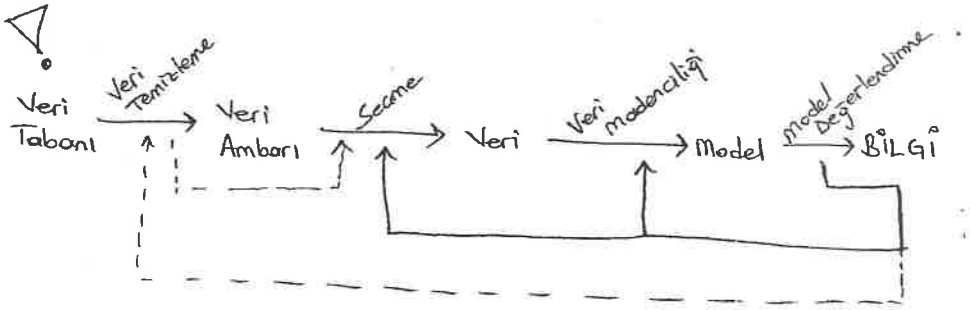
VERİ MADENCİLİĞİ NOTLARI

* Teknolojinin gelişimiyle bilgisayar ortamında ve veri tabanında tutulan veri miktarının artması, kullanıcıların beklentilerinin artması veri madenciliğinin gelişmesine yol açmıştır.

Veri madenciliği; fazla miktardaki veri içinden yararlı bilgiyi bulmak için kullanılır.

* Bilgi keşfi ise büyük veri kaynaklarından yararlı ve ilginç bilgiyi bulmak için kullanılır. Bulunan bilgi; gizli, yararlı ve önceden bilinmeyen olmalıdır.

* Teoride ⇒ bilgi keşfi > veri madenciliği
  Pratikte ⇒ bilgi keşfi ≡ veri madenciliği



Veri Kaynakları

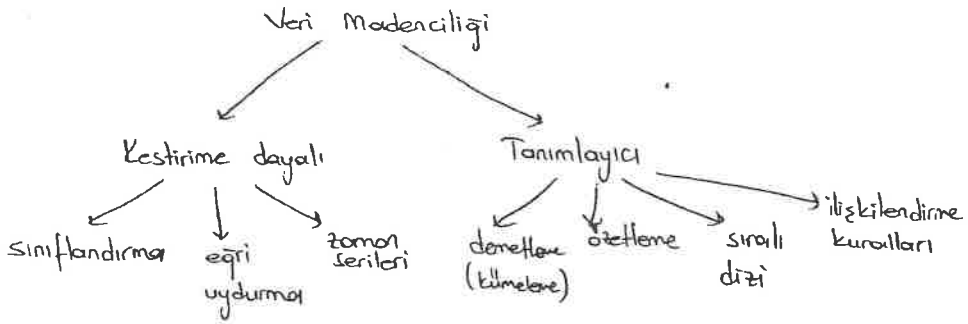
* Veri dosyaları
* Veri tabanı kaynaklı veri kümeleri
* Gelişmiş veri kümeleri (sosyal ağ, www, konumsal veriler...)

Veri Madenciliği Algoritmaları

amaç: Veriyi belli bir modele uydurmak

* tanımlayıcı → en iyi müşteri kim?
* kestirime dayalı → borsa tahmini

Veri Madenciliği Modelleri



Sınıflandırma: Veriyi önceden belirlenmiş sınıflardan birine dahil eder.
Eğri Uydurma (Regresyon): Veriyi gerçel değerli bir fonk. dönüştürür.
Zaman serileri inceleme: Zaman içinde değişen verinin değerini öngörür.
Denetleme: Benzer verileri aynı grupta toplama.
Özetleme: Veriyi alt gruplara ayırır, her alt grup için özellikler bulur.
Sıralı dizileri bulma: Veri içinde sıralı örüntüleri bulmak için kullanılır.

Bilgi Keşfinin Aşamaları

→ Uygulama alanını inceleme
→ Amaca uygun veri kümesi oluşturma
→ Veri aykılama ve önişleme
→ Veri azaltma ve veri dönüşümü
→ Veri madenciliği tekniği seçme
→ Veri madenciliği algoritmasını seçme
→ Model değerlendirme ve bilgi sunumu
→ Bulunan bilginin yorumlanması.

⚠ Veri madenciliğinde amacı büyük miktardaki veri içerisinden arama yapmak değildir bunu VTYS'leride yapıyor. Amacı; aradığımız veri mevcut ise sonuçlarını anlamaktır.

* Veri tabanı ve veri madenciliği farkı *

Veri Tabanı → Tanımlı bir sorgulama ve SQL dili kullanılır.
→ Canlı veri kullanılır.
→ Belirli bir çıkış elde edilir ve bu verinin bir alt kümesidir.

Veri Madenciliği → Tam tanımlı olmayan sorgulama ve yaygın dil yok.
→ Üzerinde işlem yapılmayan veri kullanılır.
→ Çıkış belirli değildir ve verinin alt

* Veri madenciliğinin uygulama alanları ise;

* Vt analizi ve karar verme desteği.
1. Risk analizi 2. Pazar Araştırması
3. Sahtekarlıkların saptanması gibi uygulamalardır.

1- Risk Analizi → Finans planlanması ve bilanço değerlendirmesi
→ Kaynak planlaması
→ Rekabet (rakipleri ve pazar eğilimlerini inceleme)

2- Pazar Araştırması → Uygulamalar için veri kaynağı (anket kart, kupon)
→ Hedef pazarlar bulma
→ Müşterilerin davranışlarında zaman içindeki değişiklik.
→ Coğrafi pazar incelemesi
→ Müşteri profili
→ Müşterilerin ihtiyaçlarını belirleme

3- Sahtekarlıkların Saptanması → Sigorta, bankacılık ve telekomünikasyon alanlarında geçmiş veri kullanılarak sahtekarlık yapanlar için bir model oluşturma ve benzer davranış gösterenleri belirleme.

ÖRNEK → Araba Sigortası
Sağlık Sigortası
Kredi Kartı Başvurusu.

*Veri Temizleme : Gerçek uygulamalarda veri eksik, gürültülü veya tutarsız olabilir amaç bunlardan kurtulmak..

- Eksik veri : Genellikle eksik nitelik değerleri olan silinir. Ama elle de doldurulabilir ve değer ort. ile de doldurulur.
- Gürültülü Veri : Ölçülen bir değerdeki hata gibi. Gürültüyü yok ederek sorunu çözeriz. Şu yöntemlerle yok ederiz;
  ① - Bölmeleme : Veri sıralanır buna göre bölümlere ayrılır
  ② - Kümeleme : Benzer veriler aynı demette gruplanır, aykırılar silinir.
  ③ - Eğri Uydurma : Veri bir fonksiyona uydurulur. Doğrusal eğri uydurmada bir değişkenin değeri diğer bir değişken kullanılarak bulunabilir.

* Farklı kaynaklardan olan veriler tutarlı bir şekilde birleştirilmeli ve gereksiz veriler silinmeli.

* Veri, veri madenciliği uygulamaları için uygun olmayabilir. Belirli yöntemlerle uygun hale getirilmeli.

* Veri miktarı fazla olduğunda algoritmalar çok uzun sürebilir. Veriyi azaltmak gerekebilir.

→ Nitelik birleştirme işlemi : Sorgulama için gerekli olan boyutlar kullanılıyor, bu nitelik seçme işlemi ise tüm kümenin bir alt kümesi ile işlem yapmakla olur. d boyutlu küme k<d olacak şekilde k boyutuna taşınır.

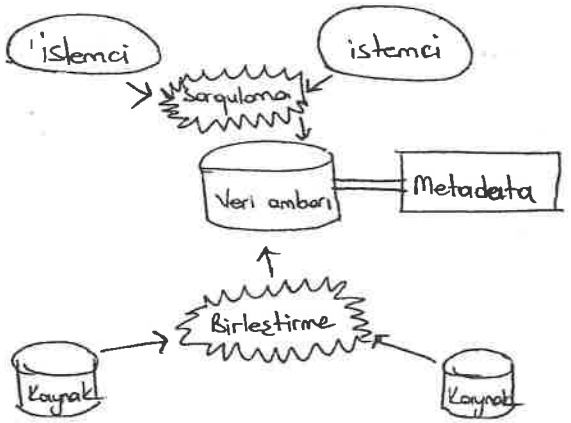
Veri Ambarları → İlişkili verilerin sorgulandığı ve analizlerinin yapılabildiği bir depodur. Vt'nını yormamak için oluşturulmuştur. Bir veri ambarı ilgili veriyi kolay, hızlı ve doğru biçimde analiz etmek için gerekli işlemleri yerine getirir.

→ amaca yönelik : Konuyla ilgili karar vermek için gerekli olmayan veriyi kullanmayarak konuya basit, özet bakış sağlar.

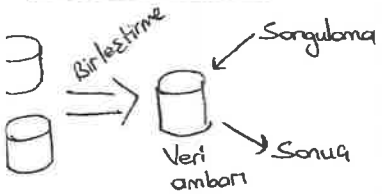
→ birleştirilmiş : Veri kaynaklarının birleştirilmesiyle oluşturulur. Dosyalar gibi.

→ zaman değişkenli : Zaman değişkeni canlı veritabanlarına göre daha uzundur.

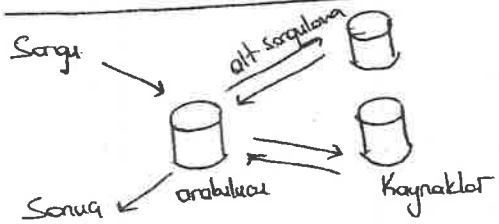
→ Değişken değil : Canlı veritabanlarından alınmış verinin fiziksel olarak başka bir ortamda saklanmasıdır.



Veri Ambarı



Birleştirilmiş Veritabanları



Veri Madenciliğinde Sorunlar

* Gizlilik ve sosyal haklar → Kişilere ait verilerin toplanarak, kişilerden izinsiz olarak kullanılması.

* Kullanıcı arabirimi → Görüntüleme ve etkileşim

* Veri madenciliği yöntemi → Farklı tipte veriler üzerinde, gürültü de, değişimli biçimde çalışma zorunda kalma.

* Başarım ve ölçeklenebilirlik →

* Veri madenciliği yöntemleri bilincsiz kullanılmamalı. Çünkü; bu yöntemler geçmiş olaylara bakarak örüntüleri bulur, gelecektekilerin garantisini vermez.

VERİ

Veri → Nesneler ve nesnelerin niteliklerinden oluşan kümedir.
* Kayıt, varlık ve örnek de denilebilir.

Nitelik → Nesnenin bir özelliğidir.

Değer Kümesi → Nitelik için saptanmış sayılar veya sembollerdir.

* Nitelik türleri → Belirli aralıkta yer alan değişkenler
→ İkili değişkenler
→ Ayrık ve sıralı değişkenler

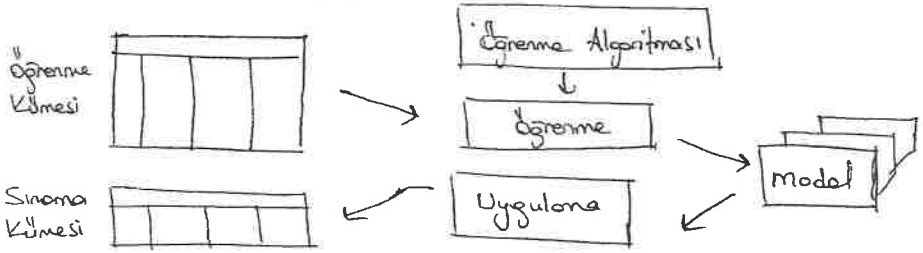
* Gerçek uygulamalarda toplanan veri kirli yani; eksik, gürültülü (maaş = -10 tl) ve tutarsız olabilir. Bunların sebepleri; kurallara uyulmaması, o anda anlık hatalar, kaynak hataları olabilir. Bu şekilde güvenilmez veri ortaya çıkar.

* Veri kaliteli ise veri madenciliği uygulamaları ile yararlı bilgi bulma şansı daha fazla. Bu yüzden veri önişleme ile veri kaliteli hale getirilmelidir.

Veri Önişleme

Veri temizleme → Eksik nitelik değerleri tamamlama gibi.
Veri birleştirme → Farklı veri kaynağındaki verileri birleştirme
Veri dönüşümü → Normalizasyon ve biriktirme.
Veri azaltma → Aynı veri madenciliği sonuçları elde edilecek şekilde veri miktarını azaltma.

3- Modeli Kullanma : Model, daha önceden görülmemiş örnekleri sınıflandırmak için kullanılır.



SINIFLANDIRMA BAŞARIMINI DEĞERLENDİRME

* Doğru sınıflandırma başarısı
* Hız
* Kararlı olması
* Ölçeklenebilirlik
* Anlaşılabilir olması
* Kuralların yapısı

SINIFLANDIRMA YÖNTEMLERİ

* Karar ağaçları
* Yapay sinir ağları
* Bayes sınıflandırıcılar
* İlişki tabanlı sınıflandırıcılar
* k-en yakın komşu yöntemi
* Destek vektör makineleri
* ⋮

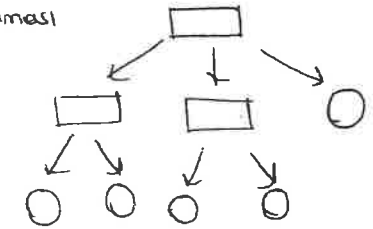
Karar Ağaçları

Akış diyagramı şeklinde ağaç yapısı

□ * Ara düğümler bir nitelik sınaması

↘ * Dallar sınama sonucu

○ * Yapraklar sınıflar



Benzerlik ve Farklılık

* Nesneler birbirine daha benzer ise daha büyük değer alırlar. Bu değer genelde 0-1 aralığında olur.

Öklid Uzaklığı : Nesneler arası farklılığı bulmak için kullanılır.

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

Minkowski Uzaklığı : Öklidin genelleştirilmiş hali.

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

$q \to 1$ olursa → Manhattan Uzaklığı

$q \to 2$ olursa → Öklid Uzaklığı

*

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

M_{00}: $i=0$, $j=0$ olduğu nitelik sayısı.

* Yalın Uyum Katsayısı → $sim(i,j) = \dfrac{M_{11} + M_{00}}{M_{00} + M_{01} + M_{10} + M_{11}}$

* Jeccard Katsayısı → $d(i,j) = \dfrac{M_{11}}{M_{01} + M_{10} + M_{11}}$

ÖRNEK: p = 1000000000
q = 0000001001

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

alın U.K $= \frac{0+7}{2+1+0+7} = 0.7$ // Jacoard K $= \frac{0}{2+1+0} = 0$ //

* Kosinüs Benzerliği : $\cos(d_1, d_2) = \frac{d_1 \cdot d_2}{\|d_1\| \cdot \|d_2\|}$ → vektör çarpım, → uzunluk

ÖRNEK: $d_1 = 3205$
$d_2 = 1000$

$d_1 \cdot d_2 = 3.1 + 2.0 + 0.0 + 5.0 = 3$
$\|d_1\| = (9+4+25)^{0.5} = 6.164$
$\|d_2\| = 1 = 1$

$\cos(d_1, d_2) = \frac{3}{6.164 \times 1} = 0,48669685$ olunur //

TEMEL SINIFLANDIRMA YÖNTEMLERİ

Bir niteliğin değerini diğer nitelikleri kullanarak belirlemeye sınıflandırma denir.

* Veri dağılımına göre bir model bulunur, başarımı belirlendikten sonra niteliğin gelecekteki değerini tahmin etmek için kullanılır

* Veri madenciliğinde;
- ayrık nitelik değ. tahmin etmek → Sınıflandırma
- sürekli nitelik " " " → öngörü"

- Gözetimli sınıflandırma → Sınıfların sayısı ve hangi nesnenin hangi sınıfta olduğu biliniyor. (Sınıflandırma)
- Gözetimsiz sınıflandırma → Hangi nesnenin hangi sınıfta olduğu bilinmiyor. Genelde sınıf sayısı da bilinmiyor. (Kümeleme)

* Kredi kartı borç harcamasının sahtekarlık olup olmadığına karar verme, ses tanıma, hastalık teshisi gibi alanlarda sınıflandırma kullanılır.

* Sınıflandırma için veri hazırlama → Veri dönüşümü, → Veri temizleme } işlemleri yapılır.

* Sınıflandırma işlemi üç aşamadan oluşur:

1) Model Oluşturma : Her nesnenin sınıf etiketi olarak tanımlanan niteliğin belirttiği bir sınıfta olduğu varsayılır. Model oluşturmak için kullanılan nesnelerin oluşturduğu veri kümesi öğrenme kümesi olarak tanımlanır.

2) Model Değerlendirme : Modelin başarısı sınama kümesi örnekleri kullanılarak belirlenir. Sınıf etiketi bilinen bir sınama kümesi örneği model kullanılarak belirlenen sınıf etiketiyle karşılaştırılır. Modelin doğruluğu, doğru sınıflandırılmış sınama kümesi örneklerinin toplam sınama kümesi örneklerine oranı olarak belirlenir.

Ağaç oluşturmada bölme temel yaklaşımlar

* Bölme kriteri
* Durma kararı
* Dallanma kriteri
* Etiketleme kuralı

Karar Ağacı Kullanılarak Sınıflandırma

* Doğrudan
* Dolaylı

Avantajı:

* Karar ağacı oluşturmak zahmetsiz
* Yorumlamak kolay
* Sürekli ve ayrık nit. için kullanılabilir

Dezavantajı:

* Sürekli nit. tahmin etmekte çok başarılı değil.
* Sınıf sayısı fazla ve örnekleme kümesi az olduğunda başarısız.

Karar Ağaçlarında Aşırı Öğrenme

Örnekleme kümesindeki örneklerin azlığı veya gürültülü olması.

Bunu engelleyen 2 yaklaşım;

- İşleni erken sona erdirme.
- Ağacı oluşturduktan sonra ağacı küçültme.

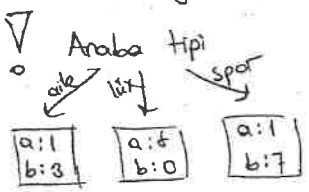
Karar ağacı yöntemleri

Genel olarak iki aşamadan oluşur.

1 - ağaç oluşturma
2 - ağaç budama

Karar ağacı oluşturma

* ağacı bütün verinin oluşturduğu tek bir düğümle başlıyor.
(a) * eğer örnekleri hepsi aynı sınıfa aitse düğüm yaprak olarak sonlanıyor ve sınıf etiketini alıyor.
Nit * eğer değil ise örnekleri sınıflara en iyi bölecek dan nitelik seçiliyor.
* İşlem sona eriyor.



a → 10
b → 10
} örnek var

farkı en yüksek olan kalitesi yüksek bölmedir.

* Ama en iyi bölen nitelik iyilik fonksiyonu ile belirlenir.

1) Bilgi Kazancı: Bütün niteliklerin ayrık değerler aldığı varsayılıyor.
2) Gini İndex: Her nitelik için olası bütün ikiye bölmeler sınanır.

(1) BİLGİ KAZANCI

* Entropi → Rastgelelik, belirsizlik, beklenmeyen durumun ortaya çıkma olasılığını gösterir.

$$H(p_1, p_2 \dots, p_n) = -\sum_{i=1}^{n} p_i \cdot \log(p_i)$$

→ Sınıflandırmada dağılım olması beklenen bir durum olduğu için entropi = 0 olmalı.

→ Örnekler aynı sınıfa aitse entropi = 0

→ Örnekler sınıflar arası eşit dağılmışsa entropi = 1

→ Örnekler sınıflar arası rastgele dağılmışsa 0 < entropi < 1 olur.

Örnek: S veri kümesinde 14 örnek : a sınıfında 9, b sınıfında 5 } örnek varsa.

$$\text{Entropi} = -\left(\frac{9}{14}\right)\log_2\left(\frac{9}{14}\right) - \left(\frac{5}{14}\right)\cdot\log_2\left(\frac{5}{14}\right) = 0{,}940$$

BİLGİ KAZANCI ;

A niteliğinin S veri kümesindeki bilgi kazancı;

$$\text{Bilgi Kazancı}(S,A) = \text{Entropi}(S) - \sum_{\text{Values}(A)} \frac{|S_v|}{|S|}\cdot \text{entropi}(S_v)$$

ÖRNEK: Önce tüm niteliklerin bilgi kazançları ayrı ayrı hesaplanır. Bilgi kazanç büyük olan kök düğüm seçilir. Aynı işlemler yapraklara ulaşılana kadar yapılır.

Mesela ; Entropi (S) = 0,940

Entropi (Sweak) = 0,811 Entropi (Sstrong) = 1.00

Gain (wind) = $0.940 - \left(\frac{8}{14}\right)^* 0{,}811 - \left(\frac{6}{14}\right)\cdot 1.00 =$ 0,048 olur.

(weak : 8) (strong : 6)

② GİNİ INDEX

* Veri kümesi S içinde n sınıf varsa ve p_j c_j sınıfının olasılığı ise ;

$$\text{gini}(S) = 1 - \sum_{j=1}^{n} p_j^2$$

* Eğer veri kümesi S_1 ve S_2 altkümelere bölünüyorsa ve her altkümede sırasıyla N_1 ve N_2 örnek varsa ;

$$\text{gini}_{\text{split}}(S) = \frac{N_1}{N}\cdot \text{gini}(S_1) + \frac{N_2}{N}\cdot \text{gini}(S_2)$$

* Gini index değeri en küçük olan seçilir

ÖRNEK ;

c1	0
c2	6

$P(C1) = 0/6 = 0$ $P(C2) = 6/6 = 1$

$\text{Gini} = 1 - P(C1)^2 - P(C2)^2 = 1 - 0 - 1 = 0$

c1	1
c2	5

$P(C1) = 1/6$ $P(C2) = 5/6$

$\text{Gini} = 1 - \frac{1}{36} - \frac{25}{36} = 1 - \frac{13}{18} = \frac{5}{18} = 0{,}278$

BÖLMELEME

*

Isı	40	48	60	72	80	90
Tenis Oynamak	No	No	Yes	Yes	Yes	No

⇒ Dinamik.

* Sürekli nitelik A sıralanır. Birbirini izleyen ancak sınıf etiketi farklı olan nitelik değerleri bulunur. En fazla kazancı sağlayan bölme seçilir.

* Statik'de ise en kaba bölmelenir. Bölmeler eşit genişlik, eşit derinlik veya demetleme yöntemi ile bulunur.

FARKLI SINIFLANDIRMA YÖNTEMLERİ

1) Örnek tabanlı yöntemler : Öğrenme kümesi saklanır, sınıflandırılacak yeni bir örnek geldiğinde öğrenme kümesi sınıf etiketini öngörmek için kullanılır.

* k-en yakın komşu yöntemi : Sınıflandırılmak istenen örneğe en yakın örnekleri bul.

⚠ Bütün örnekler n-boyutlu uzayda bir noktaya karşı düşünülür. Nesneler arasındaki uzaklık (öklid) hesaplanır. Öğrenilen fonksiyon ayrık değerli veya gerçel değerli olabilir.

* Ayrık değerli fonksiyonlarda;

k-komşu algoritması x_q örneklerine en yakın k öğrenme örneğinde en çok görülen sınıf değerini verir.

* Sürekli değerli fonksiyonlarda;

en yakın k öğrenme örneğinin ortalaması alınır.

2) Genetik Algoritmalar : Optimizasyon amaçlıdır. Bir başlangıç çözümü öneriyor, tekrarlanan her ana adımda daha iyi çözüm üretmeye çalışıyor. 5 ana parçadan oluşuyor.

→ Bireylerden oluşan bir başlangıç kümesi, P
→ Çaprazlama : Bir anne babadan yeni bireyler üretmek için yapılan işlem
→ Mutasyon : Bir bireyi rastgele değiştirme.
→ Uygunluk : En iyi bireyleri belirleme.
→ Çaprazlama ve mutasyon tekniklerini uygulayan ve uygunluk fonk. göre topluluk içindeki en iyi bireyleri seçen algoritma.

Avantajları → Paralel calışabilir.

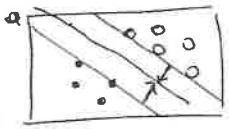
NP karmaşık problem çözümlerine uygun

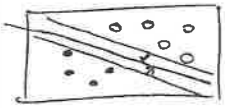
Dezavantajları → Son kullanıcının modeli anlaması zor.

Problemi GA ile çözmeye uygun hale getirmek zor.

Uygunluk fonksiyonunu belirlemek zor.

3) Destek Vektör Makineleri : Hem doğrusal olarak ayırt edilebilen hem de edilemeyen veri kümesini sınıflandırabilir.

Verileri ; a  → Bir çözüm önerisi.

b  → Başka bir çözüm önerisi.

Hangisi daha iyi ?

a , b'den daha iyi...

4) Bulanık Küme Sınıflandırıcılar :

Bulanık mantık 0.0 ve 1.0 arasında gerçel değerler kullanarak üyelik dereceleri hesaplar. Nitelik değerleri bulanık değerlere dönüştürülür. Kurallar kümesi oluşturulur.

Yeni bir örneği sınıflandırmak için birden fazla kural kullanılır. Her kuraldan gelen sonuç toplanır.

5) Öngörü :

Sınıflandırma problemleriyle aynı yaklaşım .. Model oluştur, bilinmeyen değeri hesaplamak için modeli kullan. Ama ;

Sınıflandırma → Ayrık değerli

Öngörü → Sürekli değerli.

Eğri Uydurma

Doğrusal eğri uydurma → En basit eğri uydurma yöntemidir. Veri doğrusal bir eğri ile modellenir.

$$y = w_0 + w_1 \cdot a_1 + w_2 \cdot a_2 + \cdots + w_k \cdot a_k$$

İLİŞKİLENDİRME KURALLARI

* Veri kümesi içindeki yaygın örüntülerin, nesneleri oluşturan nitelikler arasındaki ilişkilerin bulunması için vardır.

* Kural Şekli : "Gövde → Baş[destek, güven]"

satın alma(x, "ekmek") → satın alma(x, "süt")[%0.6, %65]

Yaygın Nitelikler

Nitelikler Kümesi → 1 veya daha fazla nitelikten oluşan küme.

Destek Değeri → Bir nitelikler kümesinin veri kümesinde görülme sıklığı → 5 kişiden 3'ü kola aldıysa d.d=3

Güven Değeri → $\frac{\text{Destek Değeri}}{\text{Kendisi}}$ → $g.d = \frac{3}{5}$

* Amaç; D hareket kümesinden kurallar oluşturmak;

Kuralların d.d, belirlenen en küçük d.d'nden büyük olmalı. (minsup)

Kuralların g.d, " " " g.d' " " (minconf)

eşitlikte kabulümüz " → bu eşik değerleri dışardan verilir

Brute-force Yaklaşım

Olası bütün kuralları listele

Her kural için destek ve güven değeri hesapla.

minsup ve minconf eşik değerlerinden küçükleri sil.

Hesaplama maliyeti yüksek.

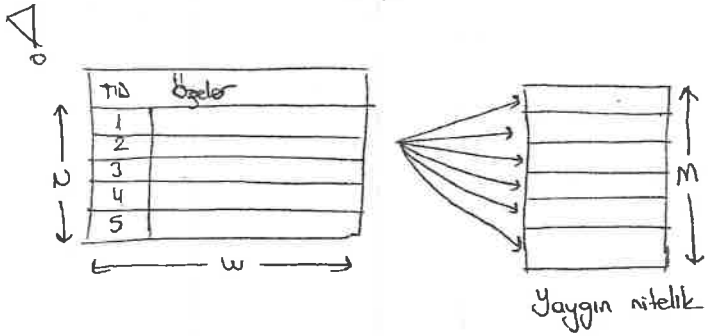
$R = 3^d - 2^{d+1} + 1$ d=Nitelik sayısı

İlişkilendirme Kural Oluşturma 2 aşamadan oluşur

1- Yaygın nitelik belirleme. (minsup ≯ a olmalı.) → abc

2- Kural oluşturma. → ab → c , b → ac ..... hepsi yazılır.

d nitelik için $2^d - 1$ yaygın nitelik oluşturulabilir.



Yaygın nitelik adayları.

* Yaygın nitelik oluşturma yöntemleri *

→ Yaygın nitelik aday sayısını (m) azaltma (Apriori Algoritması).

→ Hareket sayısını (N) azaltma.

→ Veri kümesi tarama sayısını azaltma (mN)

APRİORİ ALGORİTMASI

Temel Yaklaşım: $\forall X, Y : (X \subseteq Y) \Rightarrow s(X) \geq s(Y)$

* Bir nitelikler kümesinin destek değeri altkümesinin destek değerin. dan büyük olamaz.

* Bir yaygın nitelikler kümesinin alt kümesi de yaygın nitelikler kümesidir.

ÖRNEK

TID	Özeler
1	Ekmek, Süt
2	Ekmek, Bez, Bira, Yumurta
3	Süt, Bez, Bira, Kola
4	Ekmek, Süt, Bez, Bira
5	Ekmek, Süt, Bez, Kola
6	Şakaa :)

Çözüm

$C_1 \rightarrow$

Özeler	Sayı
Ekmek,	4
Süt	4
Bez	4
Bira	3
~~Yumurta~~	~~1~~
~~Kola~~	~~2~~

minsup = 3 ⇒ bu eşik değeri dışarıdan verilir altında olanlar silinir.

Şimdi tüm 2'li kombinasyonları yazarız.

$C_2 \rightarrow$

Özeler	Sayı
(ekmek, süt)	3
(ekmek, bez)	3
~~(ekmek, bira)~~	~~2~~
(süt, bez)	3
~~(süt, bira)~~	~~2~~
(bez, bira)	3
:)	
:)	
:)	
:)	

$C_3 \rightarrow$

Özeler	Sayı
(ekmek, süt, bez)	2
~~(ekmek, süt, bira)~~	
:)	

Destek

$C_1 + C_2 + C_3 =$

$6 + 15 + 20 = 41$

Yaygın sayı azalınca

$C_1 + C_2 + C_3$

$6 + 6 + 1 = 13$

DEMETLEME = KÜMELEME

* Nesneleri gruplara ayırma işlemidir. Aynı gruptaki nesneler birbirine daha çok benzer, farklı gruptaki nesneler birbirine daha az benzerdir.

* Eğitim verisi kullanılmaz, sınıflandırmadan farkı budur.

* Sınıflandırmada başta bilgi varken kümelemede yoktur.

* Gözetimsiz öğrenmeye girer, yani hangi nesnenin hangi sınıfa ait olduğu ve sınıf sayısı belli değil.

* Bu metod; görüntü işleme, ekonomi ve aykırılıkları belirleme gibi uygulamalarda kullanılır.

İyi bir demetleme yöntemi

Aynı demet içindeki nesneler arası fazla benzerlik olmalı.

Farklı demet içindeki nesneler arası az benzerlik olmalı.

Veri içindeki gizlenmiş örüntüleri bulabilmeli.

Gruplama için uygun demetleme kriteri bulunmalı.

Temel Demetleme Yaklaşımları

• Bölünmeli yöntemler → Veriyi bölerek, her grubu belirlenmiş ir kritere göre değerlendirir. (K-means)

• Hiyerarşik yöntemler → Veri kümelerini önceden belirlenmiş ir kritere göre hiyerarşik olarak ayırır.

• Yoğunluk tabanlı yöntemler → Nesnelerin yoğunluğuna göre emetleri oluşturur.

• Model tabanlı yöntemler → Her ~~mode~~ demetin bir modele yduğu varsayılır. Amaç bu modellere uyan verileri gruplamak.

K-Means Demetleme

Bilinen bir k değeri için k-means demetleme algoritması- in 4 aşaması vardır:

1- Veri kümesi k alt kümeye ayrılır.

2- Her demetin ortalaması hesaplanır: merkez nokta.

3- Her nesne en yakın merkez noktanın olduğu demete dahil dilir.

4- Nesnelerin demetlenmesinde değişiklik olmayana kadar dım 2'ye geri dönülür.

NOT:

* Karmaşıklığı → Yer karmaşıklığı (O((n+k).d))
Zaman " (O(k.t.n.d))

K-means Özellikleri

• Gerçeklemesi kolay

• Karmaşıklığı diğerlerine göre daha az

• K-means bazı durumlarda iyi sonuç vermeyebilir.

→ Veri grupları farklı boyutlarda ise

→ Veri gruplarının yoğunlukları farklı ise

→ Veri gruplarının şekli küresel değilse

→ Veri içinde aykırılıklar varsa en iyi sonuç vermeyebilir.

• Başlangıç şartlarına çok bağlıdır. Şartlar farklı olursa sonuçlarda farklı olur.

• Nesnelerin merkeze uzaklıkları ne kadar yakın ise çözüm o kadar iyidir.

• K-means'in kötü yanı tüm verilerin illaki bir kümeye dahil edilme şartıdır.

• Eğer veri sayısı kadar merkez olursa hata oranı en az olur ama bu istenen birşey değildir.

• Eğer tek küme elde edersek hata oranı en fazla olur.

Hata Oranı → Noktaların merkezlerine uzaklıklarını ölç, sonra hepsini topla bulduğun sonuç hata oranıdır.

* Teorene göre teker teker tüm sınıflara bakılır hangi sınıf değeri yüksek ise o sınıfa yerleştir.

$P(v_1, v_2, \dots v_n \mid S_j)$

↙ nitelikler ↳ Sınıf

ÖRNEK :

V_1	V_2	V_3	V_4	Sınıf
Yes	No	No Yes	Yes	B
Yes	No	No	No	B
No	Yes	Yes	No	M
No	No	Yes	Yes	M
Yes	No	No	Yes	B
Yes	No	No	No	M
Yes	Yes	Yes	No	M
Yes	Yes	No	Yes	M
No	No	No	Yes	B
No	No	Yes	No	M

Yandaki veri kümesi için sırasıyla gelen ⟨Yes, No, Yes, Yes⟩ örneğinin sınıfını bulalım.

Formül : $S_{hedef} = \text{argmax}\left[P(S_j)\prod_{i=1}^{n} P(v_i \mid S_j)\right]$

$P(S \mid \text{Yes, No, Yes, Yes}) = ?$

↳ tüm sınıfları teker teker dene !!

$$P(S=B \mid V_1, V_2, V_3, V_4) = \frac{4}{10}\cdot\left[\frac{3}{4}\cdot 1 \cdot \frac{1}{4}\cdot\frac{3}{4}\right] = \frac{9}{160}$$

$$P(S=M \mid V_1, V_2, V_3, V_4) = \frac{6}{10}\cdot\left[\frac{3}{6}\cdot\frac{3}{6}\cdot\frac{4}{6}\cdot\frac{2}{6}\right] = \frac{1}{30}$$

$\frac{9}{160} > \frac{1}{30}$ olduğu için bu veri B sınıfında derik..

* Eğer nitelikler sürekli değerler dursa yani evet-hayır değil de 1-2-3... gibi değerler olursa gauss dağılımını kullanırız.

ÖRNEK :

V_1	V_2	V_3	V_4	Yaş	Cinsiyet
Evet	Hayır	Evet	Hayır	38	K
Evet	Evet	Evet	Hayır	40	K
Evet	Evet	Evet	Hayır	41	K
Hayır	Hayır	Hayır	Hayır	55	K
Hayır	Evet	Hayır	Hayır	27	E
Evet	Evet	Evet	Evet	30	E
Evet	Hayır	Evet	Evet	35	E
Hayır	Hayır	Hayır	Hayır	42	E
Evet	Hayır	Hayır	Hayır	43	E
Evet	Hayır	Hayır	Hayır	45	E

Yandaki veri kümesi için

⟨Evet, Evet, Hayır, Hayır, 45⟩

Cinsiyet = ?

Normal Dağılıma göre $\Rightarrow P(v) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{1}{2}\left(\frac{v-\mu}{\sigma}\right)^2}$ (μ: ortalama, σ: standart sapma)

* $\mu = \frac{x_1 + x_2 + \dots + x_{10}}{10} = 39{,}6$

* $\sigma = \sqrt{\frac{1}{10}\cdot\left[(38-39{,}6)^2 + (40-39{,}6)^2 + \dots + (45-39{,}6)^2\right.}$

* $P(\text{Yaş} \mid \text{Cinsiyet}=k) =$ } ayrı ayrı hesaplanır Bulduğumuz değerleri;
$P(\text{Yaş} \mid \text{Cinsiyet}=e) =$ } buraya yazılır.

$$P(C=e \mid V_1, V_2, V_3, V_4, \text{Yaş}) = \frac{3}{5}\left(\frac{2}{3}\cdot\frac{1}{3}\cdot\frac{2}{3}\cdot\frac{2}{3}\cdot\frac{3}{100}\right) = 0{,}0014$$

* 0 problemi

Eğer sonuçlardan bir 0 dursa çarpım durumunda olduğu için hiç bir veri olamayız. Tüm olasılıkları 1 ile toplarız. Yani)

$V_1 = \frac{3}{5}$ $V_2 = \frac{0}{5}$ $V_3 = \frac{2}{5}$. olursa ;

$V_1 = \frac{4}{6}$ $V_2 = \frac{1}{6}$ $V_3 = \frac{3}{6}$ der işlem yaparız.

BAYES TEOREMİ

· Birbirinden bağımsız ve rastgele iki olayın birbiri ardı sıra gerçekleştiği durumlarda bu iki olayın birinin gerçekleşmesi durumunda ikinci olayın gerçekleşmesi olasılığı $P(A,B)$, $P(B,A)$ yada $P(A \cap B)$ ifadesiyle gösterilebilir.

$$P(A \cap B) = \frac{P(A/B)}{P(B)} = \frac{P(B/A)}{P(A)}$$

* Bayes teoremi rastgele bir A olayı ile diğer bir rastgele B olayı için koşullu olasılıklar ve marjinal olasılıklar arasındaki ilişkiyi tanımlar.

$$P(A/B) = \frac{P(B/A) \cdot P(A)}{P(B)}$$

* Naive Bayes Sınıflayıcı ile bayes teoremi hesaplarında dikkat edilmesi gereken en önemli fark sınıflayıcıların olasılık değerinden ziyade hedef sınıfı bulmaya odaklanmasıdır. Bu yüzden paydada bulunan değer, tüm hedef sınıflara ait olasılık hesaplarında ortak olduğunda ihmal edilebilir.

K-NN ÖRNEK

X1	X2	Y
2	4	Bad
3	6	good
3	4	good
4	10	bad
5	8	bad
6	3	good
7	9	good
9	7	bad
11	7	bad
10	2	bad

X1 = 8 ve X2 = 4 değerleri için k-nn ile hangi sınıfa ait olduğunu bulalım.. k=4

$d(i,j) = \sqrt{(2-8)^2+(4-4)^2} = 6$

$d(i,j) = \sqrt{(3-8)^2+(6-4)^2} = 5.38$

$d(i,j) = \sqrt{(3-8)^2+(4-4)^2} = 5$

(1) good

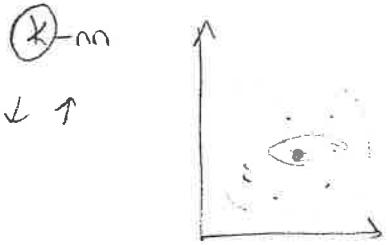
4

(3) bad

(2) bad

(1) bad

Hesaplamaları yaptıktan sonra en yakın 4 noktayı alıp tablodan karşılıklarına bakmalıyız. Good, Bad, Bad, Bad var diye BAD sınıfına dahil ederiz...



Algoritma Adımları

* K değeri dışardan istenir.

* Sınıfı bilinmeyen noktaya sınıflı veri noktalarının uzaklıkları hesaplanır. Genelde Öklid uzaklığı hesaplanır.

* Hesaplanan uzaklıklar sıralanır. En küçük uzaklığa sahip K adet sınıflı veri noktası tespit edilir.

* Seçilen K adet sınıflı veri noktası içerisinde çoğunluğa sahip sınıf belirlenir. Bu sınıf bilinmeyen nokta için sonuçtur

VERİ MADENCİLİĞİ

Büyük miktarda veri içerisinden gelecekle ilgili tahmin yapmamızı sağlayacak, yararlı modellerin oluşturulmasıdır. Bu modeller oluştuktan sonra aradığımız veriler mevcutsa bu verilerden çıkarabileceğimiz sonuçları anlamak için kullanılır.

Sınıflandırma ⟶ Riski az olan tüm kredi kartı başvurularını bul.

Kümeleme (Denetlesiz) ⟶ Harcama alışkanlığı benzer olan k.k. sahiplerini bul.

İlişkilendirme Kuralları ⟶ DVD ile birlikte sıkça satın alınan ürünü bul.

K-NN Algoritması

K en yakın komşuluk algoritması sorgu vektörünün en yakın k komşuluktaki vektör ile sınıflandırılmasının bir sonucu olan denetlemeli (kümelemeli) öğrenme algoritmasıdır.

Adımları:

* Yeni gelen birey sınıfa eklenir.
* K komşusuna bakılır.
* Öklid ile uzaklık hesaplanır.
* en yakın nerest ise birey oraya atanır

✓ $d(i,j) = \left[\sqrt{(x_{i1} - x_{j1})^2 \ldots\ldots}\right.$

ÖRNEK

*

Kayıt No	Yaş	Medeni Durum	Araba Sayısı
1	23	Bekar	1
2	25	Evli	1
3	29	Bekar	0
4	34	Evli	2
5	38	Evli	2
⋮			

Tablodan ilişkilendirme kuralları üretiniz.

① * Bekar bakmaz ; * yaş = 30-39 , medeni durum : evli ⇒ a.sayısı : 2 dir

② * Yaş : 29, Medeni durum : Bekar ⇒ Araba sayısı : 0

③ * Yaş : 30-39 ⇒ MedeniDurum : Evli

④ * arabasayısı : 0-1 ⇒ medeni durumu : bekar

① $\frac{(\text{Yaş}:30\text{-}39,\ \text{m.d}:\text{evli},\ \text{a.s}=2)}{(\text{yaş}:30\text{-}39,\ \text{m.d}:\text{evli})} = \frac{2}{2} = \%100$ → güven değer

$\frac{(\text{Yaş}:30\text{-}39,\ \text{m.d}:\text{evli},\ \text{a.s}=2)}{5} = \frac{2}{5} = \%40$ → destek değ

② $\frac{(\text{Yaş}:29,\ \text{m.d}:\text{bekar},\ \text{a.s}:0)}{(\text{Yaş}:29,\ \text{m.d}:\text{bekar})} = \frac{1}{1} = \%100$ → güven değeri

$\frac{(\text{Yaş}:29,\ \text{m.d}:\text{bekar},\ \text{a.s}=0)}{5} = \frac{1}{5} = \%20$ → destek d.

⋮

Nicel değer olan yaşı bölümleyelim; 20-29, 30-38 olsun:

Kayıt No	Yaş	M. durum	A. sayısı
1	23	Bekar	1
5	38	Evli	2

K.No	Yaş 20-29	Yaş 30-38	M.D evli	M.D bekar	Analoo 0	A. 1	A 2
1	1	0	0	1	0	1	0
5	0	1	1	0	0	0	1

Yaş. 20-29, MD. bekar $\Rightarrow A_1$

Yaş 30-38, MD. evli $\Rightarrow A_2$

ÖRNEK : Faturaların tutulduğu veri tabanı tablosundaki kayıtlar aşağıdaki şekildedir;

TID	Öğe
1	{1,2,3,4}
2	{1,2}
3	{2,3,4}
4	{2,3}
5	{1,2,4}
6	{3,4}
7	{2,4}

$\Rightarrow$

Ürün	Frekans
1	3
2	6
3	4
4	5

minsup = 3

$\Downarrow$

Ürün	Frekans
1,2	3
~~1,3~~	~~1~~
~~1,4~~	~~1~~
2,3	3
2,4	4
3,4	3

$\Leftarrow$

minsup değeri 3 olduğu için 3 lü değerleri bir değer çıkmamaktadır.

$1 \Rightarrow 2$; d.d $\rightarrow 3$

g.d $\rightarrow 3/3 = 1$

$2 \Rightarrow 1$; d.d $\rightarrow 3$

g.d $\rightarrow 3/6 = 0{,}5$

Ö	F
~~3,3,4~~	2

Kümeleme (Denetleme)

* eğitim verisi yok tüm verileri dışardan verilen k sayısına göre kümeleriz, k-means algoritması kullanılarak.
* Uzaklığa dayalı kümeleme

Sınıflandırma

* eğitim verisi kullanılarak test verisi uygun yere yerleştirilir k-nn algoritması kullanılarak.

* K-means dezavantajları

* Rastgele noktalar ile başlandığından elde edilen her sonuç kararlı değildir.
* Küme sayısı birçok kümeleme algoritmasında olduğu gibi dışardan verilir.
* Konkav olmayan küme şekillerinde başarılı değildir.
* Keskin kümeleme yapar.
* Aykırı değerlere karşı hassastır.
* Başlangıç noktasına bağımlıdır.

* K-NN dezavantajları

* Bellek tabanlı bir sınıflayıcıdır. Sınıflı veri noktaları sürekli hafızada tutulmalıdır. Veri kümesi çok büyük olduğunda hesaplama süresi de kötüleşir.
* Hesaplamaya tüm özellikler katıldığı için gereksiz veya ilgisiz özellikler sınıflandırmayı kötü yönde etkileyebilir.
* Sınıflandırma başarısı açısından genellikle yapay sinir ağları gibi gelişmiş sınıflandırma tekniklerinin gerisinde kalır.

K-MEANS

* Kümeleme yöntemidir.
* K → grup sayısı → dışardan verilir.
* kümeleme olduğu için eğitim verisi kullanılmaz, sınıflandırmadan farkı budur.
* Sınıflandırmada başta bilgi vardır kümelemede ise bilgi yoktur.

k-nn ← · → k-means eğitim verisi kullanılmaz.

K-NN

GINI İLE KARAR AĞACI ÇİZ

Sorcam dedi ama umarım şakadır. :)

C_1

Tid	Items
	1,2,5
	2,4
	2,3
	1,2,4
	1,3
	2,3
	1,3
	1,2,3,5
	1,2,3

$min_{sup} = 2$

$min_{conf.} = \%70$

Itemset	sup
1	6
2	7
3	6
4	2
5	2
	Şaka !!

L_1

C_2

Itemset	Sup
1,2	4
1,3	4
1,5	2
2,3	4
2,4	2
2,5	2

Itemset	Sup
(1,2)	4
(1,3)	4
~~(1,4)~~	~~1~~
(1,5)	2
(2,3)	4
(2,4)	2
(2,5)	2
~~(3,4)~~	~~0~~
~~(3,5)~~	~~1~~
(4,5)	0

C_3

Itemset	Sup
(1,2,3)	2
(1,2,5)	2

- $1,2 \rightarrow 5 \longrightarrow \frac{1,2,5}{1,2} \Rightarrow \frac{2}{4} \Rightarrow \%50 < \%70$
+ $1,5 \rightarrow 2 \longrightarrow \frac{1,5,2}{1,5} \Rightarrow \frac{2}{2} \Rightarrow \%100 > \%70$
+ $2,5 \rightarrow 1 \longrightarrow \frac{1,5,2}{2,5} \Rightarrow \frac{2}{2} \Rightarrow \%100 > \%70$

- $1,2 \Rightarrow 3 \longrightarrow \frac{1,2,3}{1,2} \Rightarrow \frac{2}{4} \Rightarrow \%50 < \%70$
- $1,3 \Rightarrow 2 \longrightarrow \frac{1,2,3}{1,3} \Rightarrow \frac{2}{4} \Rightarrow \%50 < \%70$
- $2,3 \Rightarrow 1 \longrightarrow \frac{1,2,3}{2,3} \Rightarrow \frac{2}{4} \Rightarrow \%50 < \%70$

"ÖRNEK 1

Gül, nur, ilk
ilk, can
nur, ilk
gül, ilk
gül, ilk, nur, can

⇒ Tamamı nitelik kümesi

gülün destek değeri ⇒ 3

gülün güven değeri ⇒ 3/5

*

TID	Öge	Sayı
1	Gül	3
2	nur	3
3	ilk	5
4	can	2

minsup = 3 dediğimiz anda "can" silinir

İkili kombinasyonlarını yazalım.

TID	Öge	Sayı
1	(gül, nur)	2
2	(gül, ilk)	3
3	(nur, ilk)	3

minsup = 3 'dü (gül, nur) gitti "

"ÖRNEK 2

TID	Items
10	A, C, D
20	B, C, E
30	A, B, C, E
40	B, E

Supmin = 2

C1

Itemset	sup
A	2
B	3
C	3
D	1
E	3

L1

Itemset	sup
A	2
B	3
C	3
E	3

C2

Itemset	sup
A, B	1
A, C	2
A, E	1
B, C	2
B, E	3
C, E	2

L2

Itemset	sup
A, C	2
B, C	2
B, E	3
C, E	2

C3

Itemset	sup
B, C, E	2

"

(1,2,5) için;

~~(1,2) ⇒ 5~~ → $\frac{(1,2,5)}{(1,2)} \Rightarrow \frac{2}{4} = \%50 < \%70$ (olmaz)

✓ (1,5) ⇒ 2 → $\frac{(1,2,5)}{(1,5)} \Rightarrow \frac{2}{2} = \%100 > \%70$ (olur.)

✓ (2,5) ⇒ 1 → $\frac{(1,2,5)}{(2,5)} \Rightarrow \frac{2}{2} = \%100 > \%70$ (olur.)

* Yani sonda elimizde kalan 2 kural var bunlar; (1,5) ⇒ 2 ve (2,5) ⇒ 1 'dir.

*Minsup değerini büyük seçersek veri kümesinden bazı örüntüler elde edilmeyebilir, önemli bilgi taşıyan veriler silinebilir

* Minsup değeri küçük seçersek yöntem karmaşıklaşır, çok fazla sayıda yaygın nitelik kümesi elde edilir

BİRLİKTELİK HAREKETİ

Soru :

TID	Items
1	1,2,5
2	2,4
3	2,3
4	1,2,4
5	1,3
6	2,3
7	1,3
8	1,2,3,5
9	1,2,3

Tablodan $min_{sup} = 2$ ve $min_{conf} = \%70$ değerleri için kurallar elde ediniz.

Not => Burada 2'nin ;

destek değeri = ~~7~~ → kaç satırda geçtiği

güven değeri = 7/9 → $\frac{\text{destek değeri}}{\text{tamamı}}$

Çözüm :

Itemset	sup
1	6
2	7
3	6
4	2
5	2

→ 1'in tabloda kaç defa geçtiği.

⇒

Itemseti 2'li ler için oluşturalım ;

Itemset	sup
(1,2)	4
(1,3)	4
~~(1,4)~~	1
(1,5)	2
(2,3)	4
(2,4)	2
(2,5)	2
~~(3,4)~~	0
~~(3,5)~~	1
~~(4,5)~~	0

support değeri 2'nin altında olanları eleyelim..
Ve buradan 3'lü oluşturalım.

⇐

Itemset	sup	
* (1,2,3)	2	✓
* (1,2,5)	2	✓

3'lü ye sadece bu kadar kaldı.

Çünkü (1,2,3) şartı sağlıyor

yani (1), (2), (3), (1,2), (1,3), (2,3)'ü silmedik

o yüzden (1,2,3) oluşturabiliriz. Şimdi kuralları incelelim..

~~(1,2,3)~~ ⇒ ~~(1,2)~~ ⇒ 3 → $\frac{(1,2,3)}{(1,2)} = \frac{2}{4} = \%50$

(ilk tablodan 1,2,3'ün birlikte olduğu durumlara bak.)

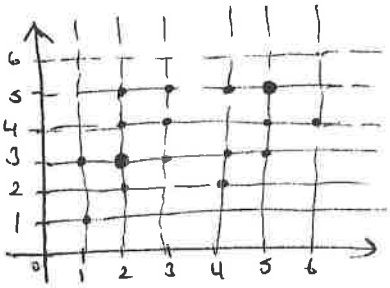
(ilk tablodan 1,2'nin birlikte olduğu durum.)

~~(1,3)~~ ⇒ 2 → $\frac{(1,2,3)}{(1,3)} = \frac{2}{4} = \%50$

(2,3) ⇒ 1 → $\frac{(1,2,3)}{\quad} = \frac{2}{4} = \%50$

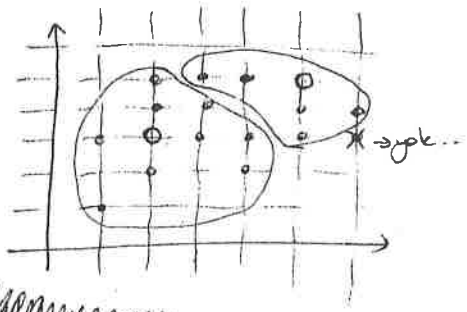
* Hepsi min_{conf} değerinin altında kalıyor, bu kuralları oluşturamayız.

K-means Örnek



k=2 olsun.

Rastgele 2 nokta koydum her nesneyi en yakın gruba dahil edelim



* Of veri madenciliği güzel alansın ama kötü derssin "

Son

Tablo 1

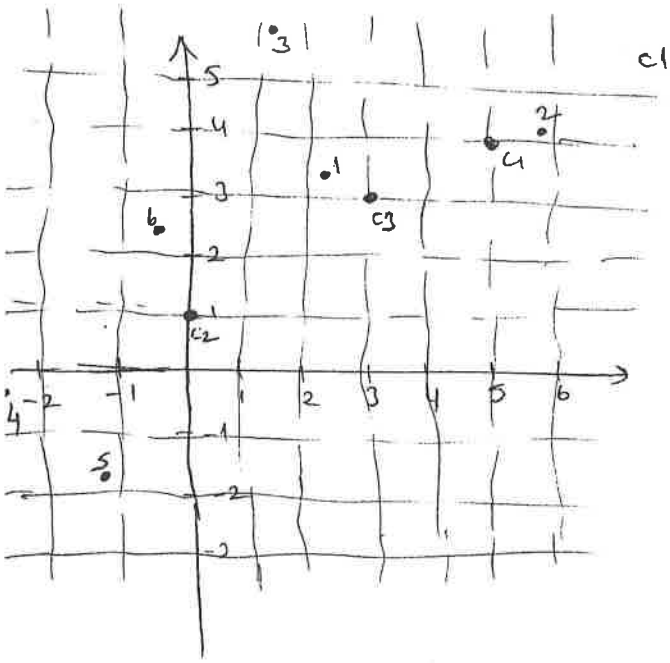
	X_1	X_2	D
1 →	2,3	3,4	a
2 →	5,3	4,2	a
3 →	1,1	5,7	a
4 →	-3,1	-0,3	b
5 →	-1,1	-1,5	b
6 →	-0,5	2,3	b

K-means yöntemiyle 3 küme merkezi seçiyoruz:

$c_1 = (5,4)$

$c_2 = (0,1)$

$c_3 = (3,3)$



1. verinin kümesini bulalım;

$c1: \sqrt{(5-2,3)^2 + (4-3,4)^2} = \sqrt{7,29 + 0,36}$

$c1 = 2,765$

~~$c2: \sqrt{(0-2,3)^2 + (1-3,4)^2} = \sqrt{5,29 + 5,76}$~~

$c2 = 6,711$ → 2. veri

$c1: \sqrt{(0-2,3)^2 + (1-3,4)^2} = 5,29 + 5,76$

$c2 = 3,324$

$c3 = \sqrt{0,49 + 1,44} = 1,389$

Devam

Toplam maliyeti hesapla = ?

Noktalardan merkezleri alıcı

Soru

Tablo 1'e göre K-nn yöntemiyle sınıflandırma yapılacaktır. Buna göre sınıfı bilinmeyen (1,2) veri noktasının K=1, K=3, K=5 için sınıflarını bularak sınıflandırma hakkında yorumunuzu yazınız.

Soru

X_1	X_2	X_3	D
K	P	G	H
K	N	G	H
B	E	G	H
B	P	G	S
B	N	T	S
K	E	T	S

Naive Bayes sınıflandırıcıya göre (K,P,T) ve (B,P,T) veri noktalarının sınıflarını bulunuz.

$$P(S=H \mid K,P,T) = \frac{3}{6}\left[\frac{3}{4} \cdot \frac{2}{4} \cdot \frac{0}{1}\right]$$

$$= 3/16$$

$$P(S=S \mid K,P,T) = \frac{3}{6}\left[\frac{1}{3} \cdot \frac{1}{3} \cdot \frac{1}{3}\right]$$

$$1/27$$

$$3/16 > 1/27$$