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

- Sınıflandırıcıların Değerlendirilmesi
- Skorlar
  - Karışıklık matrisi
  - Accuracy
  - Precision
  - Recall
  - Specificity
  - F-Score
- Eğitim ve Test Kümeleri
  - Hold-out set
  - Multiple random sampling
  - k-fold cross validation
  - Bootstrap

Sınıflandırıcıların Değerlendirilmesi

- Bir sınıflandırıcı geliştirildiğinde doğruluk değerinin belirlenmesi gereklidir.
- Doğruluğu test edilmeden bir sınıflandırıcı gerçek hayattaki problemlerde kullanılamaz.
- Bir sınıflandırıcının değerlendirilmesi için çok sayıda yöntem ve ölçüt vardır.
- **Temel değerlendirme ölçütü sınıflandırmadaki doğruluk oranıdır (accuracy)**.

$$Accuracy = \frac{\text{Number of correct classifications}}{\text{Total number of test cases}}$$

3

Sınıflandırıcıların Değerlendirilmesi

- Bazı uygulamalarda **hata oranı (error rate)** değeri kullanılmaktadır.

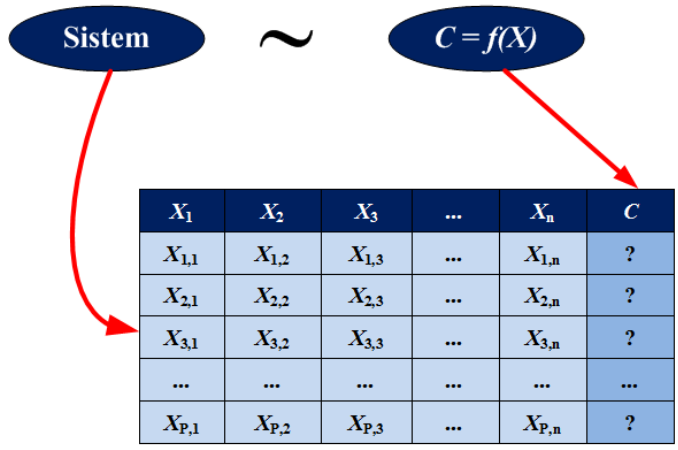
$$error\ rate = 1 - accuracy$$

- Genellikle istatistiksel testler kullanılarak farklı sınıflandırıcıların hangisinin daha uygun olduğuna karar verilebilir.
- Birden fazla sınıflandırıcıyı değerlendirmek için **aynı eğitim verisi ve aynı test verisi kullanılarak doğruluk oranları elde edilir.**

4

Sınıflandırıcıların Değerlendirilmesi

- Sınıflandırma probleminde, **giriş değerleri ile çıkış sınıfları arasında** ilişkilendirme yapan **bir fonksiyon bulunur.**



X_1	X_2	X_3	...	X_n	C
$X_{1,1}$	$X_{1,2}$	$X_{1,3}$	...	$X_{1,n}$	?
$X_{2,1}$	$X_{2,2}$	$X_{2,3}$	...	$X_{2,n}$	?
$X_{3,1}$	$X_{3,2}$	$X_{3,3}$	...	$X_{3,n}$	?
...	...	...	...	...	...
$X_{P,1}$	$X_{P,2}$	$X_{P,3}$	...	$X_{P,n}$	?

Sınıflandırıcıların Değerlendirilmesi

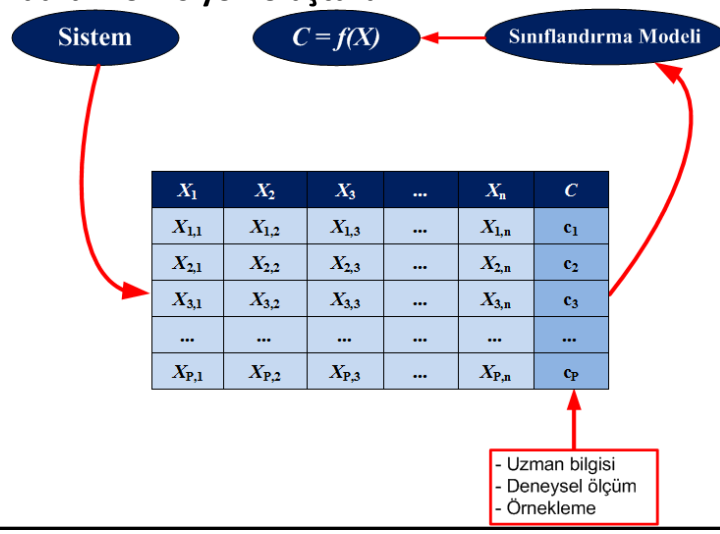
- Eğitim kümesinde **girişler ve çıkış/lar arasında** ilişkilendirmeler sağlanır.

EĞİTİM KÜMESİ

X_1	X_2	X_3	...	X_n	C
$X_{1,1}$	$X_{1,2}$	$X_{1,3}$	...	$X_{1,n}$	c_1
$X_{2,1}$	$X_{2,2}$	$X_{2,3}$	...	$X_{2,n}$	c_2
$X_{3,1}$	$X_{3,2}$	$X_{3,3}$	...	$X_{3,n}$	c_3
...	...	...	...	...	...
$X_{P,1}$	$X_{P,2}$	$X_{P,3}$	...	$X_{P,n}$	c_P

Sınıflandırıcıların Değerlendirilmesi

- **Sınıflandırma modeli,** eğitim kümesinde **giriş/ler ile çıkış/lar arasında bir fonksiyon oluşturur.**



X_1	X_2	X_3	...	X_n	C
$X_{1,1}$	$X_{1,2}$	$X_{1,3}$	...	$X_{1,n}$	c_1
$X_{2,1}$	$X_{2,2}$	$X_{2,3}$	...	$X_{2,n}$	c_2
$X_{3,1}$	$X_{3,2}$	$X_{3,3}$	...	$X_{3,n}$	c_3
...	...	...	...	...	...
$X_{P,1}$	$X_{P,2}$	$X_{P,3}$	...	$X_{P,n}$	c_P

Sınıflandırıcıların Değerlendirilmesi

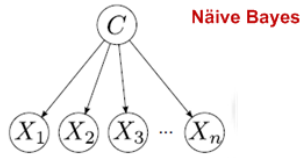
- Sınıflandırma modeli, **yeni girişler** için anlamlı sınıflandırma etiketleri belirler.

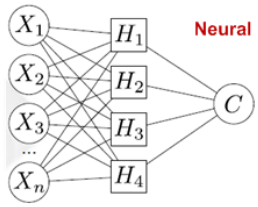
X_1	X_2	X_3	...	X_n	C
1	1	1	...	0	?
0	1	1	...	1	?
0	1	0	...	0	?
1	0	0	...	1	?
1	0	1	...	0	?

X_1	X_2	X_3	...	X_n	C
1	1	1	...	0	1
0	1	1	...	1	1
0	1	0	...	0	0
1	0	0	...	1	1
1	0	1	...	0	0

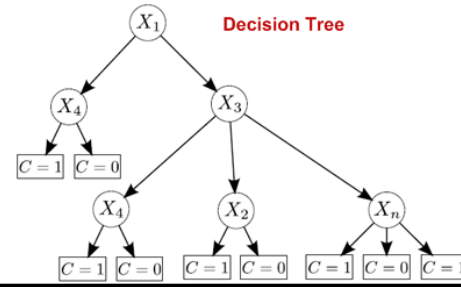
Sınıflandırıcıların Değerlendirilmesi

- Farklı sınıflandırma yaklaşımları kullanılabilir.





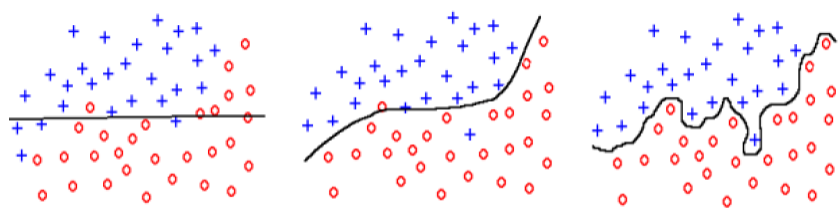
X_1	X_2	X_3	...	X_n	C
1	1	1	...	0	1
0	1	1	...	1	1
0	1	0	...	0	0
1	0	0	...	1	1
1	0	1	...	0	0



Sınıflandırıcıların Değerlendirilmesi

Underfit, fit ve overfit

- **Underfit,** eğitim kümesi için iyi düzeyde sınıflandırma yapamaz.
- **Fit,** hem eğitim kümesi hem de test kümesi için iyi düzeyde sınıflandırma yapabilir.
- **Overfit,** eğitim kümesi için iyi sınıflandırma yapar ancak test kümesi için iyi sınıflandırma yapamaz.



underfit fit overfit

Konular

- Sınıflandırıcıların Değerlendirilmesi
- Skorlar
  - Karışıklık matrisi
  - Accuracy
  - Precision
  - Recall
  - Specificity
  - F-Score
- Eğitim ve Test Kümeleri
  - Hold-out set
  - Multiple random sampling
  - k-fold cross validation
  - Bootstrap

11

Skorlar

- Bir sınıflandırıcının ne derece doğru sınıflandırma yaptığının gerçek uygulamada kullanılmadan önce belirlenmesi gereklidir.
- **Farklı eğitim kümeleri için aynı sınıflandırıcı farklı çözümler üretebilir.**
- **Aynı eğitim kümesi için de farklı sınıflandırıcılar farklı çözüm üretebilir** ve bunların en uygun olanının seçilmesi gereklidir.
- **Bir eğitim kümesi için aynı sınıflandırıcı modeli, farklı parametre değerleri için de farklı çözümler üretebilir.** Uygun parametre değerlerinin belirlenmesi gereklidir.
- **Kullanım amaçlarına**, **problemin büyüklüğüne** ve **beklenen doğruluk düzeyine göre** farklı skor değerleri kullanılarak sınıflandırıcılar değerlendirilebilir.

12

Konular

- Sınıflandırıcıların Değerlendirilmesi
- Skorlar
  - Karışıklık matrisi
  - Accuracy
  - Precision
  - Recall
  - Specificity
  - F-Score
- Eğitim ve Test Kümeleri
  - Hold-out set
  - Multiple random sampling
  - k-fold cross validation
  - Bootstrap

13

Karışıklık matrisi

- Karışıklık (confusion) matrisi ile, örnek kümedeki gerçek sınıf etiketi ile tahmin edilen sınıf etiketi sayıları gösterilir.

		TAHMİN EDİLEN		TOPLAM
		C^+	C^-	
GERÇEK	C^+	TP True Pozitif (Hits)	FN False Negatif (Miss)	N^+ Gerçek Pozitif sayısı
GERÇEK	C^-	FP False Pozitif (Miss)	TN True Negatif (Hits)	N^- Gerçek Negatif sayısı
TOPLAM		$\check{N}^+$ Tahmin Pozitif sayısı	$\check{N}^-$ Tahmin Negatif sayısı	N Toplam Örnek sayısı

14

Karışıklık matrisi

- Çok sayıdaki sınıf için örnek aşağıdadır.

		TAHMİN EDİLEN					TOPLAM
		C_1	C_2	C_3	...	C_n	
GERÇEK	C_1	T_1	F_{12}	F_{13}	...	F_{1n}	N_1
	C_2	F_{21}	T_2	F_{23}	...	F_{2n}	N_2
	C_3	F_{31}	F_{32}	T_3	...	F_{3n}	
	...	...	...	...	...	...	...
	C_n	F_{n1}	F_{n2}	F_{n3}	...	T_n	N_n
TOPLAM		$\hat{N}_1$	$\hat{N}_2$	$\hat{N}_3$	...	$\hat{N}_n$	N Toplam Örnek sayısı

15

Konular

- Sınıflandırıcıların Değerlendirilmesi
- Skorlar
  - Karışıklık matrisi
  - Accuracy
  - Precision
  - Recall
  - Specificity
  - F-Score
- Eğitim ve Test Kümeleri
  - Hold-out set
  - Multiple random sampling
  - k-fold cross validation
  - Bootstrap

16

Accuracy

- Bir sınıflandırıcının **doğru sınıflandırdığı örnek sayısının toplam örnek sayısına oranıdır.**

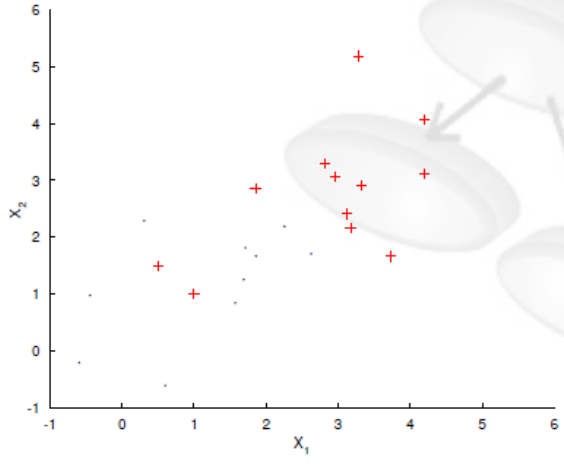
$$Accuracy = \frac{\text{Number of correct classifications}}{\text{Total number of test cases}}$$

- **Doğruluk değerlendirmesi test kümesi kullanılarak hesaplanır.**
- Eğitim sırasında kullanılmayan test verilerinde doğru sınıflandırdığı örnek sayısı alınarak doğruluk düzeyi hesaplanır.

17

Accuracy

- İyi bir sınıflandırıcının test kümesindeki tüm sınıfları doğru tahmin etmesi beklenir.

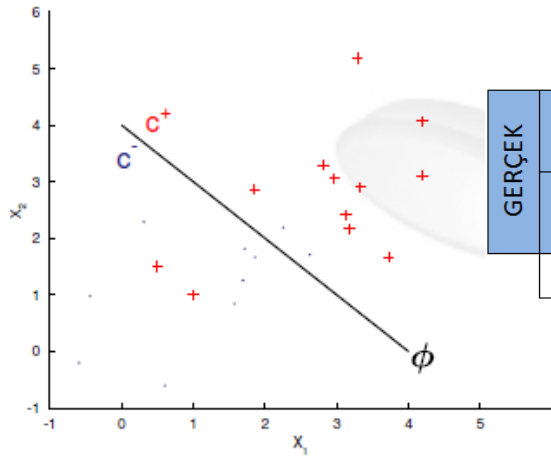


X_1	X_2	C
3,1	2,4	c^+
1,7	1,8	c^-
3,3	5,2	c^+
2,6	1,7	c^-
1,8	2,9	c^+
0,3	2,3	c^-
...	...	...

18

Accuracy

- İki sınıfa sahip bir sınıflandırıcı örneği.

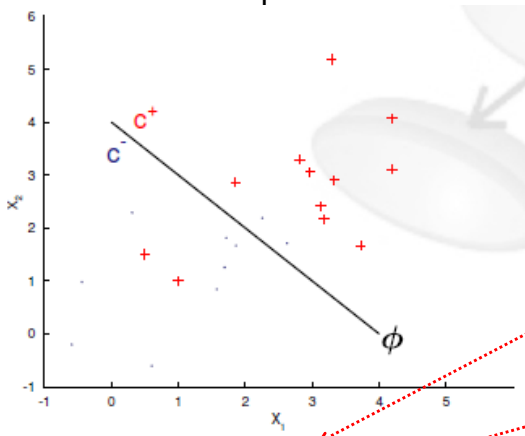


		TAHMİN		
		C^+	C^-	TOPLAM
GERÇEK	C^+	10	2	12
	C^-	2	8	10
	TOPLAM	12	10	22

19

Accuracy

- İki sınıfa sahip bir sınıflandırıcı örneği.

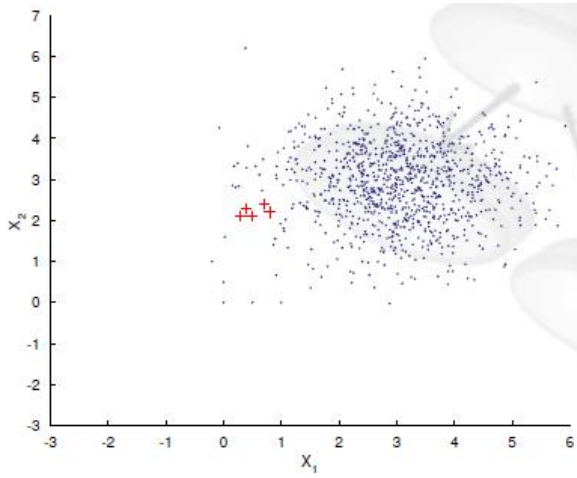


		TAHMİN		
		C^+	C^-	TOPLAM
GERÇEK	C^+	10	2	12
	C^-	2	8	10
	TOPLAM	12	10	22

- Accuracy = (TP + TN) / (N) = (10 + 8) / 22 = 0,82
- Hata = (FP + FN) / (N) = (2 + 2) / 22 = 0,18

20

Accuracy

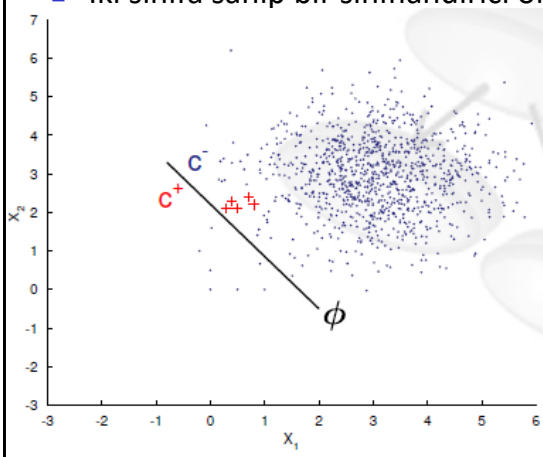


X_1	X_2	C
0,8	2,2	c^+
0,47	2,3	c^+
0,5	2,1	c^+
2,4	2,9	c^-
3,1	1,2	c^-
2,5	3,1	c^-
...	...	...

21

Accuracy

- İki sınıfa sahip bir sınıflandırıcı örneği.



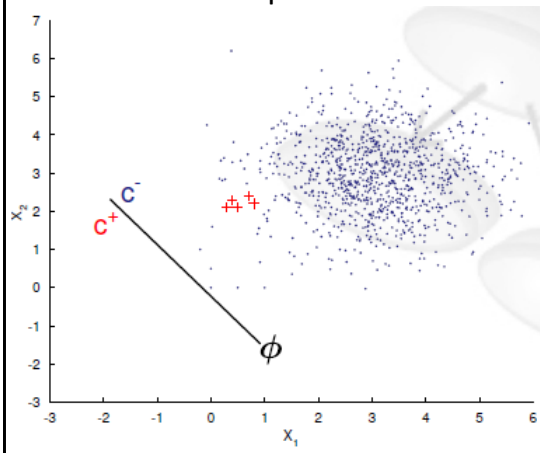
		TAHMİN		
		C^+	C^-	TOPLAM
GERÇEK	C^+	0	5	5
	C^-	7	993	1000
	TOPLAM	7	998	1005

- Accuracy = (TP + TN) / (N) = (0 + 993) / 1005= 0,99
- Hata = (FP + FN) / (N) = (7 + 5) / 1005 = 0,01 (Çok düşük hata !!!)

22

Accuracy

- İki sınıfa sahip bir sınıflandırıcı örneği.



		TAHMİN		
		C^+	C^-	TOPLAM
GERÇEK	C^+	0	5	5
	C^-	0	1000	1000
	TOPLAM	0	1005	1005

- Accuracy = (TP + TN) / (N) = (0 + 1000) / 1005= 0,995
- Hata = (FP + FN) / (N) = (0 + 5) / 1005 = 0,005 (Çok daha düşük hata !!!)

23

Konular

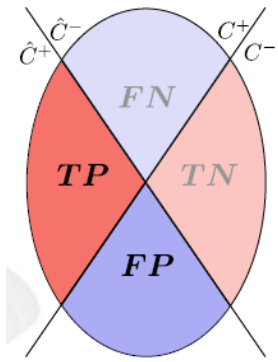
- Sınıflandırıcıların Değerlendirilmesi
- Skorlar
  - Karışıklık matrisi
  - Accuracy
  - Precision
  - Recall
  - Specificity
  - F-Score
- Eğitim ve Test Kümeleri
  - Hold-out set
  - Multiple random sampling
  - k-fold cross validation
  - Bootstrap

24

Precision

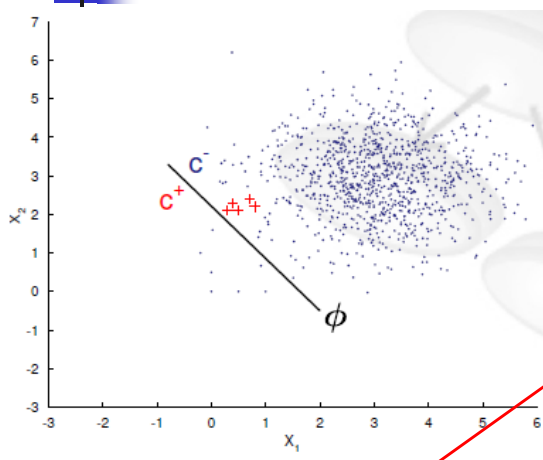
- Precision, gerçek değeri pozitif olup pozitif değere sınıflandırılan sayısının, pozitif değere sınıflandırılanların toplamına oranıdır.

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



25

Precision



		TAHMİN		
		C^+	C^-	TOPLAM
GERÇEK	C^+	0	5	5
	C^-	7	993	1000
	TOPLAM	7	998	1005

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{0}{0+7} = 0$$

(Çok kötü precison değeri !!!)

26

Konular

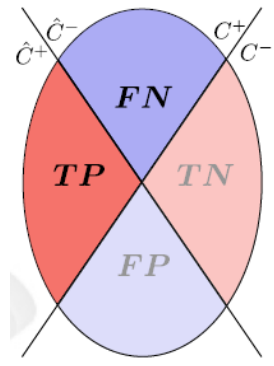
- Sınıflandırıcıların Değerlendirilmesi
- Skorlar
  - Karışıklık matrisi
  - Accuracy
  - Precision
  - Recall
  - Specificity
  - F-Score
- Eğitim ve Test Kümeleri
  - Hold-out set
  - Multiple random sampling
  - k-fold cross validation
  - Bootstrap

27

Recall

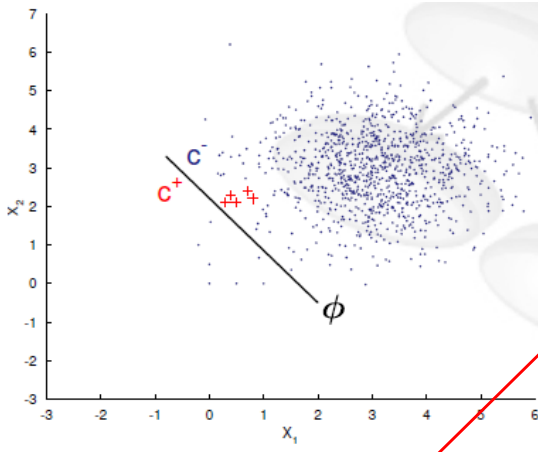
- Recall, gerçek değeri pozitif olup pozitif değere sınıflandırılan sayısının, gerçek değeri pozitif olanların tümüne oranıdır.

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



28

Recall



		TAHMİN		
		C^+	C^-	TOPLAM
GERÇEK	C^+	0	5	5
	C^-	7	993	1000
	TOPLAM	7	998	1005

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{0}{0+5} = 0$$ (Çok kötü recall değeri !!!)

29

Precision ve Recall Karşılaştırma

Spam filtreleme

- Precision: gerçekte spam olup da, spam kutusuna alınanların **spam kutusundaki tüm mesajlara oranıdır.**
- Recall: gerçekte spam olup ta spam kutusuna alınan mesajların, **gerçekte spam olan tüm mesajlara oranıdır.**

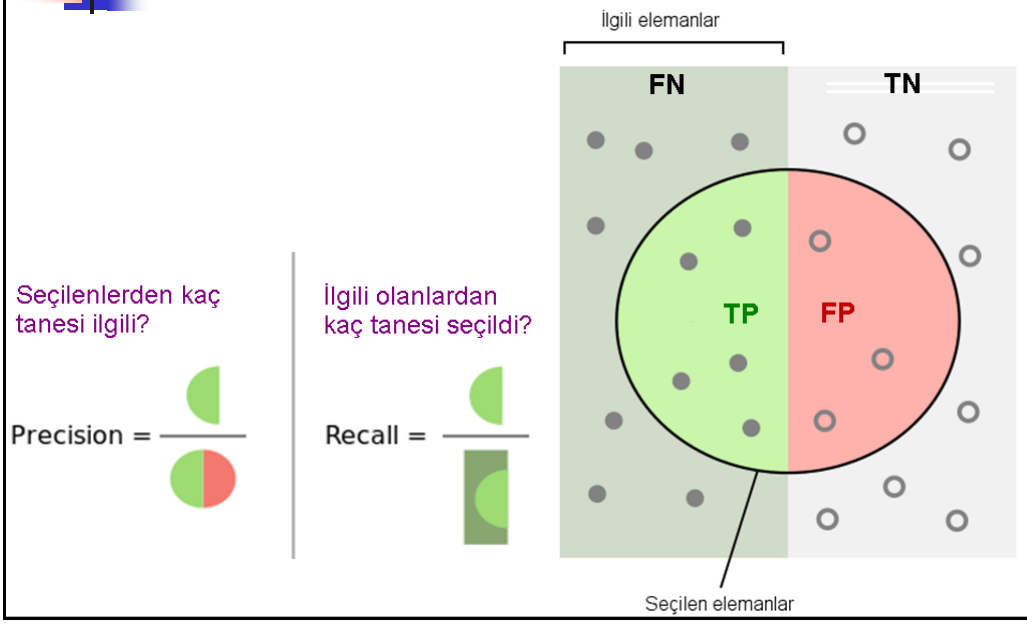
Duygu analizi

- Precision: gerçekte pozitif olup da pozitif sınıflandırılanların sayısının **pozitif olarak sınıflandırılanların tümüne oranıdır.**
- Recall: gerçekte pozitif olup da pozitif sınıflandırılanların **gerçekte pozitif olanların tümüne oranıdır.**

$$\text{Precision} = \frac{TP}{TP+FP} \qquad \text{Recall} = \frac{TP}{TP+FN}$$

30

Precision ve Recall Karşılaştırma



31

Konular

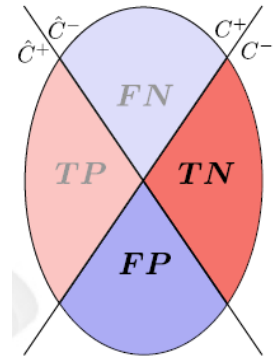
- Sınıflandırıcıların Değerlendirilmesi
- Skorlar
  - Karışıklık matrisi
  - Accuracy
  - Precision
  - Recall
  - Specificity
  - F-Score
- Eğitim ve Test Kümeleri
  - Hold-out set
  - Multiple random sampling
  - k-fold cross validation
  - Bootstrap

32

Specificity

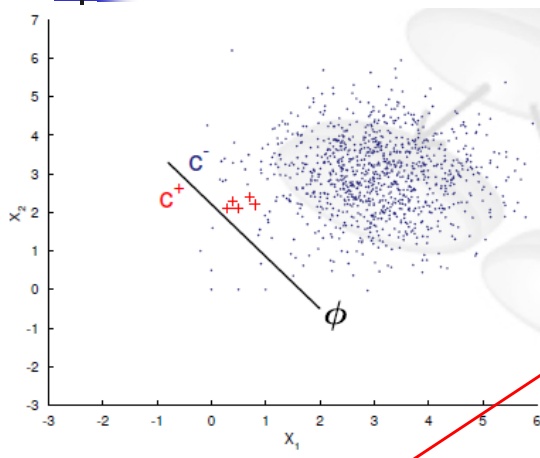
- Specifity, gerçek değeri negatif olup negatif sınıflandırılan sayısının, gerçek değeri negatif olanların tümüne oranıdır.

$$\text{Specificity} = \frac{TN}{TN + FP}$$



33

Specificity

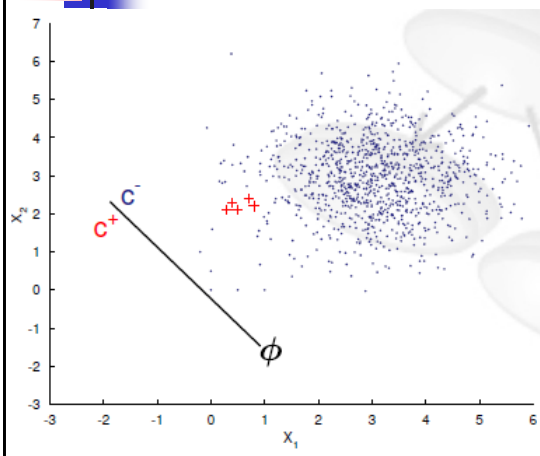


		TAHMİN		
		C^+	C^-	TOPLAM
GERÇEK	C^+	0	5	5
	C^-	7	993	1000
	TOPLAM	7	998	1005

$$\text{Specificity} = \frac{TN}{TN + FP} = \frac{993}{993 + 7} = 0{,}99$$

34

Specificity



		TAHMİN		
		C^+	C^-	TOPLAM
GERÇEK	C^+	0	5	5
	C^-	0	1000	1000
	TOPLAM	0	1005	1005

$$\text{Specificity} = \frac{TN}{TN + FP} = \frac{1000}{1000 + 0} = 1{,}00$$

35

Konular

- Sınıflandırıcıların Değerlendirilmesi
- Skorlar
  - Karışıklık matrisi
  - Accuracy
  - Precision
  - Recall
  - Specificity
  - F-Score
- Eğitim ve Test Kümeleri
  - Hold-out set
  - Multiple random sampling
  - k-fold cross validation
  - Bootstrap

36

F-Score

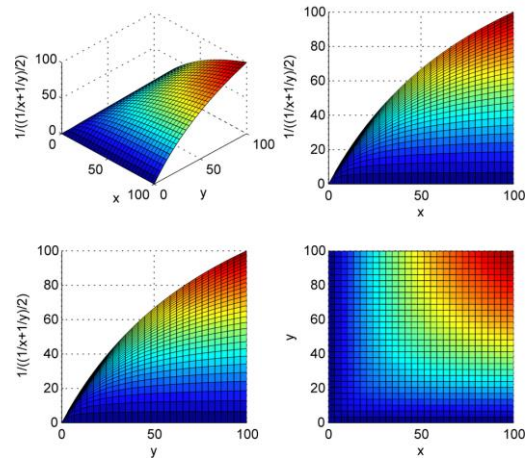
- F_1-score (Harmonic mean), iki sınıflandırıcının tek ölçüt ile değerlendirilmesi için kullanılır.

$$F-score=\frac{(\beta^2+1)\text{Precision}.\text{Recall}}{\beta^2(\text{Precision}+\text{Recall})}$$

$$F_1-score=\frac{2.\text{Precision}.\text{Recall}}{\text{Precision}+\text{Recall}}$$

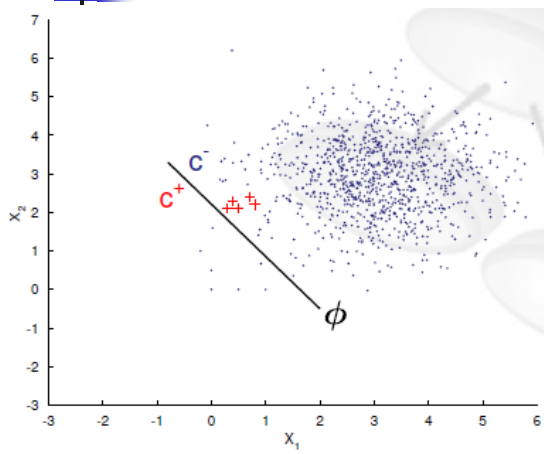
$$x_{harmonic\ mean}=\frac{n}{\sum_{k=1}^{n}\frac{1}{x_k}}$$

$$Harmonic\ mean=\frac{2}{\frac{1}{\text{Precision}}+\frac{1}{\text{Recall}}}=\frac{2.\text{Precision}.\text{Recall}}{\text{Precision}+\text{Recall}}$$



x ve y için f_1-score değişimi

F-Score



		TAHMİN		
		C^+	C^-	TOPLAM
GERÇEK	C^+	0	5	5
	C^-	7	993	1000
	TOPLAM	7	998	1005

$$F_1\text{-score}=\frac{2.\text{Precision}.\text{Recall}}{\text{Precision}+\text{Recall}}=\frac{2.0.0}{0+0}=0$$

F-Score

		TAHMİN		
		C^+	C^-	TOPLAM
GERÇEK	C^+	10	0	10
	C^-	0	5	5
	TOPLAM	10	5	15

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{10}{10+0} = 1$$

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{10}{10+0} = 1$$

$$F_1\text{-score} = \frac{2.\text{Precision}.\text{Recall}}{\text{Precision}+\text{Recall}} = \frac{2.1.1}{1+1} = 1$$

39

F-Score

		TAHMİN		
		C^+	C^-	TOPLAM
GERÇEK	C^+	0	10	10
	C^-	5	0	5
	TOPLAM	5	10	15

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{0}{0+5} = 0$$

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{0}{0+10} = 0$$

$$F_1\text{-score} = \frac{2.\text{Precision}.\text{Recall}}{\text{Precision}+\text{Recall}} = \frac{2.0.0}{0+0} = 0$$

40

F-Score

		TAHMİN		
		C^+	C^-	TOPLAM
GERÇEK	C^+	5	5	10
	C^-	5	5	10
	TOPLAM	10	10	20

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{5}{5+5} = 0{,}5$$

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{5}{5+5} = 0{,}5$$

$$F_1\text{-score} = \frac{2.\text{Precision}.\text{Recall}}{\text{Precision}+\text{Recall}} = \frac{2.(0{,}5).(0{,}5)}{0{,}5+0{,}5} = 0{,}5$$

41

F-Score

		TAHMİN		
		C^+	C^-	TOPLAM
GERÇEK	C^+	5	0	5
	C^-	15	0	15
	TOPLAM	20	0	20

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{5}{5+15} = 0{,}25$$

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{5}{5+0} = 1$$

$$F_1\text{-score} = \frac{2.\text{Precision}.\text{Recall}}{\text{Precision}+\text{Recall}} = \frac{2.(0{,}25).1}{0{,}25+1} = 0{,}4$$

42

Konular

- Sınıflandırıcıların Değerlendirilmesi
- Skorlar
  - Karışıklık matrisi
  - Accuracy
  - Precision
  - Recall
  - Specificity
  - F-Score
- Eğitim ve Test Kümeleri
  - Hold-out set
  - Multiple random sampling
  - k-fold cross validation
  - Bootstrap

43

Eğitim ve Test Kümeleri

- Bir sınıflandırıcının eğitimi ve testi için sınırlı sayıda veri elde edilebilir.
- **Eğitim verisi arttıkça** daha iyi **öğrenme ve genelleme sağlanabilmektedir.**
- **Test verisi arttıkça** sınıflandırıcının **hata olasılığı daha iyi tahmin edilebilmektedir.**
- Eğitim ve test verileri birbirinden farklı oluşturulmalıdır.
- Bir eğitim döngüsü içerisinde aynı değerlere sahip eğitim ve test veri kümesi kullanılmamalıdır.

44

Konular

- Sınıflandırıcıların Değerlendirilmesi
- Skorlar
  - Karışıklık matrisi
  - Accuracy
  - Precision
  - Recall
  - Specificity
  - F-Score
- Eğitim ve Test Kümeleri
  - Hold-out set
  - Multiple random sampling
  - k-fold cross validation
  - Bootstrap

45

Hold-out set

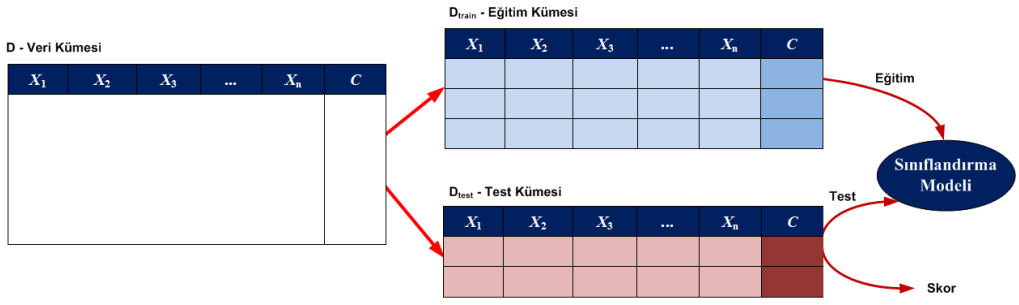
- Kullanılabilir veri kümesi D, **eğitim (D_{train})** ve **test (D_{test})** olmak üzere iki ayrı kümeye ayrılır.

  $D = D_{train} \cup D_{test}$, $D_{train} \cap D_{test} = \varnothing$
- Oluşturulan test kümesi **holdout set** olarak adlandırılır.
- Bu yöntem veri kümesi D büyükse kullanılabilir.
- **D veri kümesindeki tüm veriler bir sınıf etiketine atanmıştır.**
- **Training set** sınıflandırıcının **eğitimi için, test set** ise sınıflandırıcının **değerlendirilmesi için kullanılır.**
- **Training set için yüksek doğruluk** oranına sahip olan sınıflandırıcı, **test set için düşük doğruluk** düzeyine sahipse **overfit** yapılmıştır.
- İki küme için %50'şer veya 2/3 train ve 1/3 test için alınabilir.

46

Hold-out set

- Kullanılabilir veri kümesi D, eğitim (D_{train}) ve test (D_{test}) olmak üzere iki ayrı kümeye ayrılır.



47

Konular

- Sınıflandırıcıların Değerlendirilmesi
- Skorlar
  - Karışıklık matrisi
  - Accuracy
  - Precision
  - Recall
  - Specificity
  - F-Score
- Eğitim ve Test Kümeleri
  - Hold-out set
  - Multiple random sampling
  - k-fold cross validation
  - Bootstrap

48

Multiple random sampling

- Kullanılabilir veri kümesi D çok küçük boyutta ise, test kümesi çok daha küçük boyutta olacağından güvenilir sonuç vermez.
- Veri kümesinin küçük boyutta olduğu bu gibi durumlarda **n kez rastgele örnekleme** ile **eğitim** ve **test kümesi** oluşturulur.
- Bu durumda n tane doğruluk değeri elde edilir.
- **Sonuç doğruluk değeri**, elde edilen **doğruluk değerlerinin ortalaması alınarak hesaplanır.**

49

Konular

- Sınıflandırıcıların Değerlendirilmesi
- Skorlar
  - Karışıklık matrisi
  - Accuracy
  - Precision
  - Recall
  - Specificity
  - F-Score
- Eğitim ve Test Kümeleri
  - Hold-out set
  - Multiple random sampling
  - k-fold cross validation
  - Bootstrap

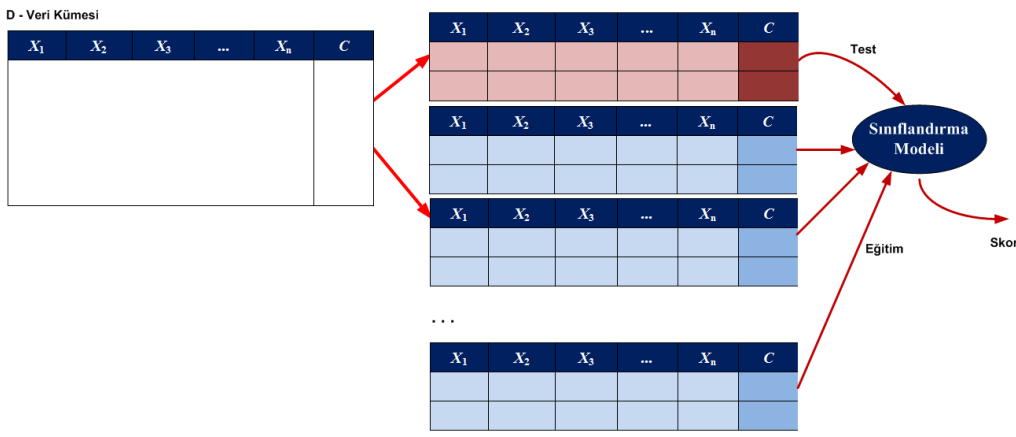
50

k-fold cross validation

- **Veri kümesi D küçük boyutta** olduğunda sık kullanılan yöntemdir.
- Bu yöntemde veri kümesi D, k adet eşit boyutta disjoint alt kümeye bölünür.
- Her (k-1) küme eğitim için kullanılırken kalan bir küme test için kullanılır.
- Bu işlem k kez tekrarlanır ve k adet doğruluk değeri elde edilir.
- Sonuç doğruluk değeri tüm doğruluk değerlerinin ortalaması alınarak hesaplanır.
- 5-fold ve 10-fold cross-validation literatürde farklı uygulamalarda yaygın kullanılmaktadır.
- **Hold-out set yöntemine göre daha iyi sonuç vermektedir.**

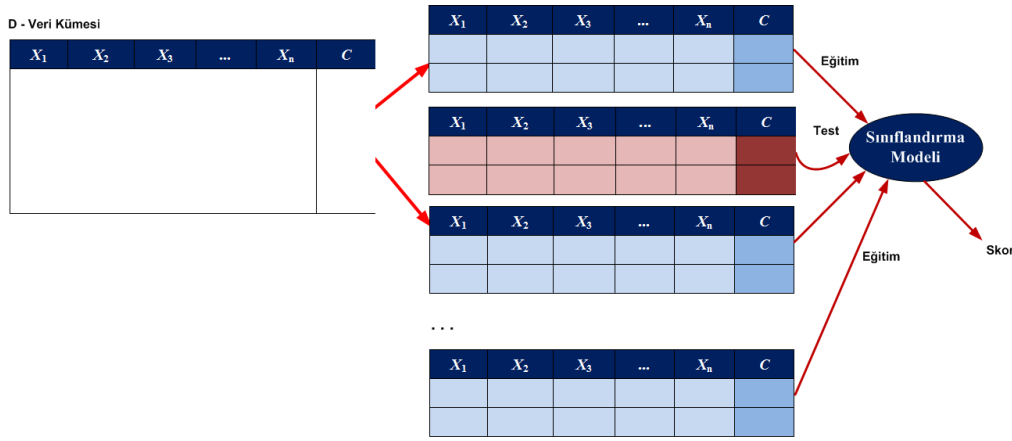
51

k-fold cross validation



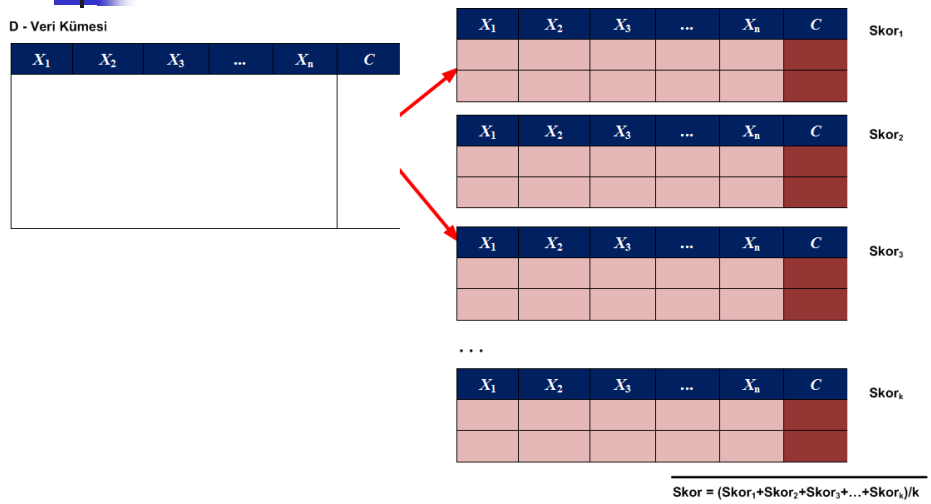
52

k-fold cross validation



53

k-fold cross validation



54

k-fold cross validation

Leaving-one-out cross validation

- k-fold cross validation yönteminin özel durumudur.
- **Eğitim kümesi, örnek sayısı kadar alt kümeye (k=N) bölünür.**
- Her iterasyonda **1 eleman dışarıda tutularak diğerlerinin tümü eğitim için kullanılır.**
- **Kalan 1 eleman ise test için kullanılır.**
- Sonuç doğruluk değeri tüm doğruluk değerlerinin toplamı alınarak hesaplanır.

55

Konular

- Sınıflandırıcıların Değerlendirilmesi
- Skorlar
  - Karışıklık matrisi
  - Accuracy
  - Precision
  - Recall
  - Specificity
  - F-Score
- Eğitim ve Test Kümeleri
  - Hold-out set
  - Multiple random sampling
  - k-fold cross validation
  - Bootstrap

56

Bootstrap

- **D veri kümesinden belirli sayıda değer rastgele seçilerek bir altküme elde edilir.**
- Elde edilen her altküme için ayrı model oluşturulur ve her birisinin doğruluk değeri ayrı ele alınır.
- **Sonuç doğruluk değeri ise tüm modellerden elde edilen doğruluk değerlerinin ortalaması alınarak hesaplanır.**
- **Test ve eğitim verileri tekrarlı olabilir.**
- **Overfit olma olasılığı vardır.**

D

	X	Y
1	4.3	2.4
2	2.1	1.1
3	5.3	2.8

D_1

	X	Y
3	5.3	2.8
1	4.3	2.4
3	5.3	2.8

D_2

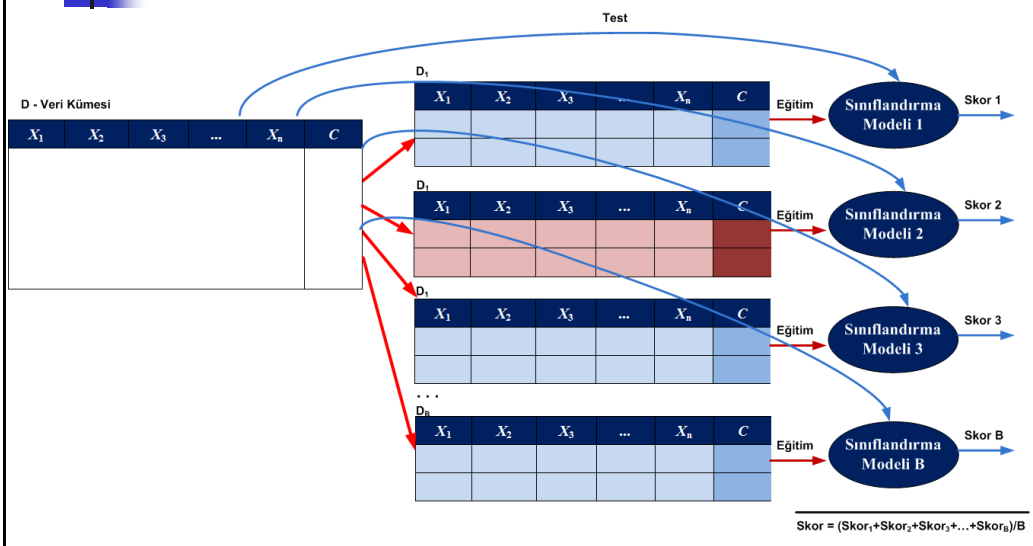
	X	Y
2	2.1	1.1
3	5.3	2.8
1	4.3	2.4

D_3

	X	Y
2	2.1	1.1
2	2.1	1.1
1	4.3	2.4

57

Bootstrap



58

Ödev

- Veri artırma (data augmentation) hakkında bir araştırma ödevi hazırlayınız.

59