

Veri Madenciliği

Güz 2023

Ders 13

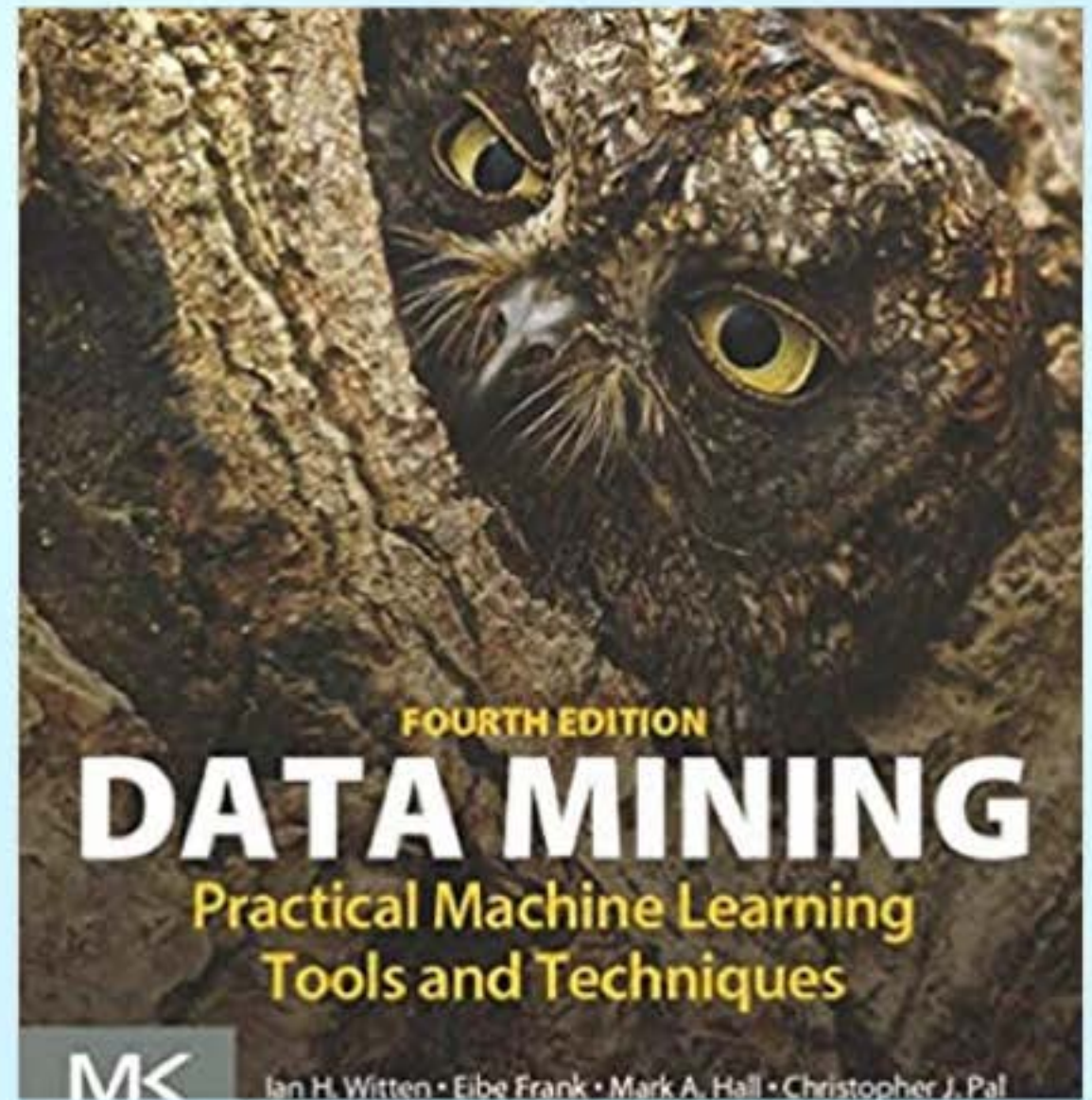
Algoritmalar–Temel Kurallar

1R–Sayısal veriler

Bayes

Dersin Kitabı

- Data Mining: Practical Machine Learning Tools and Techniques, 4th Ed., by Ian Witten, Eibe Frank, Mark Hall, and Christopher Pal (Morgan Kaufmann Publishers, 2017. ISBN: 978-0-12-804291-5)



İlkel Kuralların Çıkarılması

- **1R:** 1 seviyeli bir karar "**ağacı**" öğrenir
 - yani, hepsinin belirli bir özelliği test ettiği kurallar
- **Temel fikir**
 - Her değer için bir dal oluşturun
 - Her dal en sık görülen sınıfı atar
 - **Hata oranı:** karşılık gelen dallarının çoğunluk sınıfına ait olmayan örneklerinin oranı
 - En düşük hata oranına sahip öz niteliği seçin
-Nominal nitelikler olduğunu varsayalım

Sayısal Özniteliklerle Başa Çıkma

Sayısal verileri dönüştürmeniz gerekiyor.

- Sayısal öznitelikleri ayırıştır
- Her özelliğin aralığını aralıklara bölün
 - Örnekleri özniteliliğin değerlerine göre sıralayın
 - Sınıfın değiştiği yerlere kesme Hayırlıkları yerleştirin (çoğunluk sınıfı)
 - Bu toplam hatayı en aza indirir

Sayısal Hava Durumu Kümesini Geri Çağırın

Görünüm	Sıcaklık	Nem	Rüzgarlı	Oyun
Güneşli	85	85	Yanlış	Hayır
Güneşli	80	90	Doğru	Hayır
Bulutlu	83	86	Yanlış	Evet
Yağmurlu	70	96	Yanlış	Evet
Yağmurlu	68	80	Yanlış	Evet
Yağmurlu	65	70	Doğru	Hayır
Bulutlu	64	65	Doğru	Evet
Güneşli	72	95	Yanlış	Hayır
Güneşli	69	70	Yanlış	Evet
Yağmurlu	75	80	Yanlış	Evet
Güneşli	75	70	Doğru	Evet
Bulutlu	72	90	Doğru	Evet
Bulutlu	81	75	Yanlış	Evet
Yağmurlu	71	91	Doğru	Hayır

Örnek: sıcaklık

64	65	68	69	70	71	72	72	75	75	80	81	83	85
Evet	Hayır	Evet	Evet	Evet	Hayır	Hayır	Evet	Evet	Evet	Evet	Evet	Hayır	Hayır

64	65	68	69	70	71	72	72	75	75	80	81	83	85
Evet	Hayır	Evet	Evet	Evet	Hayır	Hayır	Evet	Evet	Evet	Evet	Evet	Hayır	Hayır

Kırılma-noktaları

$\leq 64.5 \rightarrow$ evet
 > 64.5 ve $\leq 66.5 \rightarrow$ hayır
 > 66.5 ve $\leq 70.5 \rightarrow$ evet
 > 70.5 ve $\leq 73.5 \rightarrow$ hayır
 > 73.5 ve $\leq 77.5 \rightarrow$ evet
 > 77.5 ve $\leq 80.5 \rightarrow$ hayır
 > 80.5 ve $\leq 84 \rightarrow$ evet
 $> 84 \rightarrow$ hayır

Aşırı Uyum Sorunu

- Bu prosedür gürültüye karşı çok hassastır
 - Yanlış sınıf etiketine sahip bir örnek büyük olasılıkla ayrı bir aralık üretecektir
- Basit gözüm: çoğunluk sınıfında aralık başına minimum örnek sayısını zorunlu kılın

Örnek (min = 3 ile):

64	65	68	69	70	71	72	72	75	75	80	81	83	85
Evet	Hayır	Evet	Evet	Evet	Hayır	Hayır	Evet	Evet	Evet	Evet	Evet	Hayır	Hayır

64	65	68	69	70	71	72	72	75	75	80	81	83	85
Evet	Hayır	Evet	Evet	Evet	Hayır	Hayır	Evet	Evet	Evet	Evet	Evet	Hayır	Hayır

Örnek: nem

65	70	70	70	75	80	80	85	86	90	90	91	95	96	
Evet	Hayır	Evet	Evet	Evet	Evet	Evet	Evet	Hayır	Hayır	Evet	Hayır	Hayır	Evet	Evet

65	70	70	70	75	80	80	85	86	90	90	91	95	96	
Evet	Hayır	Evet	Evet	Evet	Evet	Evet	Evet	Hayır	Hayır	Evet	Hayır	Hayır	Evet	Evet

*Daha fazla birleştirmeden önce kırılma-
noktaları*

$\leq 67.5 \rightarrow$ evet

> 67.5 ve $\leq 82.5 \rightarrow$ Evet

> 82.5 ve $\leq 85.5 \rightarrow$ hayır

> 85.5 ve $\leq 88.5 \rightarrow$ Evet

> 88.5 ve $\leq 95.5 \rightarrow$ hayır

(evet)

> 90.5 ve $\leq 95.5 \rightarrow$ hayır

$> 95.5 \rightarrow$ evet

Devam: nem

65	70	70	70	75	80	80	85	86	90	90	91	95	96	
Evet	Hayır	Evet	Evet	Evet	Evet	Evet	Evet	Hayır	Hayır	Evet	Hayır	Hayır	Evet	Evet

65	70	70	70	75	80	80	85	86	90	90	91	95	96	
Evet	Hayır	Evet	Evet	Evet	Evet	Evet	Evet	Hayır	Hayır	Evet	Hayır	Hayır	Evet	Evet

Aşırı Uyum Önleme ile, yani çoğunluk sınıfında minimum örnek sayısı

Öznitelik	Kurallar	Hata	Toplam Hata
Görünüm	Güneşli -> Hayır	2/5	4/14
	Bulutlu -> Evet	0/4	
	Yağmurlu -> Evet	2/5	
Sıcaklık	≤ 77.5 -> evet	3/10	5/14
	> 77.5 -> hayır	2/4	
Nem	≤ 82.5 -> evet	1/7	3/14
	> 82.5 ve ≤ 95.5	2/6	
	-> hayır	0/1	
	> 95.5 -> evet		
Rüzgârlı	Yanlış -> Evet	2/8	5/14
	Doğru -> Hayır*	3/6	

1R Tartışması

1R Holte tarafından bir makalede tanımlanmıştır (1993)

Çok Basit Sınıflandırma Kuralları En Sık Kullanılan Veri Kümelerinde İyi Performans Gösteriyor

Robert C. Holte, Bilgisayar Bilimleri Bölümü, Ottawa Üniversitesi

- 16 veri kümesi üzerinde deneysel bir değerlendirme içerir (sonuçların gelecekteki veriler üzerindeki performansı temsil etmesi için çapraz doğrulama kullanılarak)
- Aynı sınıf değerine sahip minimum örnek sayısı, bazı deneylerden sonra 6'ya (bizim örneğimiziz 3'ü kullandı) ayarlandı
- 1R'nin basit kuralları, çok daha karmaşık karar veren öğrencilerden çok daha kötü performans göstermedi

Sadelik önce karşılığını verir!

İstatistiksel Modelleme

- 1R'nin "Zitti": tüm özellikleri kullanın
- **İki varsayım:** öznelitlikler
 - eşit derecede önemli
 - istatistiksel olarak bağımsız (sınıf değeri göz önüne alındığında)
 - yani, bir özelliğin değerini bilmek, diğerinin değeri hakkında hiçbir şey söylemez (eğer sınıf biliniyorsa)
- Ama ... bu plan pratikte iyi çalışır

Hava Durumu Verileri için Olasılıklar

Görünüm			Sıcaklık			Nem			Rüzgârlı			Oynamak	
Evet Hayır			Evet Hayır			Evet Hayır			Evet Hayır			Evet Hayır	
Güneşli	2	3	Sıcak	2	2	Yüksek	3	4	Yanlış	6	2	9	5
Bulutlu	4	0	Ilıman	4	2	Normal	6	1	Doğru	3	3		
Yağmurlu	3	2	Serin	3	1								
Güneşli	2/9	3/5	Sıcak	2/9	2/5	Yüksek	3/9	4/5	Yanlış	6/9	2/5	9/14	5/14
Bulutlu	4/9	0/5	Ilıman	4/9	2/5	Normal	6/9	1/5	Doğru	3/9	3/5		
Yağmurlu	3/9	2/5	Serin	3/9	1/5								

Görünüm	Sıcaklık	Nem	Rüzgarlı	Oyun
Güneşli	Sıcak	Yüksek	Yanlış	Hayır
Güneşli	Sıcak	Yüksek	Doğru	Hayır
Bulutlu	Sıcak	Yüksek	Yanlış	Evet
Yağmurlu	Ilıman	Yüksek	Yanlış	Evet
Yağmurlu	Serin	Normal	Yanlış	Evet
Yağmurlu	Serin	Normal	Doğru	Hayır
Bulutlu	Serin	Normal	Doğru	Evet
Güneşli	Ilıman	Yüksek	Yanlış	Hayır
Güneşli	Serin	Normal	Yanlış	Evet
Yağmurlu	Ilıman	Normal	Yanlış	Evet
Güneşli	Ilıman	Normal	Doğru	Evet
Bulutlu	Ilıman	Yüksek	Doğru	Evet
Bulutlu	Sıcak	Normal	Yanlış	Evet
Yağmurlu	Ilıman	Yüksek	Doğru	Hayır

Olasılıklar devam

Görünüm	Sıcaklık	Nem	Rüzgarlı	Oyun
Güneşli	Serin	Yüksek	Doğru	?

"Evet" olasılığı "Evet" =

"Hayır" olasılığı =

Olasılıklar devam

$P(\text{"Evet"}) =$

$P(\text{"Hayır"}) =$

Bayes'in Kuralı

Thomas Bayes
Born: 1702 in London,
England
Died: 1761 in Tunbridge
Wells, Kent, England

E kanıtı verildiğinde H olayının olasılığı :

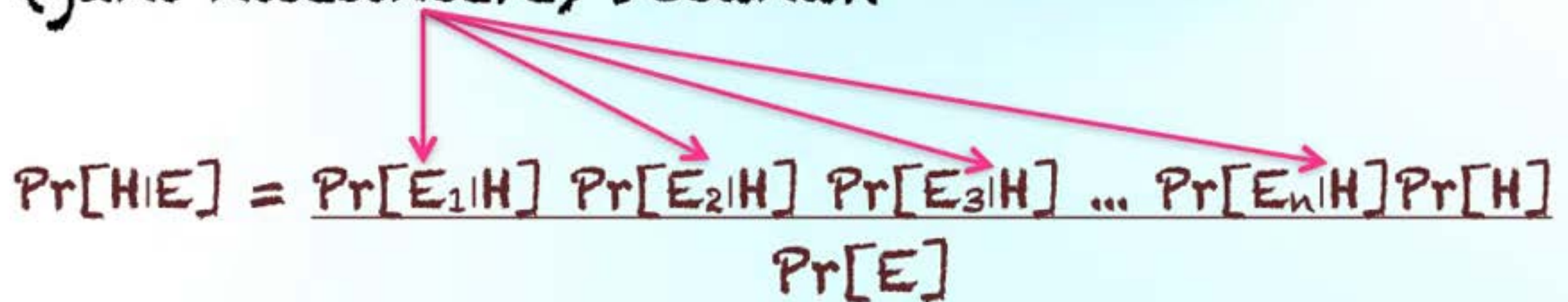
$$\Pr[H|E] = \frac{\Pr[E|H] \Pr[H]}{\Pr[E]}$$

- H'nin önceki olasılığı :
 - Kanıt görülmeden önce olayın olma olasılığı
- H'nin sonraki olasılığı :
 - Kanıt görüldükten sonra olayın olma olasılığı

Sınıflandırma için Naif Bayes

Sınıflandırma öğrenimi: Bir örnek verilen sınıfın olasılığı nedir?

- Kanıt E = örnek
- Olay H = örneğin sınıf değeri
- Naif varsayım: kanıt, bağımsız olan parçalara (yani niteliklere) bölünür.


$$\Pr[H|E] = \frac{\Pr[E_1|H] \Pr[E_2|H] \Pr[E_3|H] \dots \Pr[E_n|H] \Pr[H]}{\Pr[E]}$$

Hava Durumu Verileri Örneği

Görünüm	Sıcaklık	Nem	Rüzgarlı	Oyun
Güneşli	Serin	Yüksek	Doğru	?

$$\begin{aligned} \Pr[\text{Evet}|E] &= \Pr[\text{Görünüm=Güneşli}|E] \\ &\times \Pr[\text{Sıcaklık=Serin}|E] \\ &\times \Pr[\text{Nem=Yüksek}|E] \\ &\times \Pr[\text{Rüzgarlı=Doğru}|E] \\ &\times \Pr[\text{Evet}] \\ &\hline &\Pr[E] \\ &= \\ &\hline &\Pr[E] \end{aligned}$$

"Sıfır Frekans Sorunu"

- Her sınıf değerinde bir öznelite değeri oluşmazsa ne olur?
(örneğin "Görünüm = Bulutlu" sınıfı için "Hayır")
 - Olasılık sıfır olacak!
 - Bir posteriori olasılığı da sıfır olacak! (Hayır, diğer değerlerin ne kadar olası olduğu önemli!)
- **Çare:** her öznelite değeri sınıfı kombinasyonu için sayıma 1 ekleyin (Laplace estimator)
- **Sonuç:** olasılıklar asla sıfır olmayacak!

Laplace Estimator örneği

Görünüm	Sıcaklık	Nem	Rüzgarlı	Oyun
Bulutlu	Serin	Yüksek	Doğru	?

"Hayır" olasılığı = $0/5 * \dots\dots\dots = 0$

WEKA, yalnızca rahatsız edici nitelik veya belirli bir sınıf için değil, tüm niteliklere 1 ekler

Görünüm	.	
	Evet	Hayır
Güneşli	2	3
Bulutlu	4	0
Yağmurlu	3	2
Güneşli		
Bulutlu		
Yağmurlu		

"Hayır" olasılığı = $1/8 * \dots\dots\dots = \dots$

Değiştirilmiş Olasılık Tahminleri

- Bazı durumlarda 1'den farklı bir sabit eklemek daha uygun olabilir.
 - Örnek: **Evet** sınıfı için **Bulutlu** öznelilik (eşit ağırlıklar, yani $p = 1/3$ 'tür)

Güneşli

$$\frac{2+Y/3}{9+Y}$$

Bulutlu

$$\frac{4+Y/3}{9+Y}$$

Yağmurlu

$$\frac{3+Y/3}{9+Y}$$

- Ağırlıkların eşit olması gerekmez (ancak toplamlarının 1 olması gerekir). $p_1 + p_2 + p_3 = 1$

$$\frac{2+Yp_1}{9+Y}$$

$$\frac{4+Yp_2}{9+Y}$$

$$\frac{3+Yp_3}{9+Y}$$

Eksik Değerler

Eğitim: örnek, özneliteik deęer-sınıf kombinasyonu için sıklık sayımına dahil edilmez

Sınıflandırma: eksik özneliteiğı hesaplamadan çıkar

Örnek:

Görünüm	Sıcaklık	Nem	Rüzgarlı	Oyun
?	Serin	Yüksek	Doęru	?

$$\text{"Evet" olasılığı} = 3/9 * 3/9 * 3/9 * 9/14 = 0.0238$$

$$\text{"Hayır" olasılığı} = 1/5 * 4/5 * 3/5 * 5/14 = 0.0343$$

$$P(\text{"Evet"}) =$$

$$P(\text{"Hayır"}) =$$