VERİ MADENCİLİĞİ
Veri Önişleme

Yrd. Doç. Dr. Şule Gündüz Öğüdücü
http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195

Konular

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

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.

nitelikler

nesneler

Tid	Geri Ödeme	Medeni Durum	Gelir	Dolandırıcı
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

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

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

Konular

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

Problem

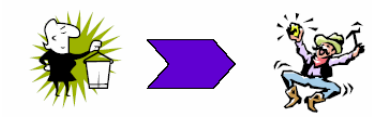
- Gerçek uygulamalarda toplanan veri kirli
  - eksik: bazı nitelik değerleri bazı nesneler için girilmemiş, veri madenciliği uygulaması için gerekli bir nitelik kaydedilmemiş
    - meslek = " "
  - gürültülü: hatalar var
    - maaş= "-10"
  - tutarsız: nitelik değerleri veya nitelik isimleri uyumsuz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

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

Sonuç

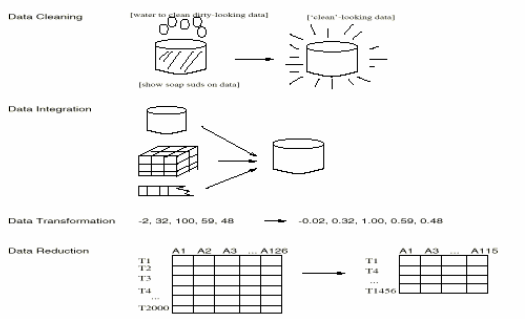
- Veri güvenilmez
  - Veri madenciliği sonuçlarına güvenilebilir mi?
  - Kullanılabilir veri madenciliği sonuçları kaliteli veri ile elde edilebilir
- Veri kaliteli ise veri madenciliği uygulamaları ile yararlı bilgi bulma şansı daha fazla.



Veri Önişleme

- 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

Veri Önişleme



Konular

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

Veriyi Tanımlayıcı Özellikler

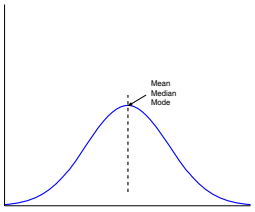
- Amaç: Veriyi daha iyi anlamak
  - Merkezi eğilim (central tendency), varyasyon, yayılma, dağılım
- Verinin dağılım özellikleri
  - Ortanca, en büyük, en küçük, sıklık derecesi, aykırılık, varyans
- Sayısal nitelikler -> sıralanabilir değerler
  - verinin dağılımı
  - kutu grafiği çizimi ve sıklık derecesi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

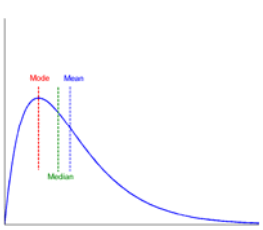
Merkezi Eğilimi Ölçme

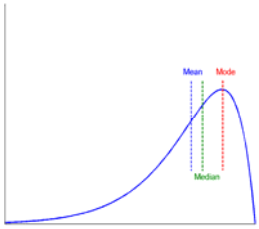
- Ortalama: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \mu = \frac{\sum x}{N}$
  - ağırlıklı ortalama $\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$
  - kırpılmış ortalama: Uç değerleri kullanmadan hesaplama
- Ortanca (median): Verinin tümü kullanılarak hesaplanır
  - veri sayısı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

Simetrik – Asimetrik Dağılım

- Simetrik ve asimetrik verinin ortalama, ortanca ve mod değerleri







Verinin Dağılımını Ölçme

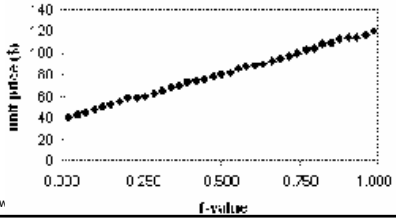
- Çeyrek, aykırılıklar, kutu grafiği çizimi
  - Çeyrek (quartile) : nitelik değerleri küçükten büyüğe doğru sıralanır.
    - Q1: ilk %25, Q3: ilk %75
  - Dörtlü aralık (Inter-quartile Range): IQR= Q3-Q1
  - Five Number Summary: min, Q1, median, Q3, max
  - Kutu Grafiği Çizimi:
    - Q1 ve Q3 aralığında 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

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

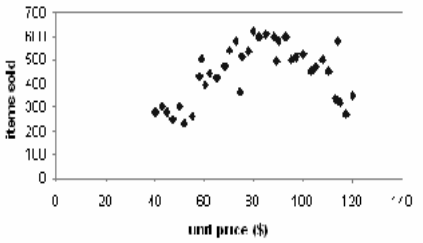
Quantile Plot

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



Scatter Plot

- İki sayısal nitelik değeri arasındaki ilişkiyi görmek



Konular

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

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

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 girişi sırasında bazı nitelikleri önemsiz görme

Eksik Veriler nasıl Tamamlanır?

- Eksik nitelik değerleri olan veri kayıtlarını kullanma
- Eksik nitelik değerlerini elle doldur
- Eksik nitelik değerleri için global bir değişken 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

Gürültülü Veri

- Ölçülen bir değerdeki hata
- Yanlış nitelik değerleri
  - hatalı veri toplama gereçleri
  - veri girişi problemleri
  - veri iletimi problemleri
  - teknolojik kısıtlar
  - nitelik isimlerinde tutarsızlık

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
- http://control.cs.berkeley.edu/abc/

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

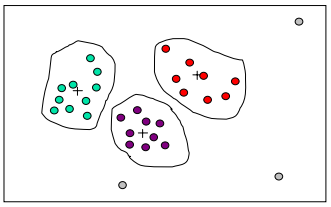
Bölme genişliği:3
1. Bölme: 4, 8, 15
2. Bölme: 21, 21, 24
3. Bölme: 25, 28, 34

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

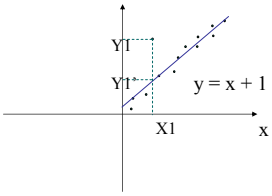
Demetleme

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



Eğri Uydurma

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



Konular

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

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

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

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

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

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

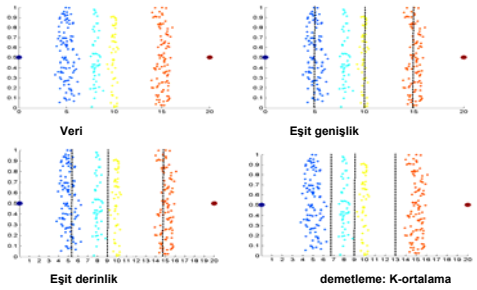
Konular

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

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

Bölmeleme ve Demetleme Yöntemleri ile Veri Düzeltme



Normalizasyon

- min-max normalizasyon

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

- z-score normalizasyon

$$v' = \frac{v - mean_A}{stand\_dev_A}$$

- ondalık normalizasyon

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

Nitelik Oluşturma

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

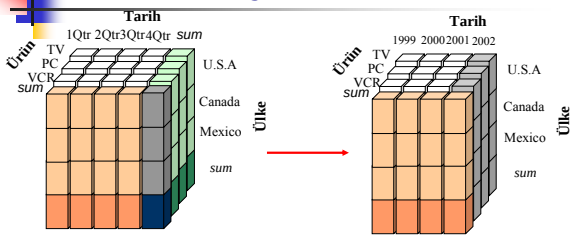
Konular

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

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

Nitelik Birleştirme



- Sorgulama için gerekli olan boyutlar kullanılıyor.

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

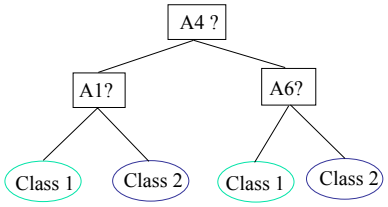
Nitelik Seçme

- 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

Örnek

Başlangıç nitelikler kümesi:

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



Seçilen nitelik kümesi: {A1, A4, A6}

Nitelik Azaltma

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

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

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

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

Veri Ayrıştırma

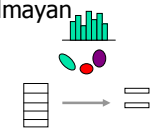
- Bazı veri madenciliği algoritmaları sadece ayrık veriler ile çalışır
- Sürekli bir nitelik değerini bölerek her aralığı etiketler
- Verinin değeri, bulunduğu aralığın etiketi ile değişir
- Veri boyutu küçülür
- Kavram oluşturmak için kullanılır

Kavram Oluşturma

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

Veri Küçültme

- Veriyi farklı şekillerde gösterme
  - parametrik
    - eğri uydurma
  - parametrik olmayan
    - histogram
    - demetleme
    - örnekleme

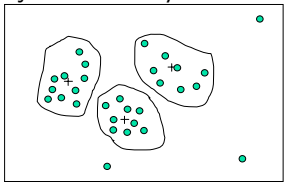


Histogram ile Veri Küçültme

- 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 $\Sigma(count_b * value_b)$
  - MaxDiff: bölme genişliğini kullanıcı belirler

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.

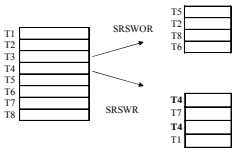


Örnekleme ile Veri Küçültme

- 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)

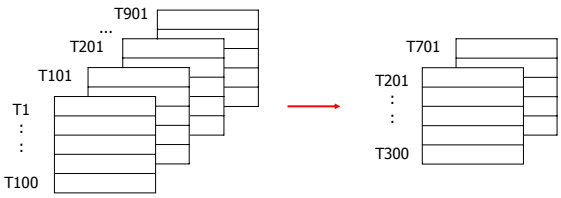
Örnek

- Örnekleme



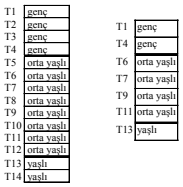
Örnek

- Demetleme



Örnek

- Katman Örnekleme



Konular

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

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

Öklid Uzaklığı

- Veri kümesi

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

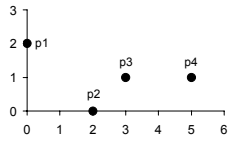
- Uzaklık matrisi

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

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

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

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

Minkowski Uzaklığı

- Öklid uzaklığının genelleştirilmiş hali

$$d(i,j) = \sqrt[q]{(|x_{i1} - x_{j1}|^q + |x_{i2} - x_{j2}|^q + \ldots + |x_{ip} - x_{jp}|^q)}$$

q: positif tam sayı

- q=1 → Manhattan uzaklığı

Uzaklık Özellikleri

- q=1⇒ Manhattan Uzaklığı
- q=2 ⇒ Öklid Uzaklığı
- Uzaklık ölçütünün sağlaması gereken özellikler:
  1. d(i,j)≥0
  2. d(i,i)=0
  3. d(i,j)=d(j,i)
  4. d(i,j)≤d(i,h)+d(h,j)
- Uzaklıklar ağırlıklı olarak da hesaplanabilir:

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

Benzerlik Özellikleri

- İki nesne arası benzerlik özellikleri
  1. sim(i,j)>=0
  2. sim(i,j)=sim(j,i)

İkili Değişkenler Arası Benzerlik

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

		Nesne *j*	
		0	1
Nesne *i*	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

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

- Jaccard katsayısı (İkili değişkenin asimetrik olduğu durumlar):

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

Örnek

p = 1 0 0 0 0 0 0 0 0 0
q = 0 0 0 0 0 0 1 0 0 1

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

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:

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

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 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
  $||d_1||$ = $(3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)^{0.5}$ = $(42)^{0.5}$ = 6.481
  $||d_2||$ = $(1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)^{0.5}$ = $(6)^{0.5}$ = 2.245

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