

VERİ MADENCİLİĞİ

Farklı Sınıflandırma Yöntemleri

Yrd. Doç. Dr. Şule Gündüz Öğüdücü
<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

1

Konular

- Sınıflandırma yöntemleri
 - Örnek tabanlı yöntemler
 - k-En Yakın Komşu Yöntemi
 - Genetik Algoritmalar
 - Karar Destek Makinaları
 - Bulanık Küme Sınıflandırıcılar
 - Öngörü
 - Eğri Uydurma
- Model Değerlendirme
- Öğrenme, sinama, geçerleme kümelerini oluşturma
- Sınıflandırıcıları birleştirme

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

2

Örnek Tabanlı Yöntemler

- Örnek tabanlı sınıflandırma:
 - Öğrenme kümesi saklanır
 - Sınıflandırılacak yeni bir örnek geldiğinde öğrenme kümesi sınıf etiketini öngörmek için kullanılır (tembel (lazy) yöntemler)
- Yöntemler
 - k-en yakın komşu yöntemi

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

3

Örnek Tabanlı Yöntemler

Öğrenme Kümesi

Nit1	NitN	Sınıf
			A
			B
			B
			C
			A
			C
			B

Yeni örnek

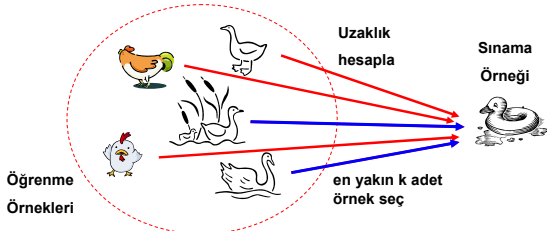
Nit1	NitN

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

4

En Yakın Komşu Yöntemi

- Temel yaklaşım: Sınıflandırılmak istenen örneğe en yakın örnekleri bul.
Örnek: ördek gibi yürüyor, ördek gibi bağıyor
=> büyük olasılıkla ördek

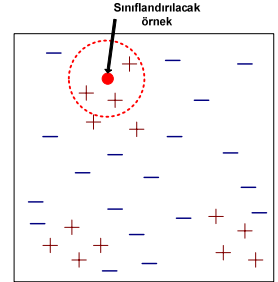


<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

5

En Yakın Komşu Sınıflandırıcı

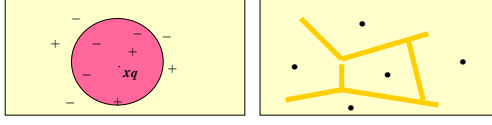
- Bütün örnekler n-boyutlu uzayda bir noktaya karşı düşürülür
- Nesneler arasındaki uzaklık (Öklid uzaklığı)
- Öğrenilen fonksiyon ayrık değerli veya gerçel değerli olabilir
- Ayrık değerli fonksiyonlarda k-komşu algoritması X_q örneğine en yakın k öğrenme örneğinde en çok görülen sınıf değerini verir
- Sürekli değerli fonksiyonlarda en yakın k öğrenme örneğinin ortalaması alınır



<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

6

K-En Yakın Komşu Yöntemi



- x_q örneği 1-en yakın komşuya göre pozitif olarak, 5-en yakın komşuya göre negatif olarak sınıflandırılır
- Voronoi diyagramları: Her öğrenme örneğini çevreleyen dışbükey çokgenlerden oluşan karar yüzeyi

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

7

K-En Yakın Komşu Yöntemi

- Uzaklık-ağırlıklı k-en yakın komşu algoritması
 - Öğrenme kümesindeki örneklere (x_i), sınıflandırılmak istenen örneğe (x_q) olan uzaklıklarına göre ağırlıklar verilmesi
 - yakın örneklerin ağırlığı daha fazla $w \equiv \frac{1}{d(x_q, x_i)^2}$
- k-en yakın komşunun ortalaması alındığı için gürültülü veriden az etkileniyor
- İlgisiz nitelikler uzaklığı etkileyebilir
 - bu nitelikler uzaklık hesaplarında kullanılmayabilir

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

8

Konular

- Sınıflandırma yöntemleri
 - Örnek tabanlı yöntemler
 - k-En Yakın Komşu Yöntemi
 - Genetik Algoritmalar
 - Karar Destek Makinaları
 - Bulanık Küme Sınıflandırıcılar
 - Öngörü
 - Eğri Uydurma
- Model Değerlendirme
- Öğrenme, sınıma, geçerleme kümelerini oluşturma
- Sınıflandırıcıları birleştirme

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

9

Genetik Algoritmalar

- Optimizasyon amaçlı
- Bir başlangıç çözümü öneriyor, tekrarlanan her ara adımda daha iyi çözüm üretmeye çalışıyor.
- Doğal evrime ve en iyi olanın yaşamını sürdürmesine dayanıyor
- Çözümü birey olarak sunuyor.
- Birey: $I = I_1, I_2, \dots, I_n - I_j$ kullanılan alfabenin bir karakteri
- gen: I_j
- Toplum: Bireylerden oluşan küme

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

10

Genetik Algoritmalar

- Genetik Algoritmalar (GA) 5 parçadan oluşuyor:
 - Bireylerden oluşan bir başlangıç kümesi, P
 - Çaprazlama (Crossover): Bir anne babadan yeni bireyler üretmek için yapılan işlem
 - Mutasyon: Bir bireyi rastgele değiştirme
 - Uygunluk (fitness): En iyi bireyleri belirleme
 - Çaprazlama ve mutasyon tekniklerini uygulayan ve uygunluk fonksiyonuna göre toplum içindeki en iyi bireyleri seçen algoritma

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

11

Çaprazlama Örnekleri

000 000	000 111	000 000 00	000 111 00
111 111	111 000	111 111 11	111 000 11
Parents	Children	Parents	Children

a) Single Crossover

a) Multiple Crossover

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

12

Genetik Algoritma

```

Input:
P //Initial Population
Output:
P' //Improved Population
Genetic Algorithm:
//Illustrates Genetic Algorithm
repeat
  N = |P|;
  P' = ∅;
  repeat
    i1, i2 = select(P);
    o1, o2 = cross(i1, i2);
    o1 = mutate(o1);
    o2 = mutate(o2);
    P' = P' ∪ {o1, o2};
  until |P'| = N;
  P = P';
until termination criteria satisfied;
    
```

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

13

GA – Avantajlar, Dezavantajlar

- Avantaj
 - Paralel çalışabilir
 - NP karmaşık problem çözümlerine uygun
- Dezavantaj
 - Son kullanıcının modeli anlaması güç
 - Problemi GA ile çözmeye uygun hale getirmek zor
 - Uygunluk fonksiyonunu belirlemek zor
 - Çaprazlama ve mutasyon tekniklerini belirlemek zor

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

14

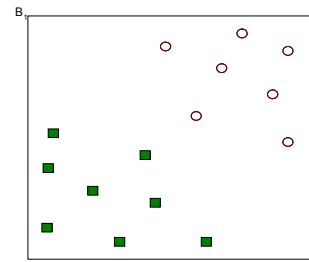
Konular

- Sınıflandırma yöntemleri
 - Örnek tabanlı yöntemler
 - k-En Yakın Komşu Yöntemi
 - Genetik Algoritmalar
 - Karar Destek Makineleri
 - Bulanık Küme Sınıflandırıcılar
 - Öngörü
 - Eğri Uydurma
- Model Değerlendirme
- Öğrenme, sınama, geçerleme kümelerini oluşturma
- Sınıflandırıcıları birleştirme

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

15

Karar Destek Makineleri

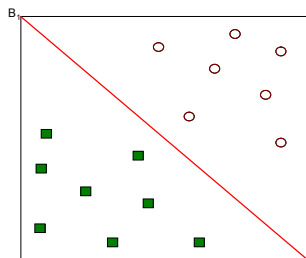


- Karar Destek Makineleri (Support Vector Machines SVM): Veriyi ayıracak doğrusal bir sınır

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

16

Karar Destek Makineleri

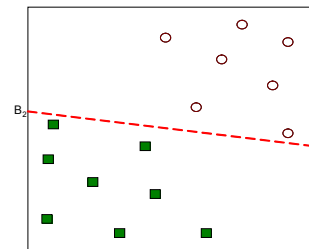


- Bir çözüm

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

17

Karar Destek Makineleri

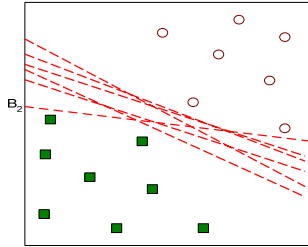


- Başka bir çözüm

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

18

Karar Destek Makineleri

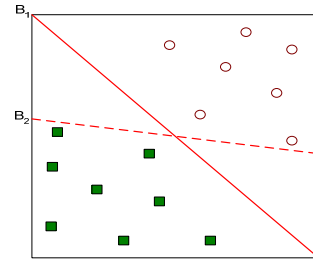


- Diğer çözümler

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

19

Karar Destek Makineleri

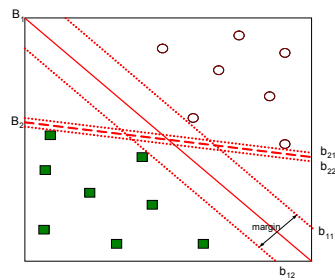


- Hangisi daha iyi? B1 mi, B2 mi?
- Daha iyi nasıl tanımlanır?

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

20

Karar Destek Makineleri

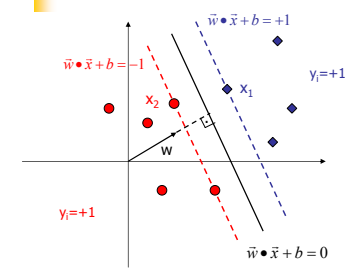


- Farklı sınıftan örnekler arasındaki uzaklığı enbüyüten düzlemi bul
=> B1, B2'den daha iyi

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

21

Karar Destek Makineleri



$$(w \cdot x) + b \geq +1, y_i = +1$$

$$(w \cdot x) + b \leq -1, y_i = -1$$

$$\Rightarrow y_i(w \cdot x + b) \geq +1$$

$$(w \cdot x_1) + b = y_i = +1$$

$$(w \cdot x_2) + b = y_i = -1$$

$$\Rightarrow w \cdot (x_1 - x_2) = 2$$

$$\Rightarrow \frac{w \cdot (x_1 - x_2)}{\|w\|} = \frac{2}{\|w\|}$$

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

22

Karar Destek Makineleri

- $\frac{2}{\|w\|^2}$ enbüyük olması isteniyor

- $y_i(w \cdot x + b) \geq +1$ olacak şekilde $\frac{\|w\|^2}{2}$ enküçük olmalı
- kısıtlı eniyileme (constraint optimization) problem

$$Lp = \frac{1}{2} \|w\|^2 - \sum_{i=1}^N a_i [y_i (w \cdot x_i + b) - 1]$$

- Problem

a_1, \dots, a_N bulunması

$\sum a_i - \frac{1}{2} \sum \sum a_i a_j y_i y_j x_i^T x_j$ en büyük olacak

kısıtlar:

(1) $\sum a_i y_i = 0$

(2) $a_i \geq 0, \forall a_i$

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

23

Eniyileme Problemi Çözümü

- Çözüm

$$w = \sum a_i y_i x_i \quad b = y_k - w^T x_k \quad \forall x_k, a_k \neq 0$$

- Sınıflandırma fonksiyonu

$$f(x) = \sum a_i y_i x_i^T x + b$$

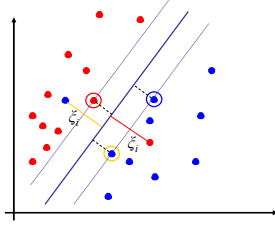
- $f(x) = 1$ ise x pozitif olarak, diğer durumlarda negatif olarak sınıflandırılıyor.

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

24

Karar Destek Makineleri

- Öğrenme kümesi doğrusal olarak ayrılmıyor
- ξ_i değişkenleri ekleniyor



$$\begin{aligned} (w \cdot x) + b &\geq +1 - \xi_i, \quad y_i = +1 \\ (w \cdot x) + b &\leq -1 - \xi_i, \quad y_i = -1 \\ &\Rightarrow y_i(w \cdot x + b) \geq +1 - \xi_i \\ \xi_i &\geq 0, \quad \forall i \end{aligned}$$

$$Lp = \frac{1}{2} \|w\|^2 - C \left(\sum_{i=1}^N \xi_i \right)$$

- Problem: $a_1 \dots a_N$ bulunması
- $\sum a_i - \frac{1}{2} \sum \sum a_i a_j y_i y_j x_i^T x_j$ en büyük olacak kısıtlar:
- (1) $\sum a_i y_i = 0$
- (2) $0 \leq a_i \leq C, \forall a_i$

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

25

Eniyileme Problemi Çözümü

- Çözüm

$$\begin{aligned} w &= \sum a_j y_j x_j \\ b &= y_k (1 - \xi_k) - w^T x_k, \quad k = \arg \max a_k \end{aligned}$$

- Sınıflandırma fonksiyonu

$$f(x) = \sum a_j y_j x_j^T x + b$$

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

26

SVM Uygulamaları

- Boser, Guyon ve Vapnik tarafından 1992 yılında önerildi. 1990'ların sonlarına doğru yaygın olarak kullanılmaya başlandı
- SVM için en yaygın eniyileme algoritmaları SMO [Platt '99] ve SVM^{light} [Joachims' 99]

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

27

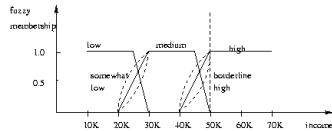
Konular

- Sınıflandırma yöntemleri
 - Örnek tabanlı yöntemler
 - k-En Yakın Komşu Yöntemi
 - Genetik Algoritmalar
 - Karar Destek Makinaları
 - Bulanık Küme Sınıflandırıcılar
 - Öngörü
 - Eğri Uydurma
- Model Değerlendirme
- Öğrenme, sınıma, geçerleme kümelerini oluşturma
- Sınıflandırıcıları birleştirme

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

28

Bulanık Küme Sınıflandırıcılar



- Bulanık mantık 0.0 ve 1.0 arasında geçerli 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

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

29

Konular

- Sınıflandırma yöntemleri
 - Örnek tabanlı yöntemler
 - k-En Yakın Komşu Yöntemi
 - Genetik Algoritmalar
 - Karar Destek Makinaları
 - Bulanık Küme Sınıflandırıcılar
 - Öngörü
 - Eğri Uydurma
- Model Değerlendirme
- Öğrenme, sınıma, geçerleme kümelerini oluşturma
- Sınıflandırıcıları birleştirme

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

30

Öngörü

- Sınıflandırma problemleriyle aynı yaklaşım
 - model oluştur
 - bilinmeyen değeri hesaplamak için modeli kullan
 - eğri uydurma
 - doğrusal
 - doğrusal olmayan
- Sınıflandırma ayırık değerli
- Öngörü sürekli değerli

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

31

Eğri Uydurma

- Doğrusal eğri uydurma:
 - en basit eğri uydurma yöntemi
 - veri doğrusal bir eğri ile modellenir.
 - veri kümesindeki niteliklerin doğrusal fonksiyonu
- $y = w_0 + w_1 a_1 + w_2 a_2 + \dots + w_k a_k$
- öğrenme kümesindeki y_i sınıfından bir x_i örneği için çıkış
- karese hatayı en küçültecek ağırlıkları bulma

$$y = w_0 x_{i0} + w_1 x_{i1} + w_2 x_{i2} + \dots + w_k x_{ik} = \sum_{j=0}^k w_j x_{ij}$$

$$\sum_{i=1}^n \left(y_i - \sum_{j=0}^k w_j x_{ij} \right)^2$$

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

32

Konular

- Sınıflandırma yöntemleri
- Model Değerlendirme
 - Hata oranı
 - Anma
 - Duyarlılık
 - F-ölçütü
 - ROC eğrileri
- Öğrenme, sinama, geçerleme kümelerini oluşturma
- Sınıflandırıcıları birleştirme

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

33

Sınıflandırma Modelini Değerlendirme

- Model başarımını değerlendirme ölçütleri nelerdir?
 - Hata oranı
 - Anma
 - Duyarlılık
 - F-ölçütü
- Farklı modellerin başarımını nasıl karşılaştırılır?
 - ROC

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

34

Sınıflandırma Hatası

- Sınıflandırma yöntemlerinin hatalarını ölçme
 - başarı: örnek doğru sınıfa atandı
 - hata: örnek yanlış sınıfa atandı
 - hata oranı: hata sayısının toplam örnek sayısına bölünmesi
- Hata oranı sinama kümesi kullanılarak hesaplanır

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

35

Model Başarımını Değerlendirme

- Model başarımını değerlendirme ölçütleri
 - modelin ne kadar doğru sınıflandırma yaptığını ölçer
 - hız, ölçeklenebilirlik gibi özellikleri değerlendirmez
- Karışıklık matrisi:

	ÖNGÖRÜLEN SINIF		
DOĞRU SINIF		Sınıf=1	Sınıf=-1
	Sınıf = 1	a	b
	Sınıf = -1	c	d

a: TP (true positive)
b: FN (false negative)
c: FP (false positive)
d: TN (true negative)

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

36

Model Başarımını Değerlendirme: Doğruluk

	ÖNGÖRÜLEN SINIF		
DOĞRU SINIF		+1	-1
	+1	a (TP)	b (FN)
	-1	c (FP)	d (TN)

Modelin başarımı:

$$\text{Dogruluk} = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{Hata Oranı} = \frac{b+c}{a+b+c+d} = \frac{FN+FP}{TP+TN+FP+FN}$$

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

37

Örnek

Sınıflandırıcı A	
TP=25	FN=25
FP=25	TN=25

Doğruluk=%50

Sınıflandırıcı B	
TP=50	FN=0
FP=25	TN=25

Doğruluk=%75

Sınıflandırıcı C	
TP=25	FN=25
FP=0	TN=50

Doğruluk=%75

Hangi sınıflandırıcı daha iyi?

- B ve C, A'dan daha iyi bir sınıflandırıcı
- B, C'den daha iyi bir sınıflandırıcı mı?

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

38

Model Başarımını Değerlendirme: Duyarlılık

	ÖNGÖRÜLEN SINIF		
DOĞRU SINIF		+1	-1
	+1	a (TP)	b (FN)
	-1	c (FP)	d (TN)

$$\text{Duyarlılık} = \frac{\text{Doğru sınıflandırılmış pozitif örnek sayısı}}{\text{Pozitif sınıflandırılmış örneklerin sayısı}}$$

$$= \frac{TP}{TP+FP}$$

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

39

Model Başarımını Değerlendirme: Anma

	ÖNGÖRÜLEN SINIF		
DOĞRU SINIF		+1	-1
	+1	a (TP)	b (FN)
	-1	c (FP)	d (TN)

$$\text{Anma} = \frac{\text{Doğru sınıflandırılmış pozitif örnek sayısı}}{\text{Doğru pozitif oranı}}$$

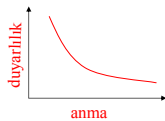
$$= \frac{TP}{TP+FN}$$

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

40

Anma / Duyarlılık

- A modeli B modelinden daha iyi anma ve duyarlılık değerine sahipse A modeli daha iyi bir sınıflandırıcıdır.
- Duyarlılık ve anma arasında ters orantı var.



<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

41

Sınıflandırıcıları Karşılaştırma

- Doğruluk en basit ölçüt
- Duyarlılık ve anma daha iyi ölçme sağlıyor
 - Model A'nın duyarlılığı model B'den daha iyi ancak model B'nin anma değeri model A'dan daha iyi olabilir.

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

42

Model Başarımını Değerlendirme: F-ölçütü

- F-ölçütü: Anma ve duyarlılığın harmonik ortalamasını alır.

$$F\text{-ölçütü} = \frac{2 * \text{duyarlılık} * \text{anma}}{\text{duyarlılık} + \text{anma}}$$

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

43

ROC (Receiver Operating Characteristic)

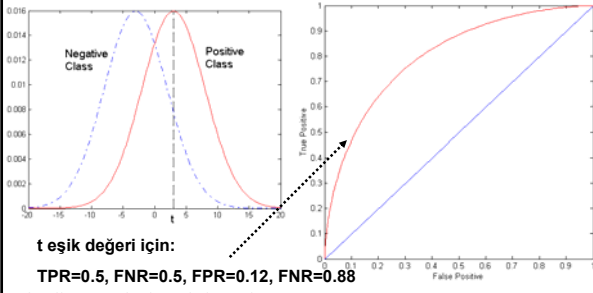
- İşaret işlemede bir sezicinin, gürültülü bir kanalda doğru algılama oranının yanlış alarm oranına karşı çizdirilen grafiği (algılayıcı işletim eğrisi)
- Farklı sınıflandırıcıları karşılaştırmak için ROC eğrileri
- Doğru pozitif (TPR - y eksen) oranının yanlış pozitif (FPR - x eksen) oranına karşı çizdirilen grafiği
 - $TPR = TP / (TP + FN)$
 - $FPR = FP / (TN + FP)$
- ROC üzerindeki her nokta bir sınıflandırıcının oluşturduğu bir modele karşı düşer

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

44

ROC Eğrisi

- İki sınıftan oluşan tek boyutlu bir veri kümesi (positive - negative)
- $x > t$ için her örnek pozitif olarak sınıflandırılıyor



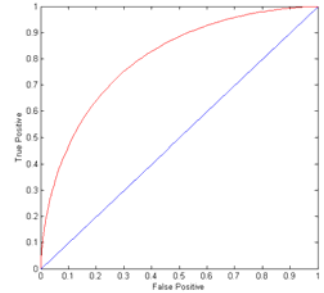
<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

45

ROC Eğrisi

(FPR, TPR)

- (0,0): Bütün örneklerin negatif sınıflandırılması
- (1,1): Bütün örneklerin pozitif sınıflandırılması
- (0,1): ideal durum
- Çapraz çizgi:
 - Rastlantısal tahmin

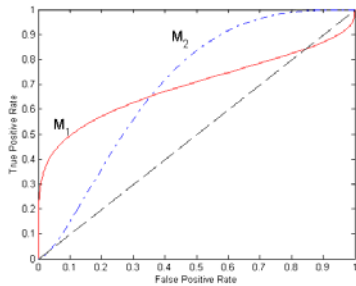


<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

46

ROC Eğrilerinin Kullanılması

- Farklı modelleri karşılaştırmak için
- M_1 veya M_2 birbirlerine üstünlük sağlamıyor
 - küçük FPR değerleri için M_1 daha iyi
 - büyük FPR değerleri için M_2 daha iyi
- ROC eğrisi altında kalan alan
 - ideal = 1
 - Rastlantısal tahmin=0.5



<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

47

ROC Eğrisinin Çizilmesi

- Her örnek için $P(+|A)$ olasılığı hesaplanır
- $P(+|A)$ değeri azalarak sıralanır
- Her farklı $P(+|A)$ değeri için bir eşik değeri uygulanır
- Her eşik değeri için TP, FP, TN, FN hesaplanır

Örnek	$P(+ A)$	Sınıf
1	0.95	+
2	0.93	+
3	0.87	-
4	0.85	-
5	0.85	-
6	0.85	+
7	0.76	-
8	0.53	+
9	0.43	-
10	0.25	+

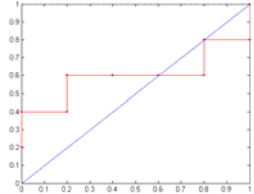
<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

48

ROC Eğrisinin Çizilmesi

Class	+	-	+	-	+	-	+	-	+	-	+	-
	0.25	0.43	0.53	0.76	0.85	0.85	0.85	0.87	0.93	0.95	1.00	
TP	5	4	4	3	3	3	3	2	2	1	0	
FP	5	5	4	4	3	2	1	1	0	0	0	
TN	0	0	1	1	2	3	4	4	5	5	5	
FN	0	1	1	2	2	2	2	3	3	4	5	
TPR	1	0.8	0.8	0.6	0.6	0.6	0.6	0.4	0.4	0.2	0	
FPR	1	1	0.8	0.8	0.6	0.4	0.2	0.2	0	0	0	

ROC Eğrisi:



<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

49

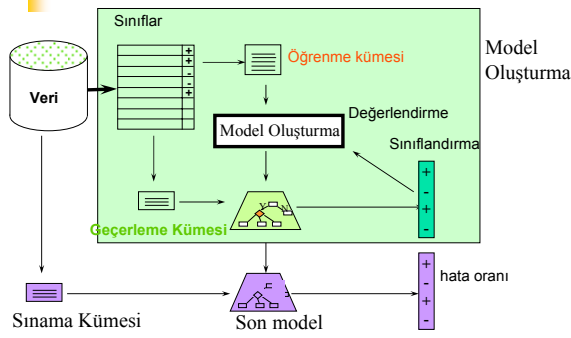
Model Parametrelerini Belirleme

- Sınama kümesi sınıflandırıcı oluşturmak için kullanılmaz
- Bazı sınıflandırıcılar modeli iki aşamada oluşturur
 - modeli oluştur
 - parametreleri ayarla
- Sınama kümesi parametreleri ayarlamak için kullanılmaz
- Uygun yöntem üç veri kümesi kullanma: öğrenme, geçerleme, sinama
 - geçerleme kümesi parametre ayarlamaları için kullanılır
 - model oluşturulduktan sonra öğrenme ve geçerleme kümesi son modeli oluşturmak için kullanılabilir

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

50

Sınıflandırma: Öğrenme, Geçerleme, Sinama



<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

51

Model Başarımını Tahmin Etme

- Örnek: Doğruluğu %25 olan bir modelin gerçek başarımı ne kadardır?
 - Sınama kümesinin büyüklüğüne bağlı
- Sınıflandırma (hileli) yazı tura atmaya benziyor
 - tura doğru sınıflandırma (başarı), yazı yanlış sınıflandırma (başarısızlık)
- İstatistikte birbirinden bağımsız olayların başarı ya da başarısızlıkla sonuçlanmaları Bernoulli dağılımı ile modellenir.
- Gerçek başarı oranını belirlemek için istatistikte güven aralıkları tanımlanmıştır.

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

52

Güven Aralığı

- p belli bir güvenle belli bir aralıkta bulunmaktadır.
- Örnek: $N=1000$ olayda $S=750$ başarı sağlanmış.
 - Tahmin edilen başarı oranı: 75%
 - Gerçek başarıya ne kadar yakın
 - %80 güven ile $p \in [73, 2 - 76, 7]$
- Örnek: $N=100$ olayda $S=75$ başarı sağlanmış.
 - Tahmin edilen başarı oranı: 75%
 - Gerçek başarıya ne kadar yakın
 - %80 güven ile $p \in [69, 1 - 80, 1]$

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

53

Ortalama Değer ve Varyans

- Başarı oranı p olan tek bir Bernoulli olayının ortalama değeri ve varyansı: $p, p(1-p)$
- N kere tekrarlanan Bernoulli olayının beklenen başarı oranı $f=S/N$
- Büyük N değerleri için, f normal dağılım
- f için ortalama değer ve varyans: $p, p(1-p)/N$
- Ortalama değeri 0 ve varyansı 1 olan X rastlantı değişkeninin $\%c$ güven aralığı :

$$\Pr[-z \leq X \leq z] = c$$
- Simetrik bir dağılım için:

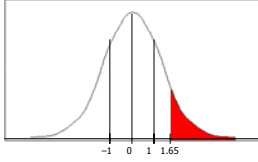
$$\Pr[-z \leq X \leq z] = 1 - 2 * \Pr[X \geq z]$$

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

54

Güven Sınırları

- Ortalama değeri 0 ve varyansı 1 olan bir normal dağılımın güven sınırları



$\Pr[X \geq z]$	z
0.1%	3.09
0.5%	2.58
1%	2.33
5%	1.65
10%	1.28
20%	0.84
40%	0.25

- $\Pr[-1,65 \leq X \leq 1,65] = 90\%$
- f 'in ortalama değerinin 0, varyansının 1 olacak şekilde dönüştürülmesi gerekir.

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

55

Dönüşüm

- f 'in ortalama değerinin 0, varyansının 1 olacak şekilde dönüştürülmesi için

$$\frac{f - p}{\sqrt{p(1-p)/N}}$$

- Güven aralığı

$$\Pr\left[-z \leq \frac{f - p}{\sqrt{p(1-p)/N}} \leq z\right] = c$$

- p 'nin değeri

$$p = \left(f + \frac{z^2}{2N} \pm z \sqrt{\frac{f}{N} - \frac{f^2}{N} + \frac{z^2}{4N^2}} \right) / \left(1 + \frac{z^2}{N} \right)$$

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

56

Örnek

- $f = 75\%$, $N = 1000$, $c = 80\%$ ($z = 1.28$):
 $p \in [0,732 - 0,767]$
- $f = 75\%$, $N = 100$, $c = 80\%$ ($z = 1.28$):
 $p \in [0,691 - 0,801]$

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

57

Konular

- Sınıflandırma yöntemleri
- Model Değerlendirme
- Öğrenme, sinama, geçerleme kümelerini oluşturma
 - holdout
 - k-kat çapraz geçerleme
 - Bootstrap
- Sınıflandırıcıları birleştirme

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

58

Verinin Dengesiz Dağılımı

- Küçük veya dengesiz veri kümeleri için örnekler tanımlayıcı olmayabilir
- Veri içinde bazı sınıflardan çok az örnek olabilir
 - tıbbi veriler: %90 sağlıklı, %10 hastalık
 - elektronik ticaret: %99 alışveriş yapmamış, %1 alışveriş yapmış
 - güvenlik: %99 sahtekarlık yapmamış, %1 sahtekarlık yapmış
- Örnek: Sınıf1: 9990 örnek, Sınıf2: 10 örnek
 - bütün örnekleri sınıf1'e atayan bir sınıflandırıcının hata oranı: $9990 / 10000 = \%99,9$
 - hata oranı yanıltıcı bir ölçüt olabilir

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

59

Dengeli Dağılım Nasıl Sağlanır?

- Veri kümesinde iki sınıf varsa
 - iki sınıfın eşit dağıldığı bir veri kümesi oluştur
 - Az örneği olan sınıftan istenen sayıda rasgele örnekler seç
 - Çok örneği olan sınıftan aynı sayıda örnekleri ekle
- Veri kümesinde iki sınıftan fazla sınıf varsa
 - Öğrenme ve sinama kümesini farklı sınıflardan aynı sayıda örnek olacak şekilde oluştur

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

60

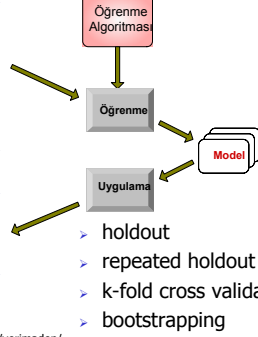
Örnek

İTİ	Nil1	Nil2	Nil3	Sınıf
1	1	Büyük	150K	0
2	0	Orta	100K	0
3	0	Küçük	70K	0
4	1	Orta	120K	0
5	0	Büyük	95K	1
6	0	Orta	80K	0
7	1	Büyük	220K	0
8	0	Küçük	85K	1
9	0	Orta	75K	0
10	0	Küçük	90K	1

Öğrenme Kümesi

İTİ	Nil1	Nil2	Nil3	Sınıf
11	0	Küçük	85K	1
12	1	Orta	80K	0
13	1	Büyük	110K	0
14	0	Küçük	95K	0
15	0	Büyük	170K	0

Sınama Kümesi



<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

61

Büyük Veri Kümelerinde Değerlendirme

- Veri dağılımı dengeli ise: Veri kümesindeki örnek sayısı ve her sınıfa ait örnek sayısı fazla ise basit bir değerlendirme yeterli
 - holdout** yöntemi: Belli sayıda örnek sınama için ayrılır, geriye kalan örnekler öğrenme için kullanılır
 - genelde veri kümesinin 2/3'ü öğrenme, 1/3'i sınama kümesi olarak ayrılır
 - öğrenme kümesi kullanılarak model oluşturulur ve sınama kümesi kullanılarak model değerlendirilir

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

62

Tekrarlı Holdout Yöntemini

- Veri kümesini farklı altkümelere bölerek holdout yöntemini tekrarlama
 - Her eğitme işleminde veri kümesinin belli bir bölümü öğrenme kümesi olarak rasgele ayrılır
 - Modelin hata oranı, işlemler sonunda elde edilen modellerin hata oranlarının ortalaması
- Problem: Farklı eğitme işlemlerindeki sınama kümeleri örtüşebilir

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

63

k-Kat Çapraz Geçerleme

- Veri kümesi eşit boyutta k adet farklı gruba ayrılır.
- Bir grup sınama, diğerleri öğrenme için ayrılır.
- Her grup bir kere sınama kümesi olacak şekilde deneyler k kere tekrarlanır.

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

64

Biri Hariç Çapraz Geçerleme

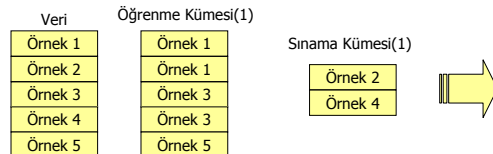
- k -kat çapraz geçerlemenin özel hali
 - k sayısı veri kümesindeki örnek sayısına (N) eşit
- Model $N-1$ örnek üzerinde eğitilir, dışarıda bırakılan 1 örnek üzerinde sınanır
- Bu işlem her örnek 1 kez sınama için kullanılacak şekilde tekrarlanır
 - model N kez eğitilir
- Model başarımı denemelerin başarımının ortalaması
- Verinin en etkin şekilde kullanımı

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

65

Bootstrap Yöntemi

- Veri kümesinden yerine koyma yöntemi ile örnekler seçilerek öğrenme kümesi oluşturulur
 - N örnekten oluşan veri kümesinden yerine koyarak N örnek seçilir
 - Bu küme öğrenme kümesi olarak kullanılır
 - Öğrenme kümesinde yer almayan örnekler sınama kümesi olarak kullanılır



<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

66

0.632 bootstrap

- N örnekten oluşan bir veri kümesinde bir örneğin seçilmeme olasılığı: $1 - \frac{1}{N}$
- Sınama kümesinde yer alma olasılığı:
$$\left(1 - \frac{1}{N}\right)^N \approx e^{-1} = 0.368$$
- Öğrenme kümesi veri kümesindeki örneklerin %63,2'sinden oluşuyor

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

67

Bootstrap Yönteminde Model Hatasını Belirleme

- Model başarımını sadece sınama kümesi kullanarak belirleme kötümser bir yaklaşım
 - model örneklerin sadece ~%63'lük bölümüyle eğitiliyor
- Model başarımı hem öğrenme kümesindeki hem de sınama kümesindeki başarımla değerlendirilir
$$\text{hata} = 0,632 \text{ hata}_{(\text{sınama})} + 0,368 \text{ hata}_{(\text{öğrenme})}$$
- İşlem birkaç kez tekrarlanarak hatanın ortalaması alınır.

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

68

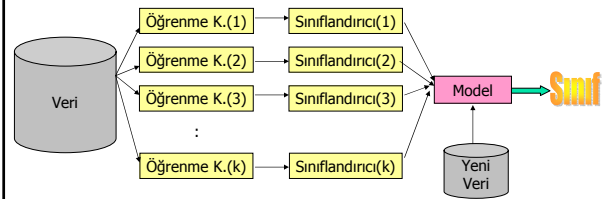
Konular

- Sınıflandırma yöntemleri
- Model Değerlendirme
- Öğrenme, sınama, geçerleme kümelerini oluşturma
- Sınıflandırıcıları birleştirme
 - Bagging
 - Boosting

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

69

Model Başarımını Artırma



- Bir grup sınıflandırıcı kullanma
 - Bagging
 - Boosting

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

70

Bagging

- N örnekten oluşan bir veri kümesinde bootstrap yöntemi ile T örnek seç
- Bu işlemi k öğrenme kümesi oluşturmak üzere tekrarla
- Aynı sınıflandırma algoritmasını k öğrenme kümesi üzerinde kullanarak k adet sınıflandırıcı oluştur
- Yeni bir örneği sınıflandırmak için her sınıflandırıcının sonucunu öğren
- Yeni örnek en çok hangi sınıfa atanmışsa o sınıfın etiketiyle etiketlenir.

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

71

Boosting

- Öğrenme kümesindeki her örneğin bir ağırlığı var
- Her öğrenme işleminden sonra, her sınıflandırıcı için yapılan sınıflandırma hatasına bağlı olarak örneklerin ağırlığı güncelleniyor
- Yeni bir örneği sınıflandırmak için her sınıflandırıcının doğruluğuna bağlı olarak ağırlıklı ortalaması alınıyor.

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

72