Veri Madenciliği

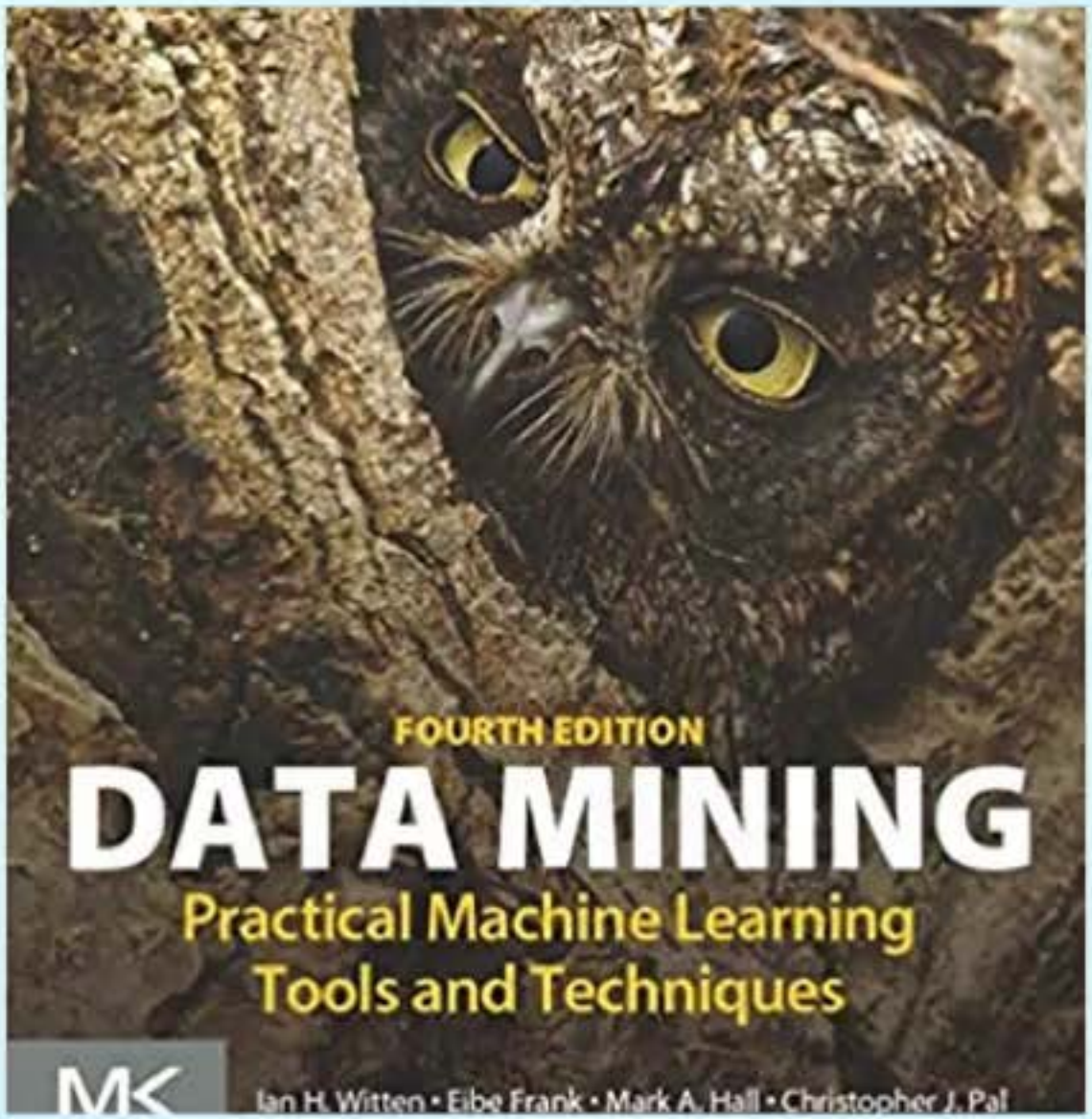
Güz 2023

Ders 6

- Çıktı:
- Tablolar, Doğrusal Modeller, Ağaçlar

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)



Çıktı:

Yapısal Desenleri Temsil Etme

- Kalıpları temsil etmenin birçok farklı yolu
  - Karar ağaçları, kurallar, örnek tabanlı, ...
- "Bilgi" gösterimi olarak da adlandırılır
- Gösterim çıkarım yöntemini belirler
- Çıktıyı anlamak, temel öğrenme yöntemlerini anlamanın anahtarıdır
- Farklı öğrenme sorunları için farklı çıktı türleri (örn. sınıflandırma, gerileme, ...)

Tablo

Çıktıyı temsil etmenin en basit yolu:

- **Girdiyle aynı biçimi kullanın!**

Sadece uygun koşullara sahip satırı bulun ve sınıfı atayın, bu durumda oynayın veya oynamayın.

Hava durumu problemi için karar tablosu:

```
Görünüm Sıcaklık Nemi Rüzgarlı Oyun
Güneşli Sıcak Yüksek Yanlış Hayır
Güneşli Sıcak Yüksek Doğru Hayır
Bulutlu Sıcak Yüksek Yanlış Evet
Yağmurlu Hafif Yüksek Yanlış Evet
Yağmurlu Serin Normal Yanlış Evet
Yağmurlu Serin Normal Doğru Hayır
Bulutlu Serin Normal Doğru Evet
.......
```

Sayısal tahmin ise, kavram aynıdır, ancak karar tablosu olarak adlandırmak yerine, regresyon tablosu olarak adlandırılır.

Tablo

Bazen bazı özellikler karar için gerekli değildir.

- Ya sıcaklığa ve rüzgarlı özelliklere ihtiyacımız yoksa?

Daha küçük, yoğunlaştırılmış bir tablo daha iyi olabilir:

Görünüm	Nem	Oyun
Güneşli	Yüksek	Hayır
Güneşli	Normal	Evet
Bulutlu	Yüksek	Evet
Bulutlu	Normal	Evet
Yağmurlu	Yüksek	Hayır
Yağmurlu	Normal	Hayır

Temel sorun: doğru kararı vermek için doğru nitelikleri seçme.

Başka Bir Basit Gösterim: Doğrusal Modeller

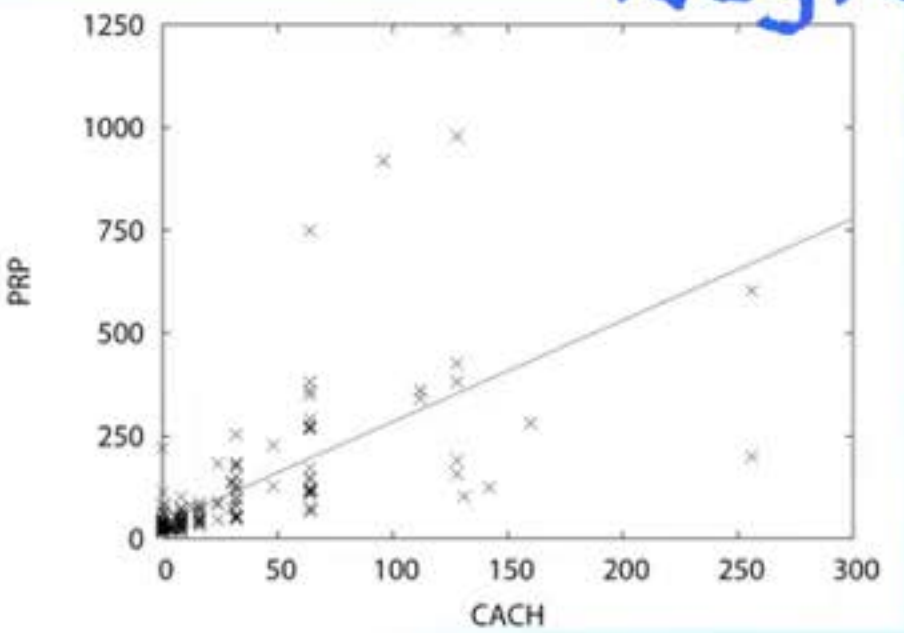
Regresyon modeli

- Tüm girişler (öznitelik değerleri) ve çıktı sayısal olduğunda kullanılır

Çıktı, ağırlıklı öznitelik değerlerinin toplamıdır

- Işin püf noktası ağırlıklar (weigths) için iyi değerler bulmaktır.

CPU Performans Verileri için Doğrusal Regresyon Fonksiyonu



CPU performansını tahmin etmek için yalnızca buradaki önbellek özniteliği kullanılır.

(İki boyutta görmek daha kolaydır.)

PRP = 37.06 + 2.47CACH

37.06 "önyargı" terimidir ve 2.47 önbellek ağırlığı gibi bir ağırlıktır. Ağırlıkları bulmak için en az kareler doğrusal regresyon yöntemi kullanılmıştır.

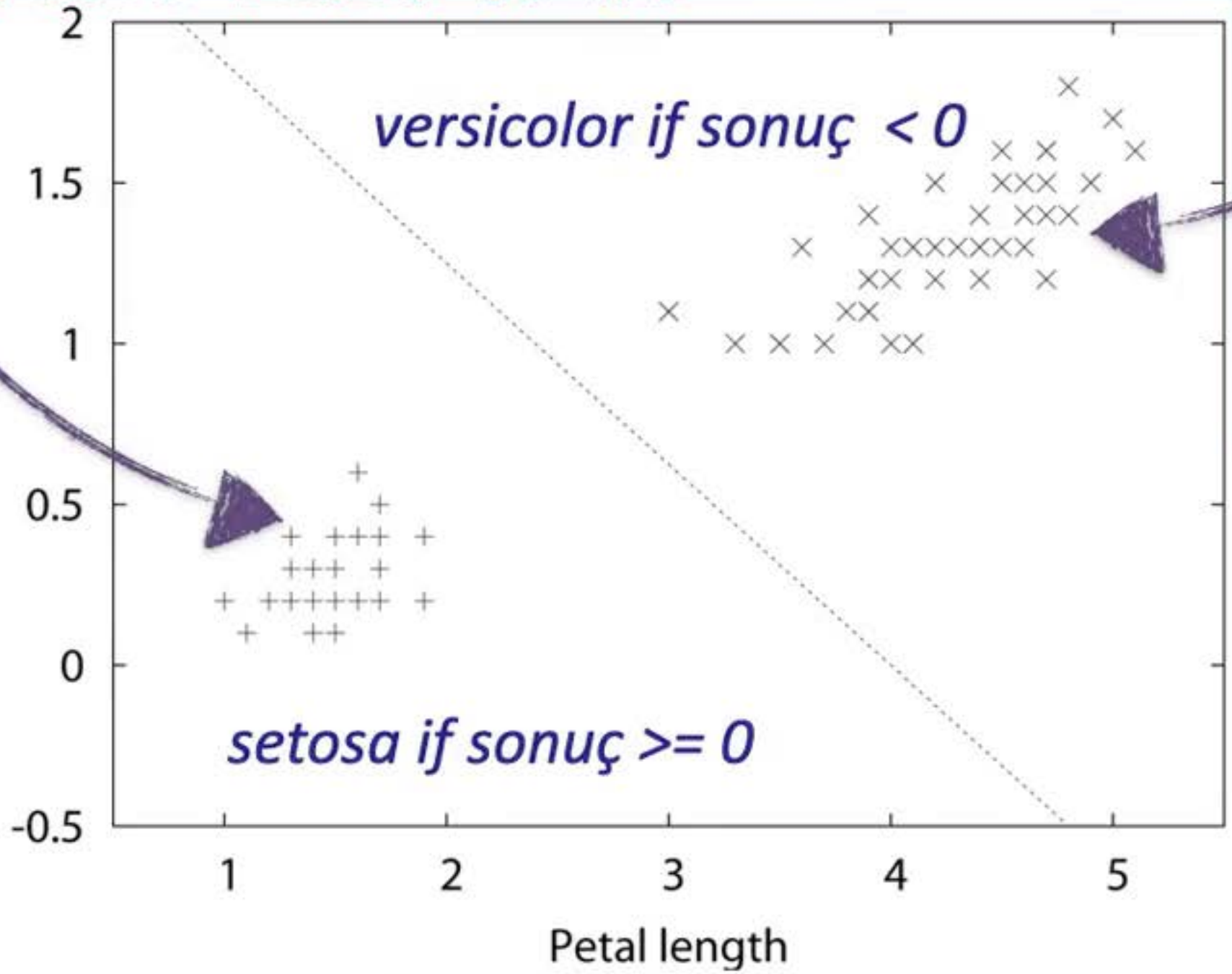
Eğitim verileri ağırlıkları bulmak için kullanılır.

Bir test örneği verildiğinde, önbellek özniteliğinin değerini ifadeye girin ve performansın değeri (çıktı/sınıf) satırda olacaktır.

İkili Sınıflandırma için Doğrusal Modeller

- Satır iki sınıfı ayırır
  - Karar sınırı – kararın bir sınıf değerinden diğerine nerede değiştiğini tanımlar
- Tahmin, özniteliklerin gözlemlenen değerlerinin ifadeye eklenmesi ile yapılır
  - If Çıktı >= 0 ise bir sınıfı, if 0 < Çıktı ise diğer sınıfı tahmin et
- Birden çok öznitelik olduğunda sınır yüksek boyutlu düzlem (hiper düzlem) olur

İris Setosas'ı Iris Versicolors'dan Ayıran Doğrusal Karar Sınırı



2.0 - 0.5PetalLength - 0.8PetalWidth = 0

Ağaç

- "Böl ve fethet" yaklaşımı ağaç üretir
- Düğümler belirli bir özniteliği sınama içerir

  Genellikle, öznitelik değeri sabitle karşılaştırılmıştır

  - Diğer olasılıklar:
  - İki özniteliğin değerlerini karşılaştırma
  - Bir veya daha fazla özniteliğin fonksiyonunu kullanma

Seçenek düğümleri (birden fazla dal seçimi), örn: bir örnek iki (veya daha fazla) yaprak açar. Bundan sonra alternatif tahminler bir şekilde birleştirilmelidir (çoğunluk oylaması).

- Yapraklar sınıflandırma, sınıflandırma kümesi veya örneklere olasılık dağılımı atar.
- Bilinmeyen örnek ağaçtan aşağı yönlendirilir.

Nominal ve Sayısal Öznitelikler

Nominal:

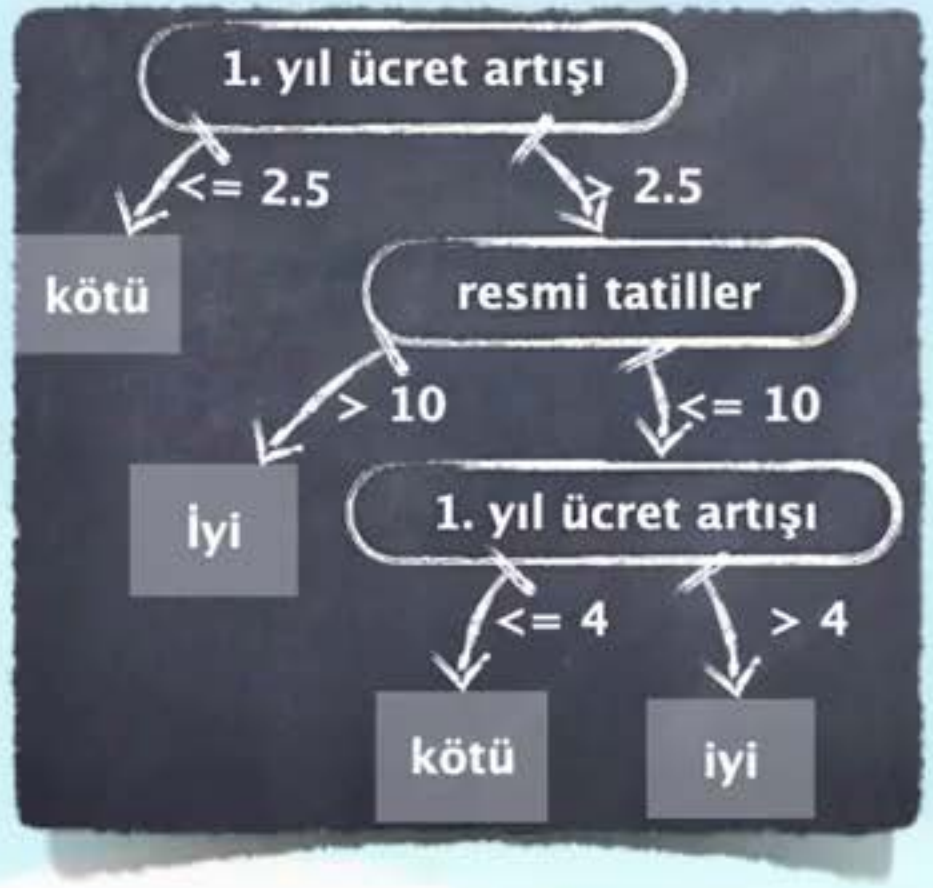
- Genellikle sayı değerlerine eşit olan çocuk sayısı ==> özniteliği birden çok kez sınanamaz

Diğer olasılık: iki alt kümeye bölmek, bu sayede birden fazla kez test edilebilir

Nominal ve Sayısal Öznitelikler

Sayısal:

- değerin sabitten büyük mü yoksa küçük mü olduğunu sınayın ==> öznitelik birkaç kez sınanabilir



Diğer olasılık: üç yönlü bölme (veya çok yönlü bölme)

- Tam sayı: *küçüktür, eşit, büyüktür*
- Reel: *altında, içinde (yani eşit olacak kadar yakın), üstünde*

Eksik Değerler

Değer yokluğunun bir önemi var mı?

- Evet ==> "eksik" ayrı bir değerdir
- Hayır ==> "eksik" özel bir şekilde ele alınmalıdır

- **Çözüm A**: en popüler dala örnek atama
- **Çözüm B**: örneği parçalara bölme
  - Parçalar, her daldan aşağı inen eğitim örneklerinin kesirine göre ağırlık alır
  - Yaprak düğümlerinden sınıflandırmalar, kendilerine filtre uygulanmış ağırlıklar kullanılarak birleştirilir


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