağılıma eşit ya da çok yakın olmalı
 - Veri madenciliği işlemi yer ve zaman karmaşıklığını azaltma
 - Sistemin başarımını artırma
- Sezgisel yöntemler kullanılarak nitelikler seçilebilir.
 - istatistiksel anlamlılık testi (statistical significance)
 - bilgi kazancı (information gain)
 - karar ağaçları

48

Örnek

Başlangıç nitelikler kümesi:

{A1, A2, A3, A4, A5, A6}



Seçilen nitelik kümesi: {A1, A4, A6}

49

Nitelik Azaltma

- Çok boyutlu veriyi daha küçük boyutlu uzaya taşıma
- d nitelikten oluşan n adet veri $D=\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ k boyutlu uzaya taşınır:

$$\mathbf{x}_i \in \mathbb{R}^d \rightarrow \mathbf{y}_i \in \mathbb{R}^k (k \ll d)$$

- Veri kümesinde yer alan bütün nitelikler kullanılır
 - Niteliklerin doğrusal kombinasyonu
- Niteliklerin ayıricılığına artırma

50

Veri Sıkıştırma

- Verinin boyutunu azaltır
 - daha az saklama ortamı
 - veriye ulaşmak daha çabuk
- Kayıplı ve kayıpsız veri sıkıştırma
 - bazı yöntemler bazı veri tiplerine uygun
 - her veri tipi için kullanılan yöntemler de var
- Eğer veri madenciliği yöntemi sıkıştırılmış veri üzerinde doğrudan çalışabiliyorsa elverişli

51

Veri Ayırıklaştırma

- Bazı veri madenciliği algoritmaları sadece ayırık veriler ile çalışır
- Sürekli bir nitelik değerini bölerek her aralığı etiketler
- Verinin değeri, bulunduğu aralığın etiketi ile değişir
- Veri boyutu küçülür
- Kavram oluşturmak için kullanılır

52

Kavram Oluşturma

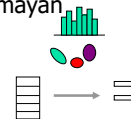
- Sayısal veriler
 - çok geniş aralıkta olabilir
 - değerleri çok sık değişebilir
- Sayısal veriler için kavram oluşturma
 - bölmeleme
 - histogram
 - demetleme
 - entropi

53

Veri Küçültme

- Veriyi farklı şekillerde gösterme

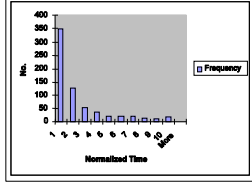
- parametrik
 - eğri uydurma
- parametrik olmayan
 - histogram
 - demetleme
 - örnekleme



54

Histogram ile Veri Küçültme

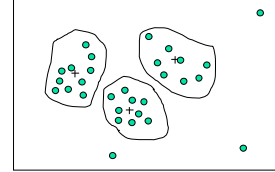
- Verinin dağılımı
- Veriyi bölerek her bölüm için veri değerini gösterir (toplam, ortalama)
 - eşit genişlik (equi-width): bölmelerin genişliği eşit
 - eşit yükseklik (equi-height): her bölmedeki veri sayısı eşit
 - v-optimal: en az varyansı olan histogram $\sum(\text{count}_b * \text{value}_b)$
 - MaxDiff: bölme genişliğini kullanıcı belirler



55

Demetleme ile Veri Küçültme

- Veri demetlere ayrılır
- Veri demetleri temsil eden örnekler (demet merkezleri) ve aykırılıklar ile temsil edilir
- Etkisi verinin dağılımına bağlı
- Hiyerarşik demetleme yöntemleri kullanılabilir.



56

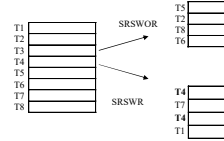
Örnekleme ile Veri Küçültme

- Büyük veri kümesini daha küçük bir alt küme ile temsil etme
- Alt küme nasıl seçiliyor?
 - yerine koymadan örnekleme (SRSWOR)
 - yerine koyarak örnekleme (SRSWR)
 - demet örnekleme (yerine koymadan veya koyarak)
 - katman örnekleme (katman: nitelik değerine göre grup)

57

Örnek

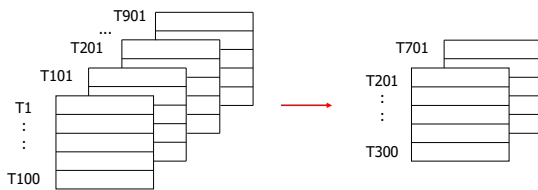
- Örnekleme



58

Örnek

- Demetleme



59

Örnek

- Katman Örnekleme

T1	genç	T1	genç
T2	genç	T4	genç
T3	genç	T6	orta yaşlı
T4	genç	T7	orta yaşlı
T5	orta yaşlı	T9	orta yaşlı
T6	orta yaşlı	T11	orta yaşlı
T7	orta yaşlı	T11	orta yaşlı
T8	orta yaşlı	T13	yaşlı
T9	orta yaşlı		
T10	orta yaşlı		
T11	orta yaşlı		
T12	orta yaşlı		
T13	yaşlı		
T14	yaşlı		

60

Konular

- Veri
- Veri Önleme
- Veriyi Tanıma
- Veri temizleme
- Veri birleştirme
- Veri dönüşümü
- Veri azaltma
- Benzerlik ve farklılık**

61

Benzerlik ve Farklılık

- Benzerlik**
 - iki nesnenin benzerliğini ölçen sayısal değer
 - nesneler birbirine daha benzer ise daha büyük
 - genelde 0-1 aralığında değer alır
- Farklılık**
 - iki nesnenin birbirinden ne kadar farklı olduğunu gösteren sayısal değer
 - nesneler birbirine daha benzer ise daha küçük
 - en küçük farklılık genelde 0
 - üst sınır değişebilir

62

Öklid Uzaklığı

- Veri kümesi

$$\begin{bmatrix} x_{11} & \dots & x_{1f} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{if} & \dots & x_{ip} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nf} & \dots & x_{np} \end{bmatrix}$$

- Uzaklık matrisi

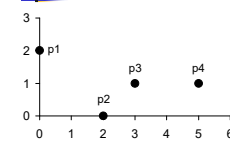
$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ \vdots & \vdots & \vdots & \\ d(n,1) & d(n,2) & \dots & 0 \end{bmatrix}$$

- Öklid uzaklığı (Euclidean Distance) nesneler arası farklılığı bulmak için kullanılır.
 - p adet niteliği (boyutu) olan i ve j nesneleri arasındaki uzaklık

$$d(i, j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \dots + |x_{ip} - x_{jp}|^2}$$

63

Örnek: Öklid Uzaklığı



nesne	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Uzaklık Matrisi

64

Minkowski Uzaklığı

- Öklid uzaklığının genelleştirilmiş hali

$$d(i, j) = \sqrt[q]{(|x_{i1} - x_{j1}|^q + |x_{i2} - x_{j2}|^q + \dots + |x_{ip} - x_{jp}|^q)} \quad q: \text{pozitif tam sayı}$$

- $q=1 \rightarrow$ Manhattan uzaklığı
- $q \rightarrow \infty$ "supremum" uzaklık ($L_{\max}$ norm, L_{∞} norm)

$$d(i, j) = \lim_{h \rightarrow \infty} \left(\sum_{f=1}^p |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = \max_f |x_{if} - x_{jf}|$$

65

Uzaklık Özellikleri

- $q=1 \Rightarrow$ Manhattan Uzaklığı (L_1 norm)
- $q=2 \Rightarrow$ Öklid Uzaklığı (L_2 norm)
- Uzaklık ölçütünün sağlaması gereken özellikler:
 - $d(i, j) \geq 0$
 - $d(i, i) = 0$
 - $d(i, j) = d(j, i)$
 - $d(i, j) \leq d(i, h) + d(h, j)$
- Uzaklıklar ağırlıklı olarak da hesaplanabilir:

$$d(i, j) = \sqrt{w_1 |x_{i1} - x_{j1}|^2 + w_2 |x_{i2} - x_{j2}|^2 + \dots + w_p |x_{ip} - x_{jp}|^2}$$

66

Benzerlik Özellikleri

- İki nesne arası benzerlik özellikleri
 - $\text{sim}(i,j) \geq 0$
 - $\text{sim}(i,j) = \text{sim}(j,i)$

67

Uzaklık Matrisleri

Manhattan (L_1)

point	attribute 1	attribute 2
x1	1	2
x2	3	5
x3	2	0
x4	4	5

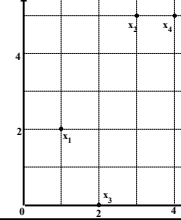
L	x1	x2	x3	x4
x1	0			
x2	5	0		
x3	3	6	0	
x4	6	1	7	0

Euclidean (L_2)

L2	x1	x2	x3	x4
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

Supremum

L _∞	x1	x2	x3	x4
x1	0			
x2	3	0		
x3	2	5	0	
x4	3	1	5	0



İkili Değişkenler Arası Benzerlik

- İkili bir değişkenin 0 veya 1 olarak iki değeri olabilir.
- Bir olasılık tablosu oluşturulur:

Nesne i		Nesne j	
		0	1
0	M_{00}	M_{01}	
1	M_{10}	M_{11}	

M_{00} : i nesnesinin 0, j nesnesinin 0 olduğu niteliklerin sayısı
 M_{10} : i nesnesinin 1, j nesnesinin 0 olduğu niteliklerin sayısı
 M_{01} : i nesnesinin 0, j nesnesinin 1 olduğu niteliklerin sayısı
 M_{11} : i nesnesinin 1, j nesnesinin 1 olduğu niteliklerin sayısı

- Yalın uyum katsayısı** (simple matching coefficient): ikili değişkenin simetrik olduğu durumlarda

$$\text{sim}(i,j) = \frac{M_{11} + M_{00}}{M_{00} + M_{01} + M_{10} + M_{11}}$$

- Jaccard katsayısı** (İkili değişkenin asimetrik olduğu durumlar):

$$d(i,j) = \frac{M_{11}}{M_{01} + M_{10} + M_{11}}$$

69

Örnek

$p = 1000000000$
 $q = 0000001001$

$M_{01} = 2$
 $M_{10} = 1$
 $M_{00} = 7$
 $M_{11} = 0$

Yalın Uyum Katsayısı:

$$(M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00}) = (0+7) / (2+1+0+7) = 0.7$$

Jaccard Katsayısı:

$$(M_{11}) / (M_{01} + M_{10} + M_{11}) = 0 / (2 + 1 + 0) = 0$$

70

Sıralı Değişkenler

- f niteliği M_f farklı değer alan sıralı bir değişken
- x_i nesnesinin f nitelik değeri $r_{if} \in \{1, \dots, M_f\}$
- x_i nesnesinin f nitelik değeri şu şekilde bir değer dönüştürülür:

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

71

Farklı Tipte Nitelikler Arası Uzaklık

$$d(i,j) = \frac{\sum_{f=1}^p \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^p \delta_{ij}^{(f)}}$$

p : nitelik sayısı
 $d(i,j)$: i ve j nesneleri arası uzaklık

$\delta_{ij}^{(f)}$

- $=0$: f niteliği her iki nesne 0 ise ve asimetrik ikili ise
- $=0$: f niteliği herhangi bir nesne eksikse
- diğer durumlarda 1

- f niteliği sayısal nitelik ise: $d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{\max_f - \min_f}$

- f niteliği nominal veya ikili ise:

- $d_{ij}^{(f)} = 0$: $x_{if} = x_{jf}$

- $d_{ij}^{(f)} = 1$: diğer durumlarda

- f niteliği sıralı değişken

- z_{if} değeri hesaplanır ve sayısal nitelik olarak değerlendirilir.

72



Kosinüs Benzerliği

- d_1 ve d_2 iki doküman. Kosinüs benzerliği

$$\cos(d_1, d_2) = d_1 \bullet d_2 / ||d_1|| ||d_2||$$

$d_i \bullet d_j$: iki dokümanın vektör çarpımı

$||d_i||$: d_i dokümanının uzunluğu

- **Örnek**

$$d_1 = 3 \ 2 \ 0 \ 5 \ 0 \ 0 \ 0 \ 2 \ 0 \ 0$$

$$d_2 = 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 2$$

$$d_1 \bullet d_2 = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$$

$$||d_1|| = (3^2 + 2^2 + 0^2 + 5^2 + 0^2 + 0^2 + 0^2 + 2^2 + 0^2 + 0^2)^{0.5} = (42)^{0.5} = 6.481$$

$$||d_2|| = (1^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 1^2 + 0^2 + 2^2)^{0.5} = (6)^{0.5} = 2.245$$

$$\cos(d_1, d_2) = .3150$$

73