Büyük Veri İçin İstatistiksel Öğrenme (Statistical Learning for Big Data)

M. Ali Akcayol
Gazi Üniversitesi
Bilgisayar Mühendisliği Bölümü

Bu dersin sunumları, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Trevor Hastie, Robert Tibshirani, Jerome Friedman, Springer, 2017." ve "Mining of Massive Datasets, Jure Leskovec, Anand Rajaraman, Jeffrey David Ullman, Stanford University, 2011." kitapları kullanılarak hazırlanmıştır.

Konular

- Sınıflandırma Problemleri
- Sınıflandırıcı Tasarımı
- Temel Sınıflandırıcı Türleri
- İstatistiksel Öğrenme
- Denetimli Öğrenme
- Denetimsiz Öğrenme
- Regresyon

Sınıflandırma Problemleri

- Sınıflandırma, günlük hayattaki problemlerde yaygın bir şekilde kullanılmaktadır.
- **Sınıflandırma problemlerinin çözümünde** belirlenmiş iki veya daha fazla **sınıf, özellikler** veya **nitelikler kümesinin tanımlı olması gereklidir.**
- **İstatistiksel yöntemler, makine öğrenmesi yöntemleri** veya **yapay sinir ağları** sınıflandırma problemleri için kullanılmaktadır.

3

Sınıflandırma Problemleri

İstatistiksel Yöntemler

- İstatistiksel yöntemler genellikle bir **olasılık modeline dayanır** ve her bir **sınıfın olma olasılığını belirler.**
- **Model** genellikle **probleme özgü** oluşturulur.
- **Kullanıcılar** sadece **parametre seçimi** gibi özelliklere yönelik işlemleri **yaparlar.**
- Günümüzde **modern istatistiksel yöntemler**, joint distribution (birleşik dağılım) gibi **birden fazla sınıfın aralarındaki ilişkiyi** de göz önüne alarak sınıflandırma kuralı sağlarlar.

4

Sınıflandırma Problemleri

Makine Öğrenmesi

- **Makine öğrenmesi yöntemleri** genellikle **mantıksal** veya **binary operatörleri kullanır.**
- Bir grup **örnek üzerinden öğrenme işlemi gerçekleştirilir.**
- Örneğin, **karar ağaçları** bir grup **mantıksal operatöre göre sınıflandırma yapar** ve kural tabanlı bir makine öğrenmesi yöntemidir.
- Makine öğrenmesi **sınıflandırma deyimlerini olabildiği kadar basit bir şekilde oluşturmaya çalışır.**
- Kullanıcılar kuralları kolay bir şekilde anlayabilirler.

5

Sınıflandırma Problemleri

Yapay Sinir Ağları

- **Yapay sinir ağları**, insan beyninin **anlama** ve **muhakeme yeteneğini taklit etmeyi amaçlar.**
- İnsanların **dil becerileri, ticari uygulamalar, bilimsel** ve mühendislik disiplinlerine yönelik **örüntü tanıma, modelleme** ve **tahmin** gibi çok farklı uygulama alanları bulunmaktadır.
- Yapar sinir ağları genellikle **çok katmanlı ve düğümlerin (node, neuron) birbiriyle bağlantılı** olduğu yapıdadır.
- Her **düğüm bir** veya **birden fazla giriş alabilir.**
- **Düğümlerden bir kısmı yapay sinir ağının çıkışını oluşturur.**
- Ağın tamamı çok karmaşık yapıya sahiptir.

6

Konular

- Sınıflandırma Problemleri
- Sınıflandırıcı Tasarımı
- Temel Sınıflandırıcı Türleri
- İstatistiksel Öğrenme
- Denetimli Öğrenme
- Denetimsiz Öğrenme
- Regresyon

7

Sınıflandırıcı Tasarımı

- Sınıflandırıcıların tasarımında, **doğruluk**, **hız**, **kapsayıcılık** ve **öğrenme süresi** oldukça **önemlidir.**

Doğruluk

- Bir sınıflandırıcı **çoğu girişler için doğru sınıfları belirleyebilir,** ancak **bazı girişler için hata da yapabilir.**
- Bu **hataların sıklığı** ve hatalı sonuçların önemi geliştirilen **sınıflandırıcının** performansını ve **kullanılabilirliğini etkilemektedir.**
- **Bazı sınıflandırıcılar için hata oranının kontrol edilebilmesi** de **önemlidir.**

8

Sınıflandırıcı Tasarımı

Hız

- **Bazı uygulama alanlarında sınıflandırıcının hızı en önemli gereksinimdir.**
- Özellikle **gerçek zamanlı uygulamalarda** sınıflandırıcının çok kısa sürede sonuç üretmesi gereklidir.
- Eğer, **bir sınıflandırıcı %90 doğruluğa sahipse** ve %95 doğruluğa sahip bir sınıflandırıcıya göre **100 kat daha hızlı sonuç üretiyorsa, doğruluk düzeyi düşük de olsa tercih edilebilir.**

9

Sınıflandırıcı Tasarımı

Kapsayıcılık

- Sınıflandırıcının sahip olduğu bir **kuralın kolay anlaşılabilir olması gereklidir**, aksi takdirde uygulanması sırasında hatalara neden olabilir.
- Problemin **tüm durumlarını içerecek şekilde** güvenilir sonuçlar üretmesi gerekir.

Öğrenme süresi

- Sınıflandırıcının özellikle **çabuk** ve **sık değişen ortamlarda** sınıflandırma kurallarını **hızlı öğrenmesi gereklidir.**
- Gerçek zamanlı **değişen şartlara hızla uyarlanabilir olması** sistemin doğruluğu açısından oldukça önemlidir.

10

Konular

- Sınıflandırma Problemleri
- Sınıflandırıcı Tasarımı
- Temel Sınıflandırıcı Türleri
- İstatistiksel Öğrenme
- Denetimli Öğrenme
- Denetimsiz Öğrenme
- Regresyon

11

Temel Sınıflandırıcı Türleri

- **İlk sınıflandırıcılar** sınıfların birbirinden ayrılması için **iki boyutlu**, **üç boyutlu** veya **çok boyutlu uzayda bir grup doğru kullanmaktaydı.**
- **Temel sınıflandırıcılar doğrusal yöntemlerle** girişlere göre **hedef sınıfı belirlemektedir.**
- Temel sınıflandırıcılar **çözüm uzayını doğrusal olarak bölerler.**
- **Fisher doğrusal ayrıştırıcı**, **karar ağaçları** ve **kural tabanlı yöntemler temel sınıflandırıcılardır.**

12

Temel Sınıflandırıcı Türleri

Fisher Doğrusal Ayrıştırıcı

- Fisher doğrusal ayrıştırıcı **en temel sınıflandırıcıdır.**
- **Iris veriseti** için başarılı sonuçlar vermektedir.
- Iris veriseti üç sınıf etiketine sahiptir: **SETOSA, VERSICOLOR** ve **VIRGINICA**.
- Farklı türdeki **50 çiçeğin çanak yaprakları (sepal)** ve **taç yapraklarının (petal)** ölçülerine göre sınıflandırma yapılmaktadır.
- Fisher doğrusal ayrıştırıcısına göre VERSICOLOR ve VIRGINICA için **aşağıdaki iki kural yazılabilir.**

  1- EĞER Petal Genişliği < 3,272 – 0,3254 * Petal Uzunluğu THEN VERSICOLOR

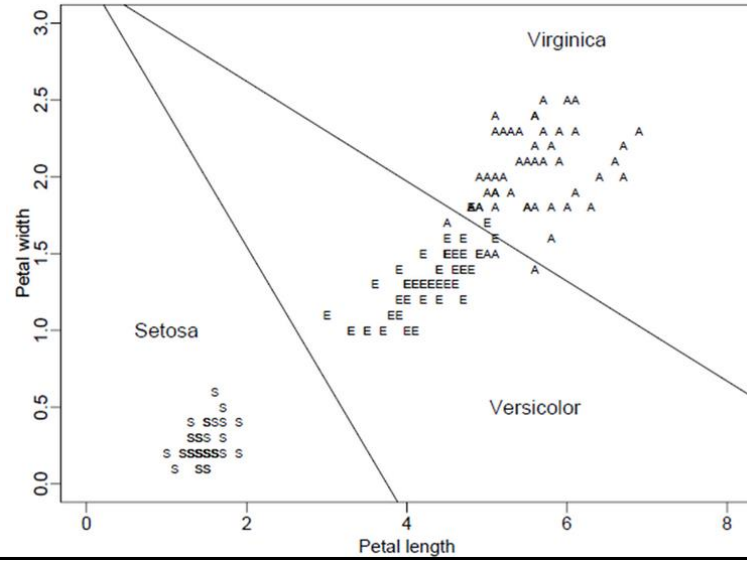
  2- EĞER Petal Genişliği > 3,272 – 0,3254 * Petal Uzunluğu THEN VIRGINICA

- Kurallar uygulandığında **6 gözlem değeri hatalı** sınıflandırılmaktadır.

13

Temel Sınıflandırıcı Türleri

Fisher Doğrusal Ayrıştırıcı



14

Temel Sınıflandırıcı Türleri

Karar Ağaçları ve Kural Tabanlı Yöntemler

- Oluşturulan kurallar ile **çözüm uzayı kutular halinde parçalanır.**
- **Her aşamada bir kutu test edilir** ve kalan kutular azaltılmış olur.
- Iris veriseti için aşağıdaki kurallar yazılabilir:

  1- EĞER Petal Uzunluğu < 2,65 THEN SETOSA

  2- EĞER Petal Uzunluğu > 4,95 THEN VIRGINICA

  3- EĞER 2,65 < Petal Uzunluğu < 4,95 THEN

  EĞER Petal Genişliği < 1,65 THEN VERSICOLOR

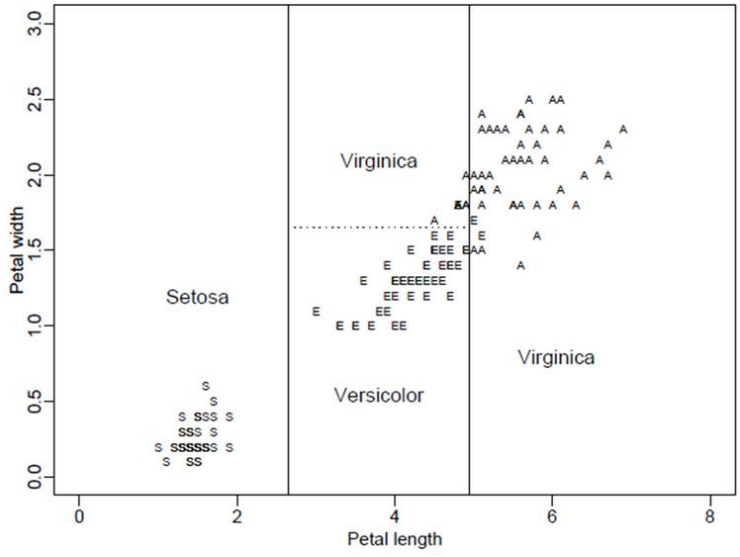
  EĞER Petal Genişliği > 1,65 THEN VIRGINICA

- Yukarıda oluşturulan sınıflandırma kuralları **3 hata** ile sınıflandırma işlemini yapabilir.

15

Temel Sınıflandırıcı Türleri

Karar Ağaçları ve Kural Tabanlı Yöntemler



16

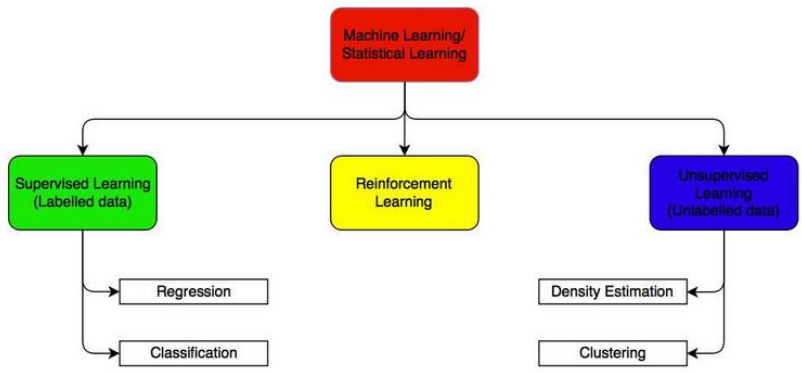
Konular

- Sınıflandırma Problemleri
- Sınıflandırıcı Tasarımı
- Temel Sınıflandırıcı Türleri
- İstatistiksel Öğrenme
- Denetimli Öğrenme
- Denetimsiz Öğrenme
- Regresyon

17

İstatistiksel Öğrenme

- İstatistiksel öğrenme, veri içerisinde **önceden bilinmeyen örüntülerin keşfedilmesi için kullanılan yöntemler kümesidir.**
- Makine öğrenmesi yöntemleri bilgi keşfi amacıyla **istatistiksel öğrenme yapar.**



18

İstatistiksel Öğrenme

- Makine öğrenmesi yöntemleri **denetimli, denetimsiz** veya **yönlendirmeli öğrenme** yapabilir.
- **Denetimli öğrenmede,** öğrenen sisteme **giriş veri seti** ile **birlikte sınıf etiketinin de verilmesi** zorunludur.
- **Denetimsiz öğrenmede,** sadece **giriş veri seti verilir** ve **veriler arasındaki ilişkilendirme kuralı** (uzaklık, benzerlik veya başka bir ilişkilendirme ölçütü) **verilir.**
- **Yönlendirmeli öğrenmede,** modelden elde edilen **sonucun kalitesini** veya değerini **ölçmek** için **uygunluk fonksiyonu** (fitness function) **tanımlanır.**
- **Denetimli öğrenme** yöntemlerine **yapay sinir ağları, karar ağaçları, destek vektör makineleri** örnek verilebilir.
- **Denetimsiz öğrenme** yöntemlerine **k-means** örnek olarak verilebilir.
- **Yönlendirmeli öğrenmeye** ise, **genetik algoritma, tavlama benzetimi** örnek olarak verilebilir.

19

Konular

- Sınıflandırma Problemleri
- Sınıflandırıcı Tasarımı
- Temel Sınıflandırıcı Türleri
- İstatistiksel Öğrenme
- Denetimli Öğrenme
- Denetimsiz Öğrenme
- Regresyon

20

Denetimli Öğrenme

- **Denetimli (gözetimli) öğrenme,** makine öğrenmesinde sınıflandırma veya tümevarımlı (inductive) öğrenme şeklinde ifade edilir.
- Denetimli öğrenmede **hedef değerler** (targets) ile **giriş değerleri** (inputs) birlikte **eğitim kümesi** (training set) olarak sağlanır.
- Eğitim kümesinin boyutu ve giriş değerleri ile çıkış değerleri arasındaki örnek ilişki sayısı, **eğitim kümesinin tüm sistem davranışını ifade edebilmesini sağlayacak şekilde olmalıdır.**
- **Yetersiz eğitim kümesi** verisi ile iyi öğrenmiş bir model elde etmek mümkün değildir.
- **Öğrenme işleminde bir kayıt kümesi kullanılır** ve özellikler kümesi olarak gösterilir.

21

Denetimli Öğrenme

- $\mathbf{A = \{A_1, A_2, \ldots, A_{|A|}\}}$
- Burada, |A| kümedeki eleman sayısını gösterir.
- Bir **veri kümesi** aynı zamanda **hedef C özelliğine de (sınıf) sahip olabilir.**
- $\mathbf{C \cap A = \varnothing}$ dir ve aşağıdaki gibi ifade edilir:

  $\mathbf{C = \{c_1, c_2, \ldots, c_{|C|}\}, |C| \geq 2}$
- Verilen bir D veri kümesi için **öğrenmedeki amaç, A'daki özellikler ile C'deki sınıflar arasındaki ilişkiyi gösteren** bir sınıflandırma/tahmin için **fonksiyon oluşturmaktır.**
- Elde edilen bu fonksiyon, **sınıflandırma modeli, tahmin modeli** veya **sınıflandırıcı** olarak adlandırılır.

22

Denetimli Öğrenme

Örnek

- Bir banka şubesinin müşterilerini kredi verilebilirlik açısından sınıflandırdığını varsayalım.
- Banka yeni bir başvuru yapıldığında önceki bilgilerine göre krediye uygunluk durumunu belirleyecektir.
- Eğitim kümesinde belirli sayıda 15 gözlem verisi kullanılarak bir model geliştirilebilir.
- Bu verilere göre oluşturulacak bir öğrenen model ile **yeni gelen bir müşteri için kredi uygunluk durumunun belirlenmesi** veya tahmin edilmesi **gerçekleştirilebilir.**

23

Denetimli Öğrenme

Örnek

ID	Age	Has_job	Own_house	Credit_rating	Class
1	young	false	false	fair	No
2	young	false	false	good	No
3	young	true	false	good	Yes
4	young	true	true	fair	Yes
5	young	false	false	fair	No
6	middle	false	false	fair	No
7	middle	false	false	good	No
8	middle	true	true	good	Yes
9	middle	false	true	excellent	Yes
10	middle	false	true	excellent	Yes
11	old	false	true	excellent	Yes
12	old	false	true	good	Yes
13	old	true	false	good	Yes
14	old	true	false	excellent	Yes
15	old	false	false	fair	No

24

Konular

- Sınıflandırma Problemleri
- Sınıflandırıcı Tasarımı
- Temel Sınıflandırıcı Türleri
- İstatistiksel Öğrenme
- Denetimli Öğrenme
- Denetimsiz Öğrenme
- Regresyon

25

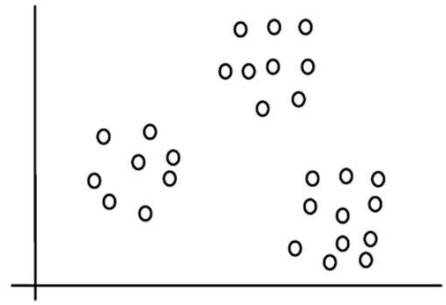
Denetimsiz Öğrenme

- Denetimli öğrenmede, giriş verileri ile çıkış niteliği arasındaki ilişkiyi ortaya çıkartır.
- Elde edilen model ile **yeni verilerle ileriye dönük tahmin yapılması amaçlanmaktadır.**
- Denetimsiz öğrenmede, eğitim sürecinde **hedef nitelik bulunmamaktadır.**
- Denetimsiz öğrenmede **veriler arasındaki** bazı **yapısal ilişkilerin** veya **örüntülerin** ortaya çıkartılması amaçlanmaktadır.
- Örneğin, kümelemede veri içerisindeki benzer örneklerin yakınlıklarına göre kümeler oluşturulur.
- Birbirine belirlenen değerden daha uzak olanlar ayrı kümelere atanır.
- **Apriori algoritması** ile yapılan birliktelik kural madenciliği **unsupervised learning olarak nitelendirilir.**

26

Denetimsiz Öğrenme

- Kümeleme, **denetimsiz öğrenme olarak adlandırılır.**
- Aşağıdaki veri kümesinde uzaklıklara göre üç küme görülmektedir. Farklı özellikler gözönüne alınırsa küme sayısı daha fazla veya daha az olabilir.



- Eksenler, yaş ve boy, gelir ve harcama, eğitim ve gelir gibi birbiriyle ilişkili veriler olabilir ve çıkış ise doğal olarak oluşan üç sınıf olabilir.
- Sağlık, psikoloji, tarım, sosyoloji, biyoloji, arkeoloji, pazarlama, sigortacılık, kütüphane gibi **çok farklı alanlarda kullanılmaktadır.**

27

Konular

- Sınıflandırma Problemleri
- Sınıflandırıcı Tasarımı
- Temel Sınıflandırıcı Türleri
- İstatistiksel Öğrenme
- Denetimli Öğrenme
- Denetimsiz Öğrenme
- Regresyon

28

Regresyon

- Doğrusal regresyon modelinde, **giriş değişkeni/değişkenleri ile çıkış değişkeni/değişkenleri arasındaki ilişki doğrusaldır.**
- **Parametrelerin** tüm örnekler veya gözlem değerleri için **uygun değerleri hesaplanır.**
- Çözüm uzayı da **doğrularla sınıf etiketlerini birbirinden ayırır.**
- **Doğrusal bir model:**

  $f(x) = mx + b$

  şeklinde tanımlanır.
- Burada x giriş parametresi, $\boldsymbol{f(x)}$ **hesaplanan çıkış değeri**, $\boldsymbol{m}$ **ve** $\boldsymbol{b}$ **ise parametrelerdir.**

29

Regresyon

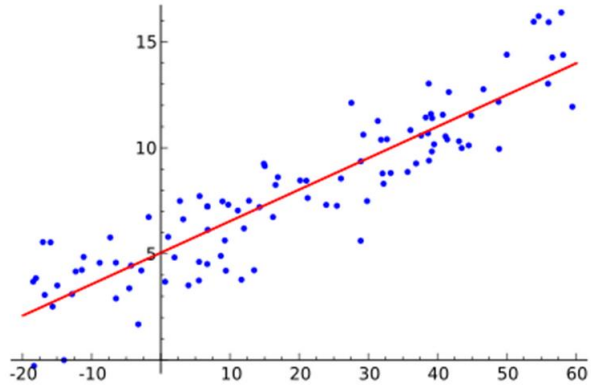
- Eğitim kümesindeki tüm x değerleri için **hesaplanan** $\mathbf{y} = \mathbf{f(x)}$ **değerleri ile beklenen çıkış değeri olan** $\hat{\mathbf{y}}$ **arasındaki fark minimuma indirilmeye çalışılır.**
- Hata farkı olarak en yaygın kullanılan **hataların farklarının karelerinin toplamı (sum squared error)** ile ifade edilir.

$$SSE = \sum (y_i - \hat{y}_i)^2$$

30

Regresyon

- Aşağıdaki şekilde doğrusal regresyon modeli ile elde edilen örnek bir sınıflandırma çözümü görülmektedir.



- Şekilde **doğru ile örnekler iki sınıfa ayrılmıştır.**
- **Çözüm uzayı** doğrusal bir fonksiyonla **ikiye bölünmüş durumdadır.**

31

Ödev

- Öğrenen modellerin genelleme performansı için model seçimi ve değerlendirme ölçütleri hakkında detaylı bir araştırma ödevi hazırlayınız.

32