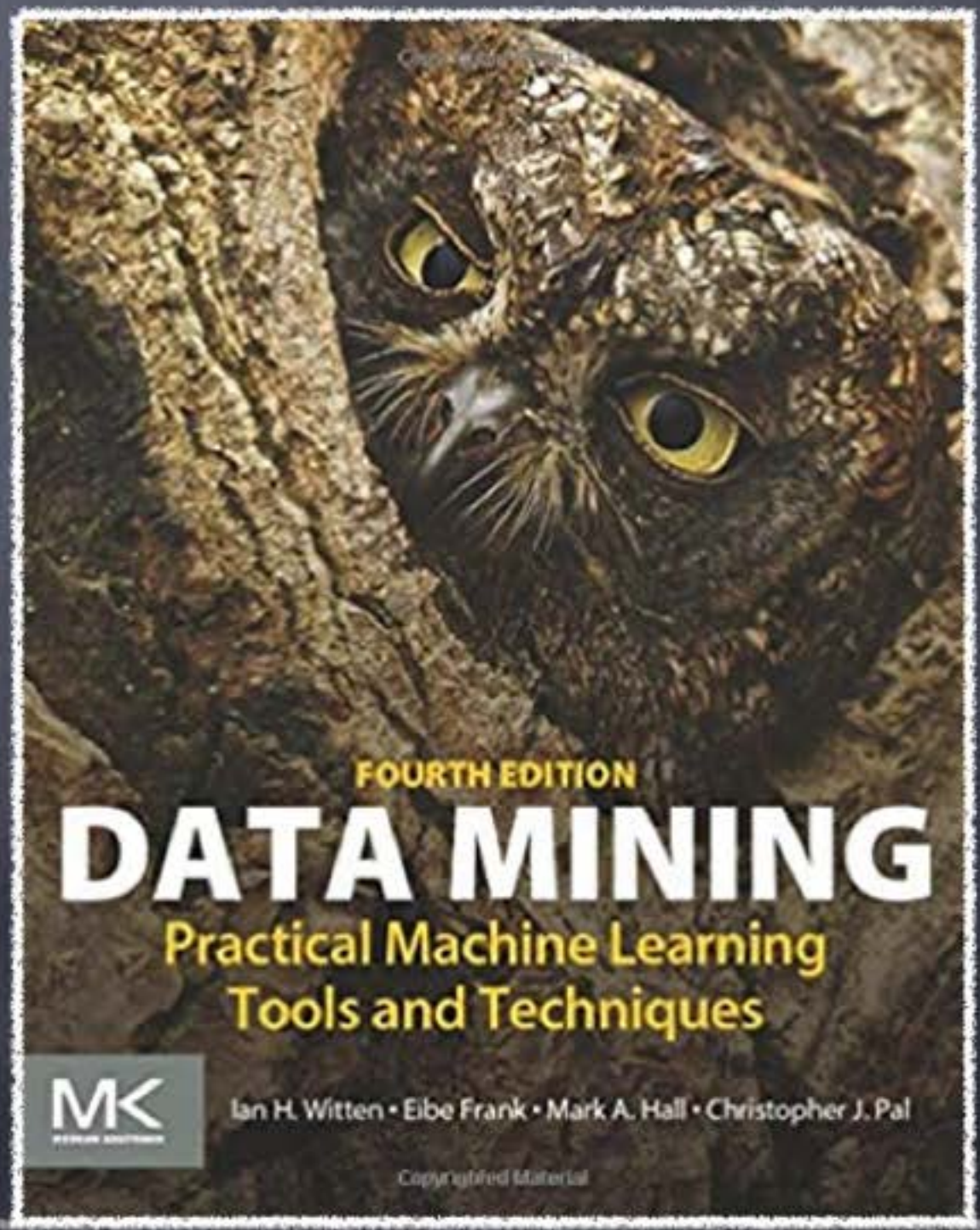
Veri Madenciliği

Güz 2023
Ders 3

Saha Uygulamaları
Etik

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)



Saha Uygulamaları

Öğrenmenin sonucu - veya öğrenme yönteminin kendisi - pratik uygulamalarda kullanılır

- Kredi başvurularının işlenmesi
- Petrol birikintileri için tarama görüntüleri
- Elektrik arzı tahmini
- Makine arızalarının teşhisi
- Pazarlama ve satış (Büyük $$)
- Web Madenciliği (PageRank, sorgu içeriği madenciliği vb.)

Bunlar, araştırma problemlerine veya bu derste inceleyeceğimiz basit problemlere karşı kullanıma sunulan gerçek ML problemleridir.

Saha Uygulamaları

- Ham petrol ve doğal gazın ayrılması (parametrelerin belirlenmesinde kullanılan kurallar artık bir gün yerine 10 dakika sürüyor)
- Rotogravür baskıda bant oluşumunun azaltılması (insan eliyle yapılcak işin 500'den 30'a düşürülmesi)
- Telefon arızaları için uygun teknisyenleri bulma (10 milyon $ tasarruf sağlamıştır)
- Bilimsel uygulamalar: biyoloji, astronomi, kimya (hücre yapısını analiz etmek, ilaç keşfini hızlandırır, gök cisimlerini kataloglamak vb.)
- TV programlarının otomatik seçimi
- Yoğun bakım hastalarının takibi

Kredi Başvurularının İşlenmesi (American Express)

Karar vermek

Yargı içeren kararlar.

Verilen: Finansal ve kişisel bilgilerle soruşturma

Soru: Kredi verilmeli mi?
Basit istatistiksel yöntem, kredi görevlilerine gönderilen sınırdaki vakaların (diğer kalanlar %10) %90'ını kapsar

Ancak: Kabul edilen sınırda vakaların %50'si temerrüde düşmüştür!

Çözüm: Tüm sınır vakaları reddedilsin mi?
Hayır! Sınırdaki vakalar en aktif müşterilerdir....

Makine Öğrenimine Giriş

Sınırda vakalar için 1000 eğitim örneği

20 özelliği:

- yaş
- mevcut işyerindeki geçirdiği yıllar
- şimdiki adreste geçirdiği yıllar
- banka ile çalıştığı yıllar
- Sahip olunan diğer kredi kartları, ...

Öğrenilmiş Kurallar: vakaların %67'sinde doğru

Uzman İnsanlar: vakaların yalnızca %50'sini düzeltmiştir

Kararları müşterilere açıklamak için kurallar kullanılabilir

Nitelikleri üret

Görüntü Tarama

Verilen: kıyı sularının radar uydu görüntüleri
Sorun: bu görüntülerdeki Petrol birikintilerini tespit edilmesi

Petrol birikintileri, değişen boyut ve şekil ile karanlık bölgeler olarak görünür

Kolay değil: Benzer karanlık bölgelere hava koşulları neden olabilir (ör. şiddetli rüzgar)

Yüksek eğitimli personel gerektiren pahalı süreç

Sınıflandırma şeması, konuşlandırılan şey değil, öğrenme şemasının kendisidir.

Makine Öğrenimine Giriş

Normalleştirilmiş görüntüden karanlık bölgeleri ayıklanması

Girdi: görüntüler
Çıktı: İşaretli bölgelere sahip daha küçük resimler.
Özelikleri normalleştirmek VE üretmek için görüntü işleme algoritmalarını kullanın.

Özellikleri:
- bölgenin büyüklüğü
- şekil, alan
- yoğunluk
- sınırların keskinliği ve pürüzlülüğü
- diğer bölgelerin yakınlığı
- arka plan hakkında bilgi

Öğrenme şeması özniteliklere uygulandıktan sonra

Kısıtlamalar/Sorunlar:
- Birkaç eğitim örneği—Petrol birikintileri nadirdir!
- Dengesiz veriler: çoğu karanlık bölge kaygan değildir
- Aynı görüntüdeki bölgeler bir toplu iş oluşturur—Farklı gruplar farklı arka planlara sahiptir
- Gereksinim: ayarlanabilir yanlış alarm oranı

Yük Tahmini

Daha hızlı yap

Elektrik tedarik şirketlerinin gelecekteki güç talebi tahminine ihtiyacı var

Her saat için min/maks yük tahminleri ==> önemli tasarruf

Verilen: "Normal" iklim koşullarını varsayan manuel olarak oluşturulmuş yük modeli

Sorun: Hava koşullarına göre ayarlamak

Statik model şunlardan oluşur:

- yıl için temel yük
- yıl boyunca yük periyodikliği
- tatillerin etkisi

On beş yıllık veri artı "ek" faktörleri.

Makine Öğrenimine Giriş

Tahmin, "ençok benzeyen" günler kullanılarak düzeltilmiştir.

Özellikler:

- sıcaklık
- nem
- Rüzgar hızı
- bulut örtüsü okumaları
- artı gerçek yük ile öngörülen yük arasındaki fark

Ortaya çıkan sistem, uzman tahminci ile aynı değeri vermiştir, ancak hesaplanması saatler yerine saniyeler almıştır.

Statik modele eklenen sekiz "ençok benzeyen" gün arasındaki ortalama fark

Lineer regresyon katsayıları, benzerlik fonksiyonunda özeliklerin ağırlıklarını oluşturur.

Kuralları değiştirerek uzmanın yardım etmesine izin verin

Makine Arızalarının Teşhisi

Teşhis: Uzman sistemlerin klasik alanı

Verilen: Bir cihaz montajının çeşitli noktalarında ölçülen titreşimlerin Fourier analizi

Problem: Hangi arıza mevcut?

EĞER bir hata değilse, o zaman hata NEDİR?

Elektromekanik motorların ve jeneratörlerin
önleyici bakımı
Bilgi çok gürültülü

Şimdiye kadar: uzman/el yapımı kurallarla teşhis

Makine Öğrenimine Giriş

Mevcut: Uzmanın teşhisi 600 arıza
~300 yetersiz, geri kalanı eğitim için kullanılıyor

- Uzmanın alan bilgisi olmadığı için bu ilk kurallardan memnun değildir.

- Daha fazla arka plan / gürültü bilgisi, tatmin edici olan daha karmaşık kurallarla sonuçlanmıştır.

Öğrenilmiş kurallar, el yapımı olanlardan daha iyi performans göstermiştir.

para kazanma

Pazarlama ve Satış I

Şirketler, büyük miktarlarda pazarlama ve satış verilerini hassas bir şekilde kaydeder

Uygulamalar:

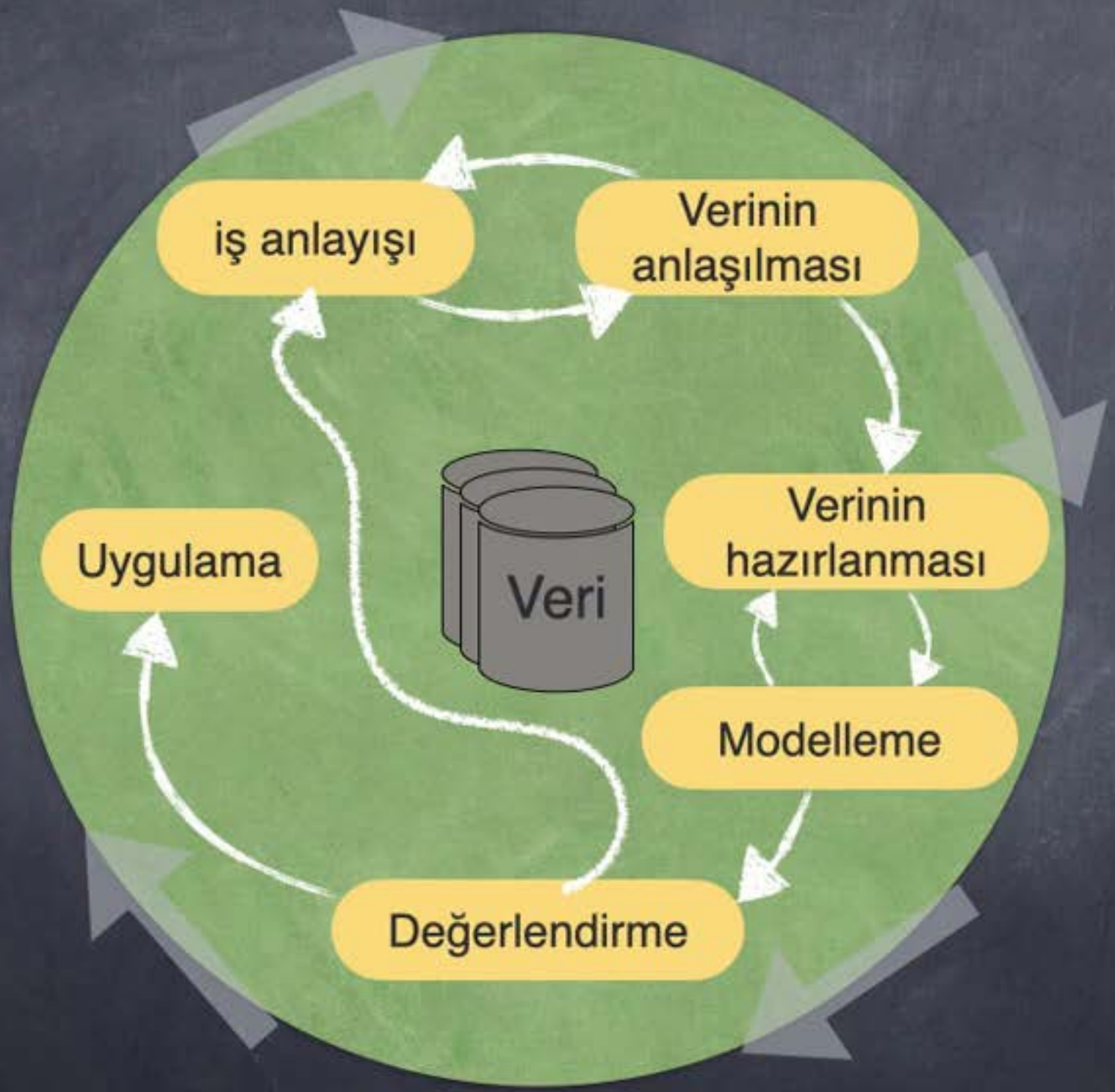
- Müşteri sadakati:
  Davranışlarındaki değişiklikleri tespit ederek hata yapma olasılığı olan müşterilerin belirlenmesi (örneğin bankalar/telefon şirketleri)
- Özel teklifler:
  Karlı müşterilerin belirlenmesi (örneğin, tatil sezonunda ekstra paraya ihtiyaç duyan güvenilir kredi kartı sahipleri)

Pazarlama ve Satış II

- Pazar Sepeti Analizi
  İlişkilendirme teknikleri, bir işlemde birlikte ortaya çıkma eğiliminde olan öğe gruplarını bulur (ödeme verilerini analiz etmek için kullanılır)
- Satın alma kalıplarının tarihsel analizi
- Potansiyel müşterilerin belirlenmesi
  Promosyon postalarına odaklanma (hedeflenen kampanyalar, kitlesel olarak pazarlanan kampanyalardan daha ucuzdur)

Perşembe (sadece birkaç ürün) ve Cumartesi günleri (tam alışveriş) Meyve suyu/Bebek Bezi'ni hatırlatın

Veri Madenciliği Süreci



Makine Öğrenimi ve İstatistik

Tarihsel fark (büyük ölçüde basitleştirilmiş):

- İstatistikler: hipotezleri test etme
- Makine öğrenimi: doğru hipotezi bulma

Ama: büyük örtüşme

- Karar ağaçları
- En yakın komşu yöntemleri

Bugün: bakış açıları birleşti

- Çoğu ML algoritması istatistiksel teknikler kullanır

Arama Olarak Genelleme

Tüm olası kural kümelerinde arama olarak Veri Madenciliği:

- Tüm olası kural kümelerinin kümesi çok büyük
- Tüm seti aramak zor
- Verilerdeki gürültü tüm kural kümelerini ortadan kaldırabilir

Aramalardaki önyargılar, sorunların çözülmesine yardımcı olur

- Dil Önyargısı (Kural kümeleri nasıl tanımlanır?)
- Arama önyargısı (En iyi kural yerine iyi bir kural bulma)
- Aşırı uyumdan kaçınma önyargısı (Daha basit ağaçlar yeni örneklere genelleme yapmakta daha iyi olabilir)

Veri Madenciliği ve Etik I

Pratik uygulamalarda ortaya çıkan etik sorunlar

- **Verileri anonimleştirmek zordur**
  Amerikalıların %85'i sadece posta kodu, doğum tarihi ve cinsiyetten tanımlanabilir
- **Veri madenciliği genellikle ayrımcılık yapmak için kullanılır**
  Örneğin. kredi başvuruları: bazı bilgileri (ör. cinsiyet, din, ırk) kullanmak etik değildir
- **Etik durum uygulamaya bağlıdır**
  Örneğin. tıbbi uygulamada aynı bilgiler OK
- **Özellikler sorunlu bilgiler içerebilir**
  Örneğin. alan kodu ırkla ilişkili olabilir

Veri Madenciliği ve Etik II

- Önemli sorular:
  - Verilere kimlerin erişmesine izin verilir?
  - Veriler hangi amaçla toplandı?
- Bundan meşru olarak ne tür sonuçlar çıkarılabilir?
- Sonuçlara uyarılar eklenmelidir
- Tamamen istatistiksel argümanlar asla yeterli değildir!
- Kişiler Veri Madenciliği'nin sonuçlarını kendi bilgileriyle birlikte almalı ve bunları nasıl ve uygulanıp uygulanmayacağına karar vermelidir.