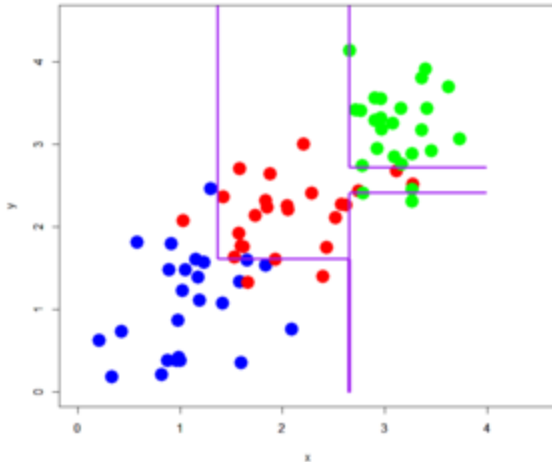
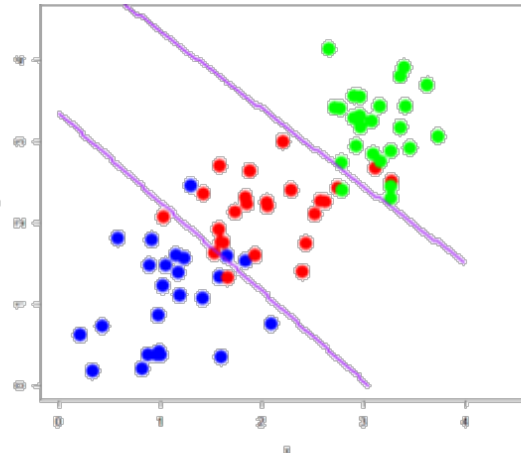


Regresyon ve Sınıflandırma

- Temel fark
 - Sınıflandırmada sıralı olmayan kategorik bir hedef değişken vardır.
 - Regresyon probleminde sürekli ya da sıralı bir hedef değişken vardır.
- Tüm regresyon yaklaşımları, sınıflandırma problemini çözmek için kullanılabilir.
- Sınıflandırma problemi doğru karar verme sınırını bulmaktan ibarettir (benim bakış açım)



veya



← **Ayrımcı**
yaklaşım karşı
Üretici
yaklaşım

Discriminative
versus
Generative

Sınıflandırma için doğrusal regresyon kullanımı

□ Doğrusal regresyon

$$\hat{Y} = \hat{\beta}_0 + \sum_{j=1}^p X_j \hat{\beta}_j$$

kesen ← → katsayılar

Öznitelik vektörünü kullanarak

$$X^T = (X_1, X_2, \dots, X_p)$$

$$\hat{Y} = X^T \hat{\beta},$$

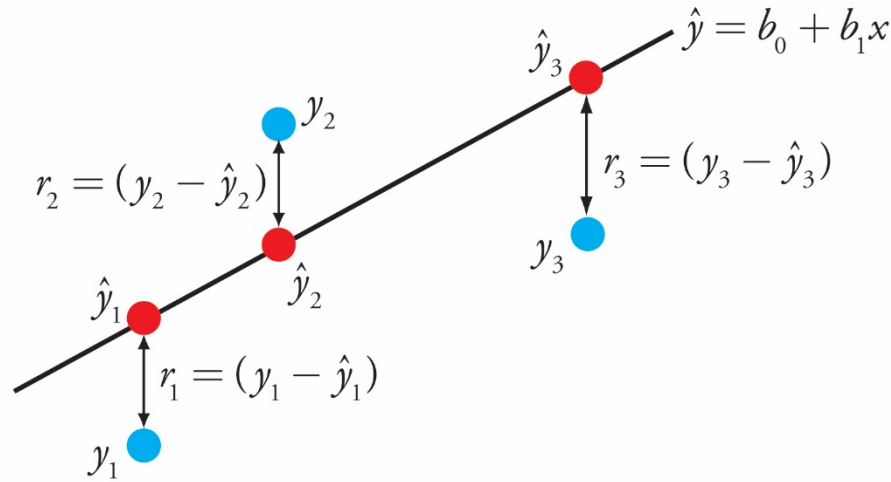
*Küçük kareler yöntemi (least-squares) ile elde edilir

Sürekli bir tahmin elde edebiliriz.

*Least-squares toplam hatayı enazlamayı hedefler

Sınıflandırma için doğrusal regresyon kullanımı

- Hataların karelerinin toplamını enazlama

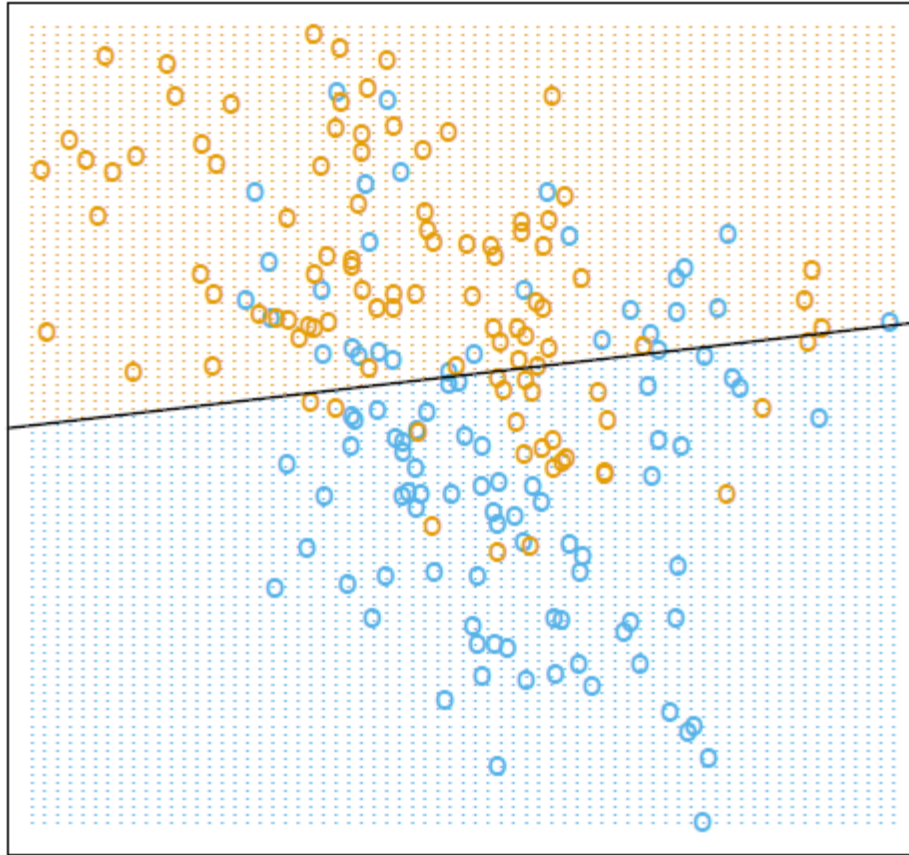


- İki sınıflı bir problemimiz olsun

X1	X2	Class
7.40	1.91	BLUE
3.92	0.24	ORANGE
2.15	1.08	ORANGE
-2.36	0.70	BLUE
.	.	.
.	.	.
.	.	.
0.09	-1.75	ORANGE
0.71	0.67	BLUE

Sınıflandırma için doğrusal regresyon kullanımı

Linear Regression of 0/1 Response



(BLUE = 0, ORANGE = 1),

ORANGE
 $\{x : x^T \hat{\beta} > 0.5\}$

Karar
verme sınırı
 $\{x : x^T \hat{\beta} = 0.5\}$

X1	X2	Class
7.40	1.91	0
3.92	0.24	1
2.15	1.08	1
-2.36	0.70	0
.	.	.
.	.	.
.	.	.
0.09	-1.75	1
0.71	0.67	0

Sınıflandırma için doğrusal regresyon kullanımı

□ Problemler

- Doğrusal regresyon varsayımları
 - Normal dağılan hatalar
- Sadece iki sınıflı problemler için çalışır
 - İki'den fazla farklı sınıf için?
- Kategorik ya da ordinal değişkenler
 - Sayısal gösterimi gerektirir
- Doğrusal olma zorunluluğu
 - Polinom terimler eklemek (örn. X^2)
 - Etkileşim terimleri eklemek (örn. XY)
- Pratikte bir kesme noktası (threshold) belirlemeyi gerektirir.

Sınıflandırma

Lojistik regresyon

- Neden? (Doğrusal regresyon tercih edilmiyor)
 - Y sadece iki farklı değer alabildiğinden e Normal dağılmamaktadır

$$\hat{Y} = \hat{\beta}_0 + \sum_{j=1}^p X_j \hat{\beta}_j + \text{hata (e)}$$

- Olasılıklar birden büyük ya da sıfırdan küçük çıkabilir.
- Lojistik regresyon bir dönüşüm ile $[0,1]$ tahminler üretir

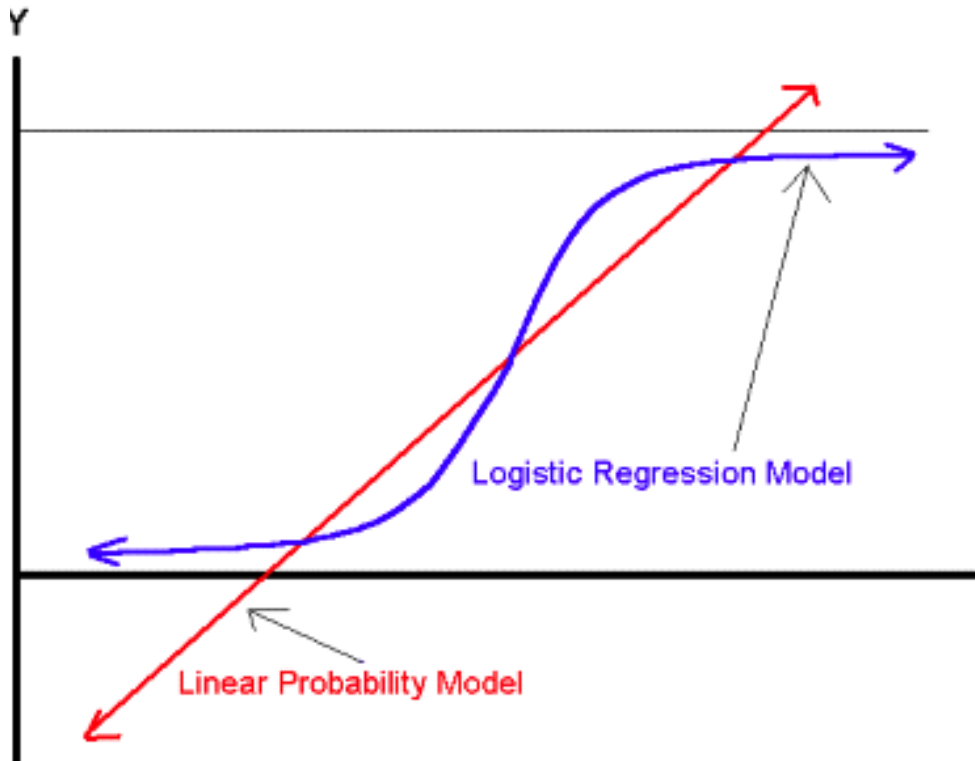
$$\Pr(G = 1|X = x) = \frac{\exp(\beta_0 + \beta^T x)}{1 + \exp(\beta_0 + \beta^T x)},$$

$$\Pr(G = 2|X = x) = \frac{1}{1 + \exp(\beta_0 + \beta^T x)}.$$

Sınıflandırma

Lojistik regresyon

Lojistik ve Doğrusal Regresyon



Doğrusal regresyon

$$\hat{Y} = \hat{\beta}_0 + \sum_{j=1}^p X_j \hat{\beta}_j.$$

Lojistik regresyon

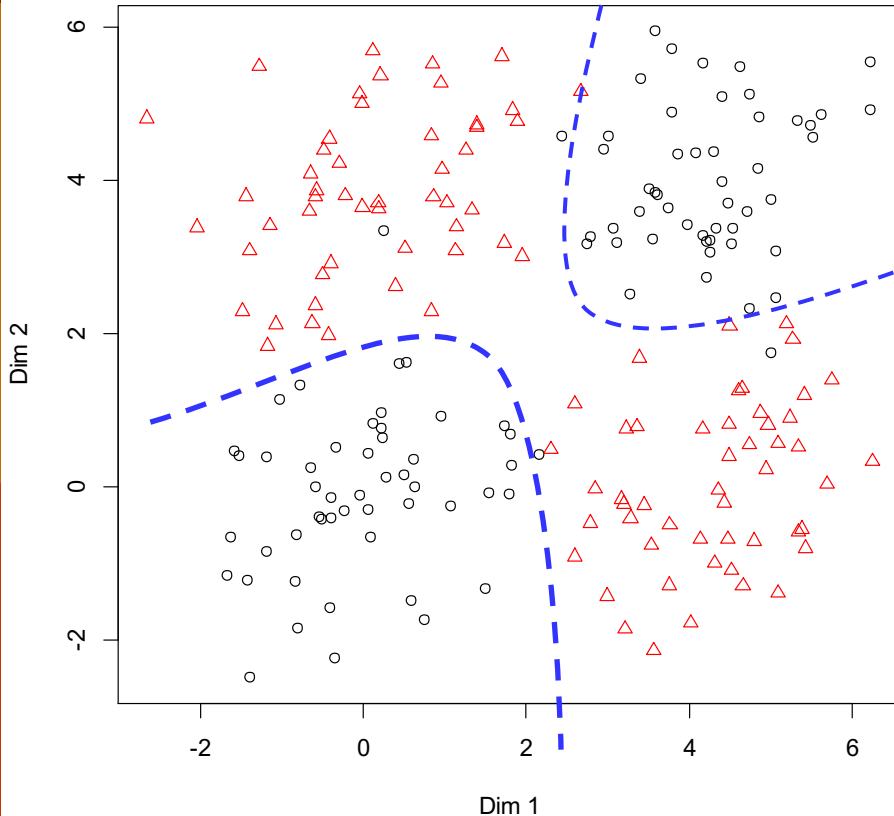
$$\Pr(G = 1|X = x) = \frac{\exp(\beta_0 + \beta^T x)}{1 + \exp(\beta_0 + \beta^T x)}$$

$$\Pr(G = 2|X = x) = \frac{1}{1 + \exp(\beta_0 + \beta^T x)}$$

Sınıflandırma

Doğrusal olmayan durumlar

A nonlinear case



Doğrusal karar verme sınırları çalışmazsa ne yapmalı?

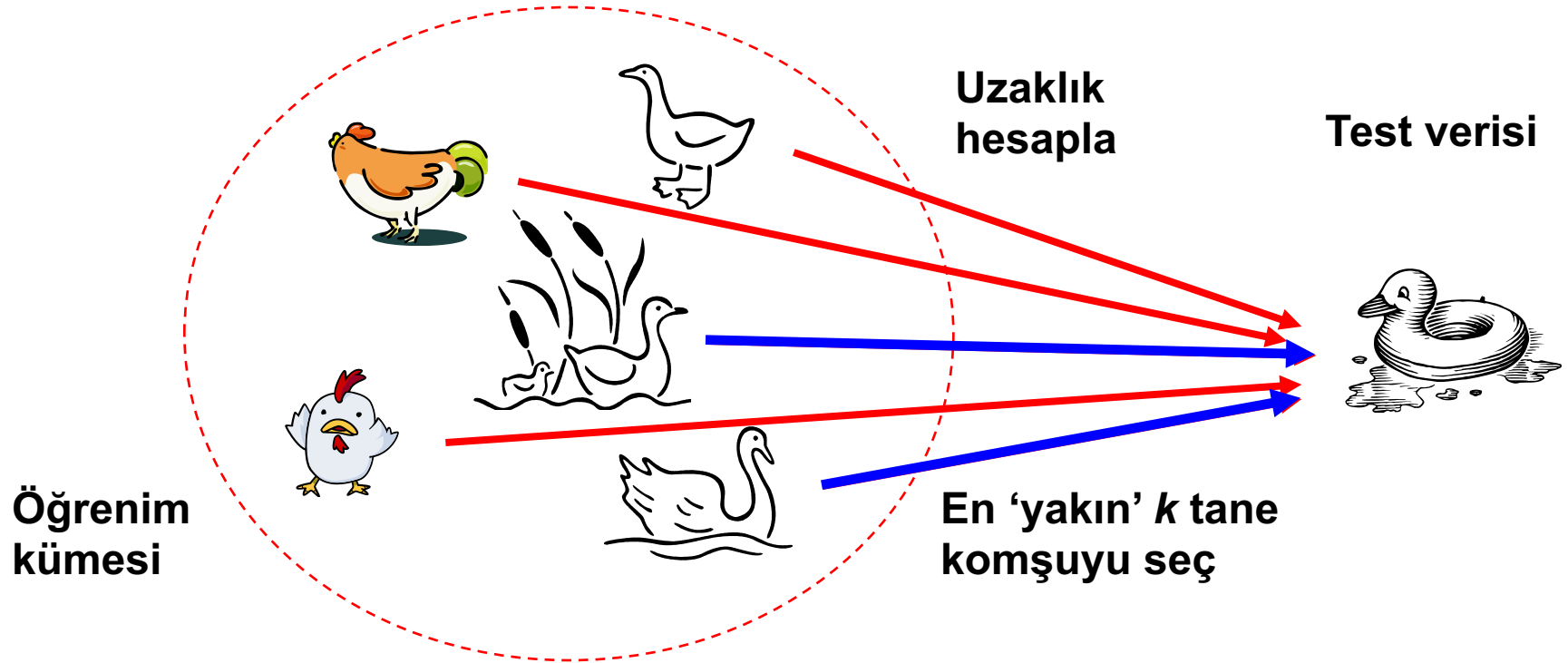
- Doğrusal olmayan terimlerin eklenmesi
 - Neler olabilir?
- Doğrusal olmayan ilişkileri modelleyebilen yöntemlerin kullanımı
 - Birden fazla model vardır.
 - En kolayı ve bilineni En Yakın Komşu (Nearest Neighbor-NN) yöntemidir.

Sınıflandırma

En Yakın Komşu

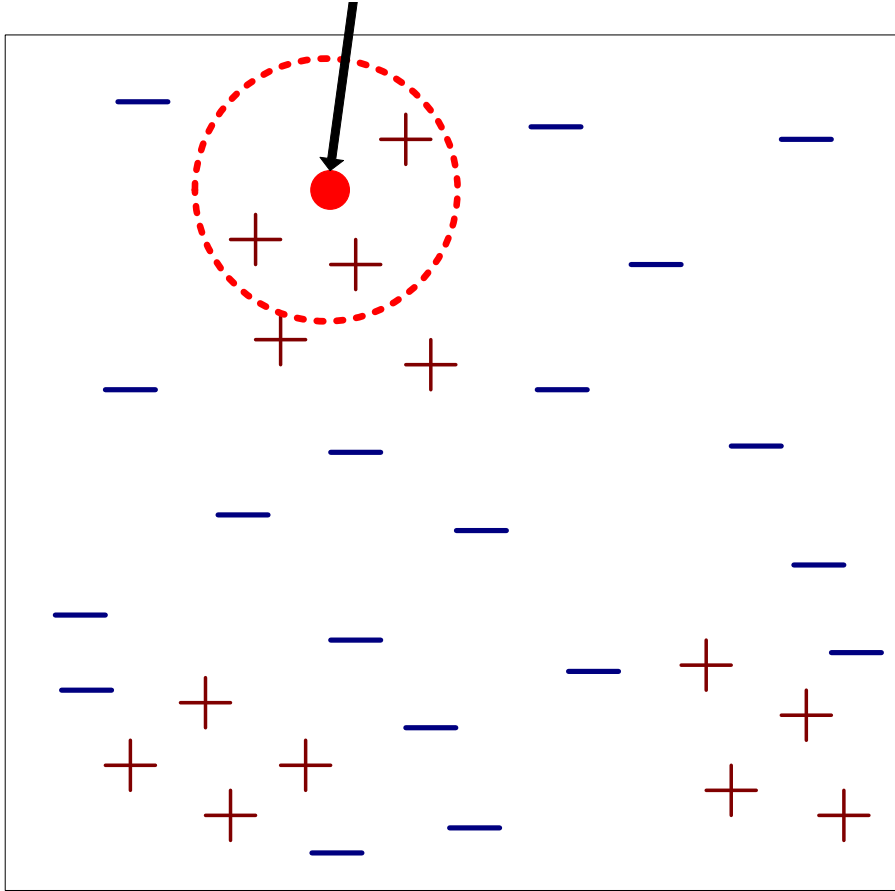
□ Temel fikir

- Eğer test hayvanı ördek gibi yürüyorsa, ördek gibi ses çıkarıyorsa, büyük ihtimal ile ördektir.



Sınıflandırma

En Yakın Komşu



- Üç şey gerektir
 - Kaydedilmiş öğrenme verisi
 - Uzaklık ölçüsü
 - k değeri, seçilecek en yakın komşu sayısı
- Sınıflandırma yapmak için
 - Öğrenim verisine uzaklıkları hesapla
 - En yakın k komşuyu bul
 - En yakın komşuların sınıfını kullanarak oylama sonucu sınıfa karar ver

Sınıflandırma

En Yakın Komşu

- **k**-en yakın komşu tahmini

$$\hat{Y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i,$$

$N_k(x)$ örnek x için k en yakın öğrenim verisi seti

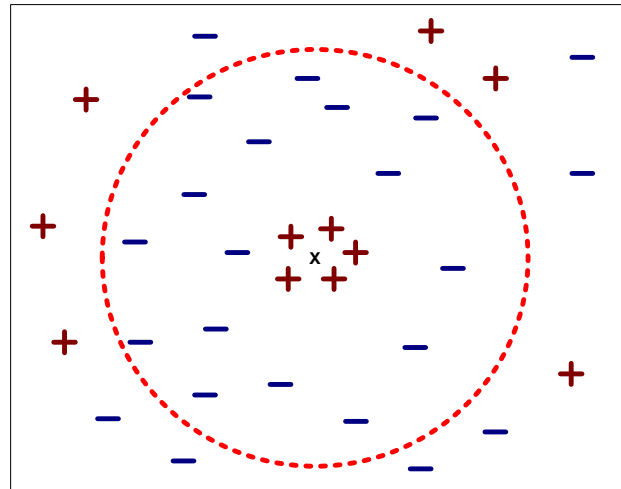
- En yakın noktaların y değerlerinin ortalaması
 - Regresyon problemi çözümü?
- Sınıflandırma için?
 - Mod alınabilir
- Ağırlık ortalaması?
 - Nasıl?

Sınıflandırma

En Yakın Komşu

□ k nasıl seçilmeli?

- Çok küçük k , gürültüden etkilenme
- Çok büyük k , komşuluk diğer sınıfları içerebilir.

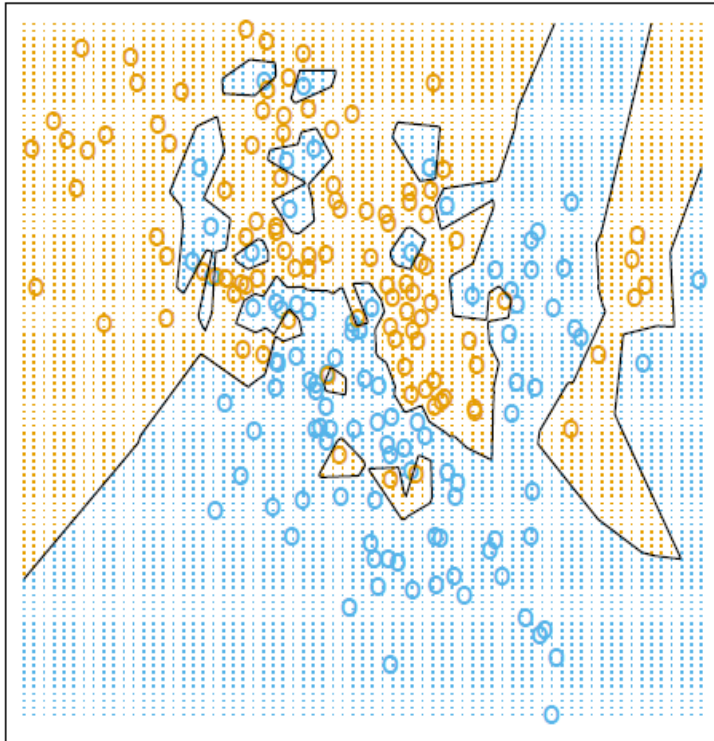


Sınıflandırma

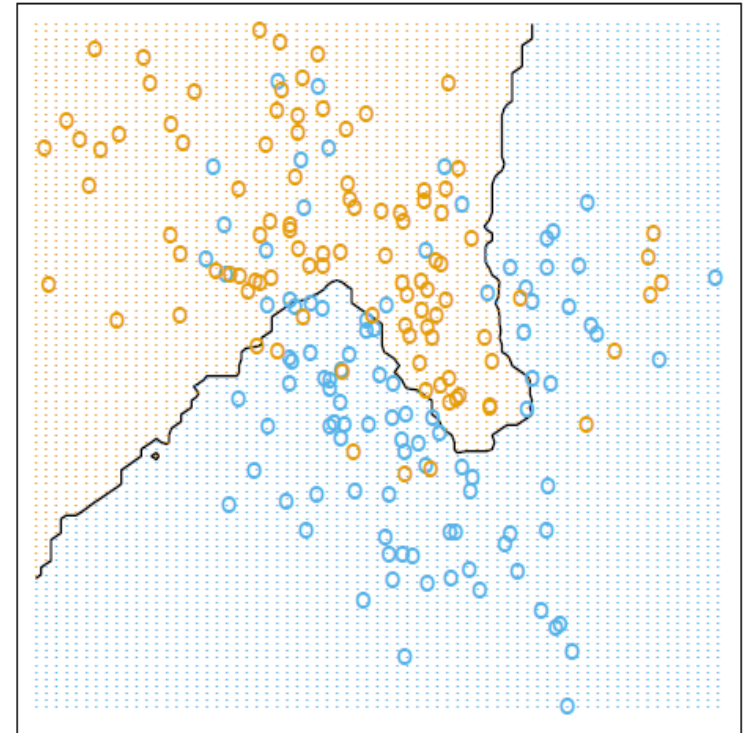
En Yakın Komşu

- Farkı k değerleri için karar verme sınırları

1-Nearest Neighbor Classifier



15-Nearest Neighbor Classifier



Sınıflandırma

En Yakın Komşu

- Tembel öğrenici (Lazy learner)
 - Altında bir model yok
 - Yorumu açık değil
 - Test verisindeki noktanın öğrenme verisindeki her noktaya uzaklığı hesaplanmalıdır
 - Gerçek zamanlı uygulamalar için problemlidir
 - Özellikle öğrenme verisi çok büyükse
 - Hafıza kullanımı açısından etkin değil
 - Öğrenme verisini saklamayı gerektirir
- Bir uzaklık/benzerlik ölçütü gerektirir
 - Öznitelik sayısı arttıkça anlamlı sonuç elde etmek zorlaşır (curse of dimensionality)
- Doğrusal olmayan sınırlar bulabilir

Sınıflandırma

En Yakın Komşu

- Örnek: Zaman serilerinde en yakın komşu bulma
 - EKG veri seti
 - http://www.cs.ucr.edu/~eamonn/time_series_data/
 - İki sınıflı sınıflandırma problemi. EKG verisi ile ritm bozukluğu yaşayan hastaları, sağlıklı hastalardan ayırma problemi
 - Zaman üzerinde 96 gözlem içeren 100 öğrenme verisi
 - 100 test verisi

Performans deęerlendirme

- Ölçütler
 - Performansı hangi ölçütler ile deęerlendirebiliriz?
- Performans ölçme yöntemleri nedir?
 - Modelin başarılı tahminler üretip üretmeyeceęi konusunda nasıl fikir elde edebiliriz?
- Farklı modelleri nasıl karşılaştırabiliriz?
 - Farklı modeller arası en iyi modeli nasıl seçeriz?

Performans deęerlendirme

- Sınıflandırma algoritması belirlemek birden çok aşamayı içerir
 - Öğrenme verisi kullanarak model oluşturulur
 - Test örnekleri üzerinde deęerlendirme yapılır
 - Yeni örnekler üzerinde tahminler elde edilir.

Öğren ve test et paradigması!

Değerlendirme kriteri

- Tahmin doğruluğu: Modelin yeni/model oluşturmada kullanılmamış veri üstünde doğru tahminler üretme başarısı
 - doğruluk=test verisinde doğru tahmin edilen gözlem yüzdesi
- Yorumlanabilirlik: Modelin anlaşılabilirliği ve tahmin problemi hakkında fikir vermesi
- Gürbüzlük(Robustness): Modelin farklı koşullar altında doğru tahmin yapabilmesi
 - Parametre değerleri değiştiğinde
 - Gürültülü ya da kayıp veriler olduğunda

Değerlendirme kriteri

- Hız: modeli kurarken ve kullanırken geçen zaman
- Ölçeklenebilirlik: veri seti büyüse dahi etkin biçimde model kurabilme
- Basitlik: Modelin çalışma prensiplerinin kolay anlaşılması

Model deęerlendirme

Performans ölçümü için metrikler

- Modelin tahmin kabiliyetine odaklanır
 - Hızlı ya da ölçeklenebilir olmayı hesaba katmaz
- Öğrenme kümesine ait olmayan veriler üzerinde deęerlendirmek anlamlıdır
 - N_t – test verisi sayısı
 - N_c – doęru tahmin edilen test örneęi

- Tahmin doęruluęu:
$$\eta = \frac{N_c}{N_t}$$

- Tahmin hatası:
$$\varepsilon = \frac{N_t - N_c}{N_t}$$

Model değerlendirme

Performans ölçümü için metrikler

- Kararsızlık matrisi:
 - İkili (2-sınıf) sınıflandırma

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a	b
	Class=No	c	d

- a: TP (gerçek pozitif)
- b: FN (yanlış negatif)
- c: FP (yanlış pozitif)
- d: TN (gerçek negatif)

- Kararsızlık matrisine bağlı olarak birden çok değerlendirme metriği kullanılmaktadır

Original classes	Predicted		
	K_1	K_2	K_3
K_1	50	0	0
K_2	0	48	2
K_3	0	4	46

3 sınıf içeren
problem örneği

Model değerlendirme

Performans ölçümü için metrikler

- Doğruluk kullanımı uygulamaya bağlı olarak problemlili olabilir
 - İki sınıflı problem
 - Sıfır sınıfından örnek sayısı= 9990
 - Bir sınıfından örnek sayısı = 10
 - Eğer model herşeyi sıfır olarak tahmin ederse tahmin başarısı= 99.9 %

□ Maliyet matrisi

$C(i | j)$: j sınıfından bir örneği i sınıfı olarak tahmin etmek

	PREDICTED CLASS		
	$C(i j)$	Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	$C(\text{Yes} \text{Yes})$	$C(\text{No} \text{Yes})$
	Class=No	$C(\text{Yes} \text{No})$	$C(\text{No} \text{No})$

Model deęerlendirme

Performans ölçümü için metrikler

- Farklı hata tiplerine deęişen maliyetler
- Tüm hataların maliyeti aynı deęildir
 - Kredi geri ödemeleri
 - Kredisini geri ödemeyen insan müşteri kaybından daha maliyetlidir
 - Tıbbi testler
 - Yanlış teşhisin hastalığı teşhis edememe durumuna göre maliyeti farklıdır
 - Spam
 - Spam e-postayı kaçırmak ile önemli bir maili spam diye gözden kaçıрма maliyeti farklıdır
- Algoritmaları yanlış pozitif ve yanlış negatifin bize yaratacağı maliyetleri hesaba katarak öğrenmek gerekir.

Model değerlendirme

Performans ölçümü için metrikler

□ Sınıflandırma maliyetini hesaplama

Cost Matrix	PREDICTED CLASS		
	C(i j)	+	-
ACTUAL CLASS	+	-1	100
	-	1	0

Model M ₁	PREDICTED CLASS		
		+	-
ACTUAL CLASS	+	150	40
	-	60	250

Doğruluk = 80%

Maliyet = 3910

Model M ₂	PREDICTED CLASS		
		+	-
ACTUAL CLASS	+	250	45
	-	5	200

Doğruluk = 90%

Maliyet = 4255

Model deęerlendirme yntemleri

- Performansı lebilen gvenilir bir lt nasıl bulunur?
 - Model performansı **ğrenme algoritmasının** yanı sıra birden ok etkene baęımlı olabilir:
 - Sınıf daęılımı
 - Yanlıř sınıflandırmanın maliyeti
 - ğrenme ve test etme veri byklę
- Tek nemli karar deęildir
- Problem nitelikleri nemlidir

Model deęerlendirme yntemleri

Kestirim stratejileri

- rnekleme ii (In-sample) deęerlendirme
 - Eldeki tm veriyi model parametrelerini ğrenmek iin kullanır
 - Ama modelin eldeki veriyi ne kadar iyi aıkladığı anlamaktır
 - Parametre sayısını azaltmak amalanır
- rnekleme dıŐı (out-of-sample) deęerlendirme
 - Split data into training and test sets
 - Focus: how well does my model predict things
 - Prediction error is all that matters
- İstatistik genelde rnekleme ii deęerlendirmeye odaklanır, makine ğrenmesi/veri madencilięi yntemleri ise rnekleme dıŐı deęerlendirmeyi n planda tutar.

Model değerlendirme yöntemleri

Kestirim stratejileri

- Dışarda bırakma (holdout)
 - Öğrenme için verinin 2/3 ve test için 1/3 ünü kullanma
- Cross validation (çapraz eşleme - bağımsız geçerlilik sınaması)
 - Veriyi k tane ayırık kümelere böl
 - k -fold: her seferin de bir kümeyi test verisi dışarda bırak, geri kalan $k-1$ kümeyi öğrenme verisi olarak kullan
 - Birini dışarda bırak (Leave-one-out): $k=n$
- Rastgele alt örneklem alma (Random subsampling)
 - Tekrarlayan dışarda bırakma
- Katmanlı örneklem alma (Stratified sampling)
 - Fazla örneklem alma (oversampling) ve az örneklem alma (undersampling)

Model deęerlendirme yntemleri

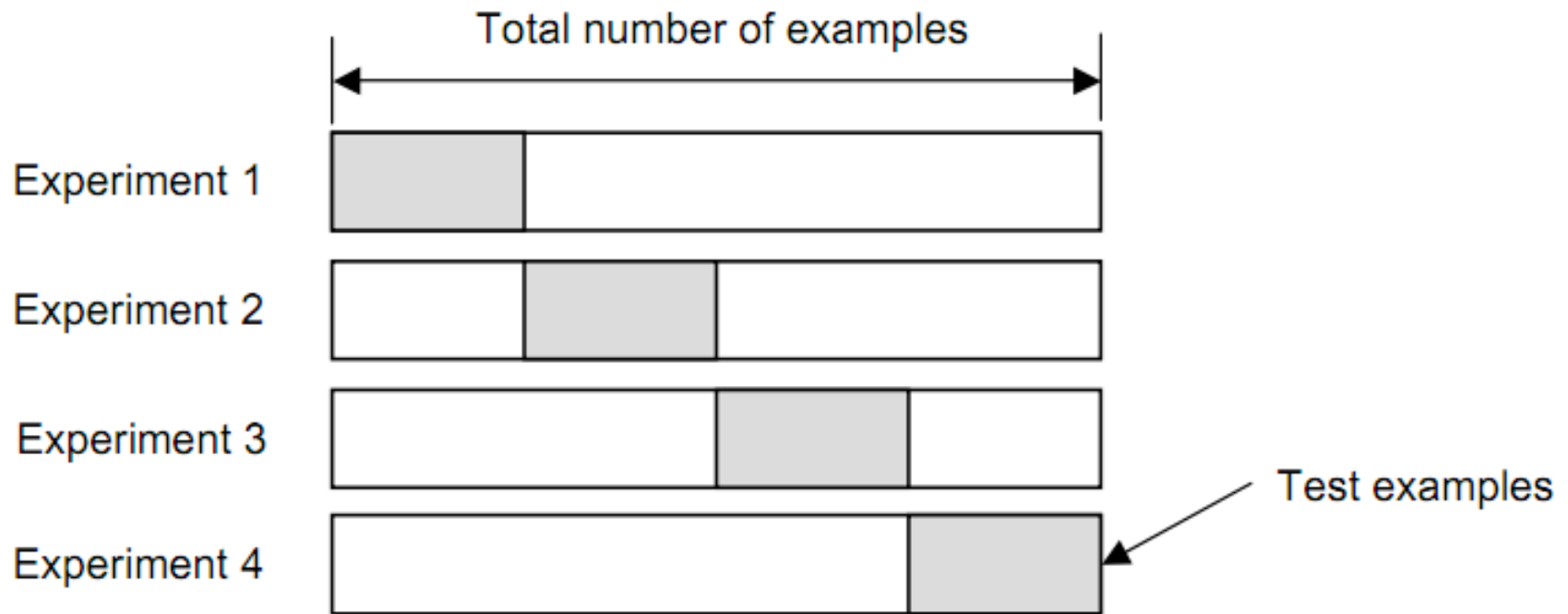
Kestirim stratejileri

- Dıřarda bırakma yntemi genellikle verinin 1/3 n test verisi 2/3 n ise ęrenme verisi olarak kullanır
- Sınıf daęılımı aısından dengesiz veri setleri (class imbalance) rastgele alınan rneklemler verideki gerek daęılımı yansıtmayabilir.
- *Katmanlı rneklem (Stratified sample):*
 - Her alt kmede sınıflar eřit oranlarda ele alınsın

Model değerlendirme yöntemleri

Kestirim stratejileri

□ Cross-validation



Model deęerlendirme yntemleri

Kestirim stratejileri

- Deęerlendirme iin kullanılan standart yntem: katmanlı 10'lu apraz eřleme (cross-validation)
- Neden 10? Birok veri setinde yeterli olduęu grlmřtr.
- Tekrarlayan katmanlı apraz eřleme model performansı hakkına daha iyi fikir verir.
 - Tek bir rastgele seimle oluřturulmuř apraz eřleme sonularına gvenmektense, birden ok rastgele seim yapmak
- Tekrarlarrrde elde edilen ortalama performans deęerlendirme iin en gvenilir olttr.

Model deęerlendirme yntemleri

Kestirim stratejileri

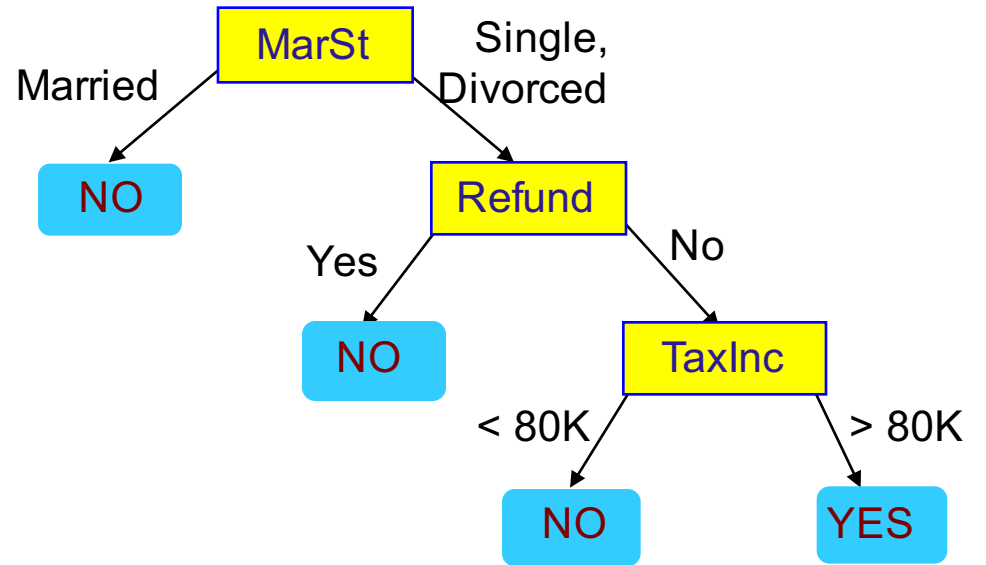
- Bir rneęi dıřarda bırakan aęraz eřleme (Leave-One-Out (LOO) cross-validation)
 - Veriyi en iyi řekilde kullanır
 - Rassallık iermez
 - Hesaplama zamanı aısından ktdr ama en gvenilir performans kestirimini saęlar
- Dıřarda bırakma yntemi ile elde edilen ngr rastgele seilen rneklemler zerinde tekrarlanarak daha iyi hale getirilebilir.
 - Her tekrarda katmanlı rastgele seim yap
 - Farklı tekrarlar zerinden yapılan hataların ortalamasını al
- Tekrarlayan dıřarda bırakma yntemi

Sınıflandırma

Ağaç-tabanlı yaklaşımlar

kategorik kategorik sürekli sınıf

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



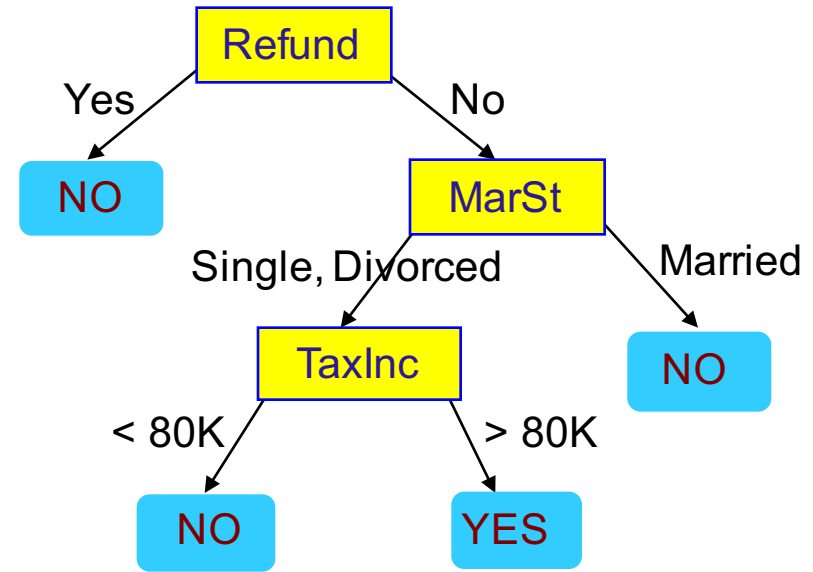
Aynı veriyi açıklayan birden fazla ağaç olabilir

Sınıflandırma

Ağaç-tabanlı yaklaşımlar

kategorik kategorik süreklî sınıf

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

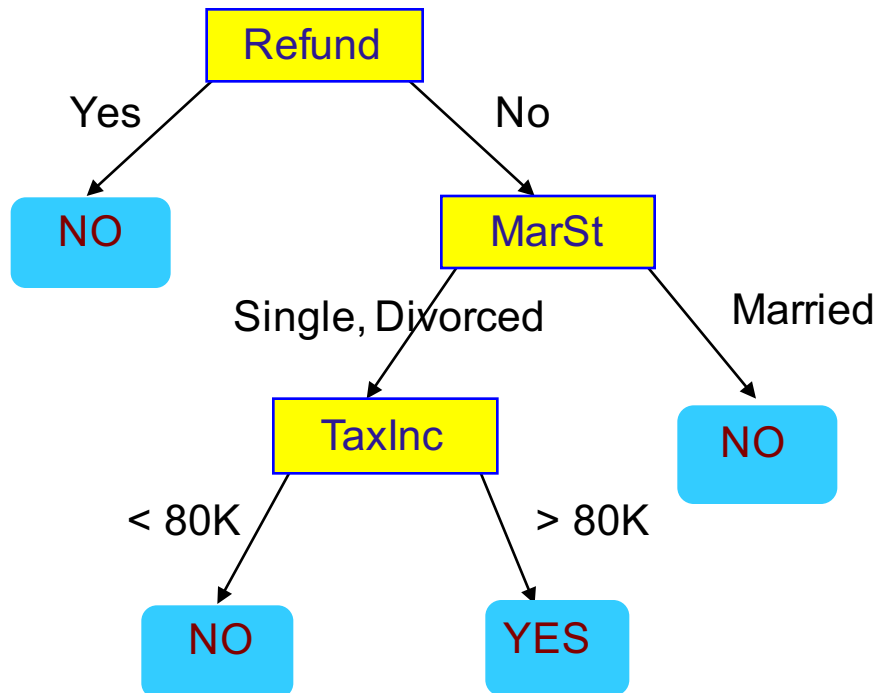


Öğrenme verisi

Model: Karar ağacı

Sınıflandırma

Ağaç-tabanlı yaklaşımlar



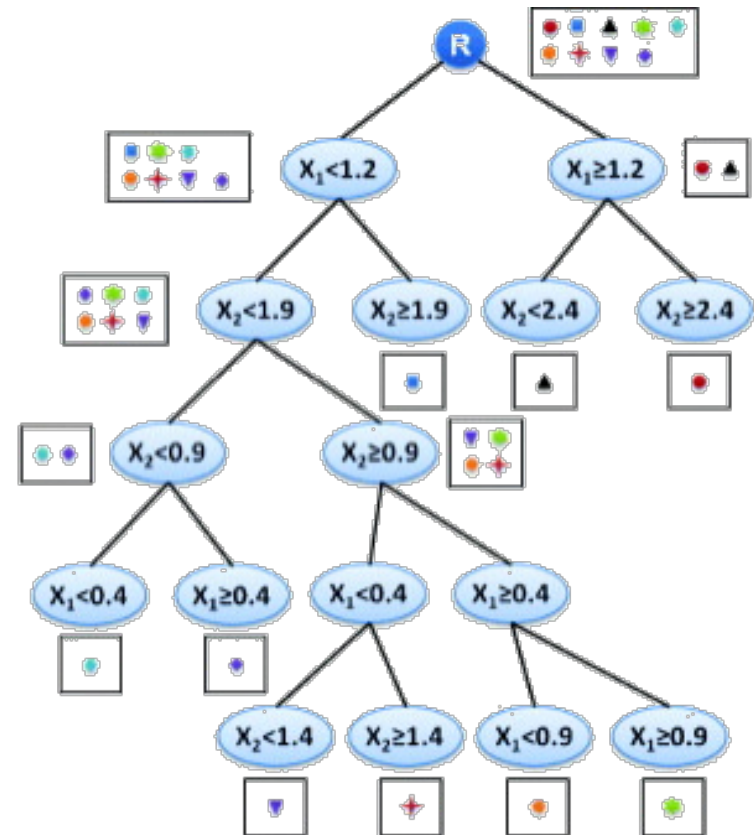
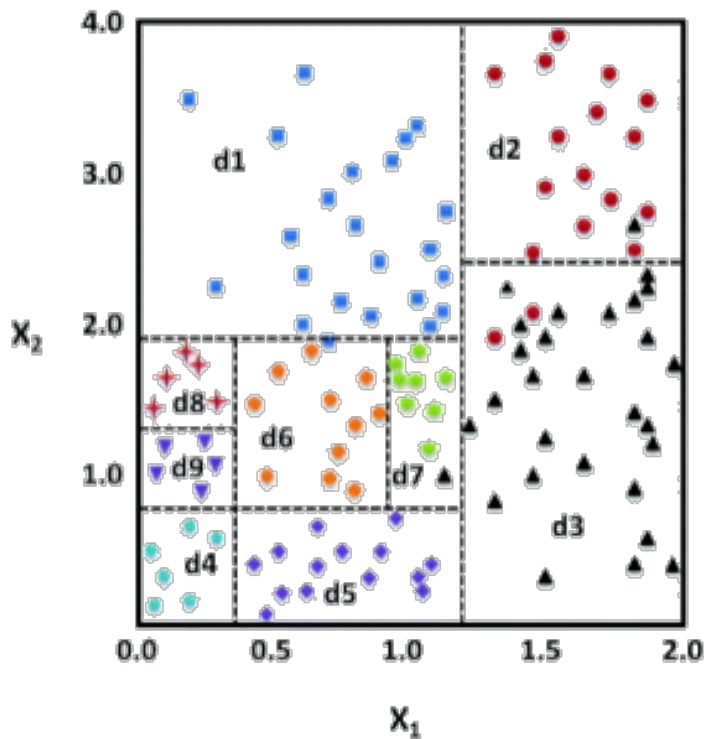
Test Verisi

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Sınıflandırma

Ağaç-tabanlı yaklaşımlar

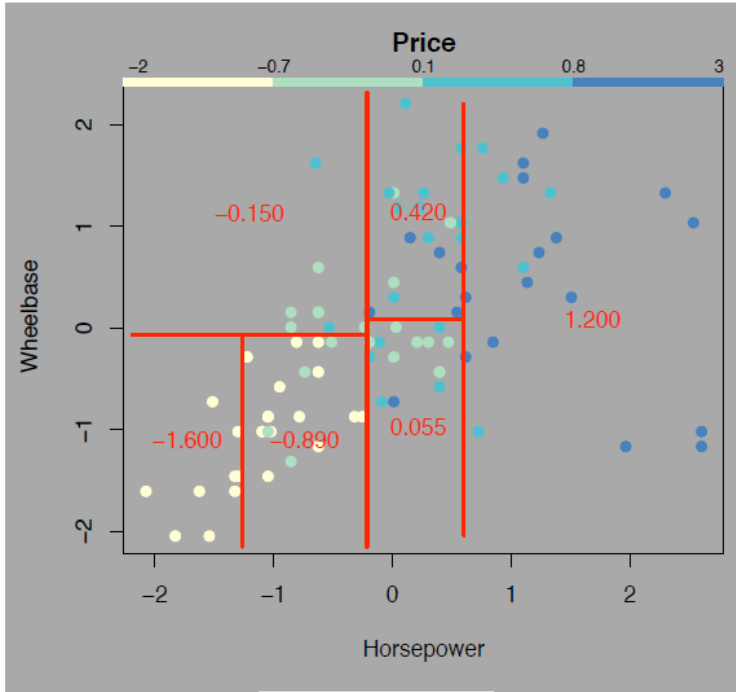
- Öznitelik uzayını bölme
 - Ardışık bölmelerle karar verme sınırı belirleme



Regresyon

Ağaç-tabanlı yaklaşımlar

İki değişkenimiz olsun (wheelbase and horsepower)



Bu değişkenlerle araba fiyatını tahmin eden bir karar ağacı kuralım

$$m(\mathbf{x}) = \sum_{i=1}^l k_i \times I(\mathbf{x} \in D_i)$$

X1	X2	Y
0.499	0.844	0.039
0.325	0.963	0.399
0.905	0.015	0.409
.	.	.
.	.	.
0.879	0.730	0.281

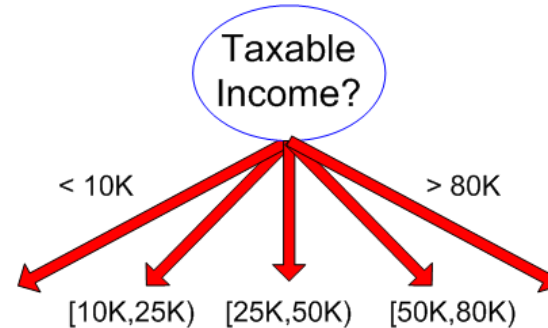
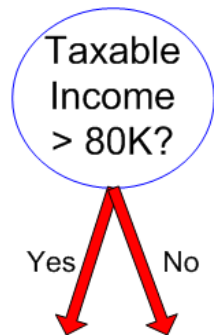
Toplam hatayı en aza indirecek bir D_i bulma

Ağaç-tabanlı yaklaşımlar

Öğrenme

□ Meseleler

- Uzayı nasıl bölmeli?
 - Tek seferde kaç parçaya bölmeli?
 - İkili bölümler
 - Çoklu bölümler
 - Bölme kuralı ne olmalı?
 - Hangi öznelik kullanılacak?
 - Kural ne olacak?
- Bölme ne zaman durdurulacak



Ağaç-tabanlı yaklaşımlar

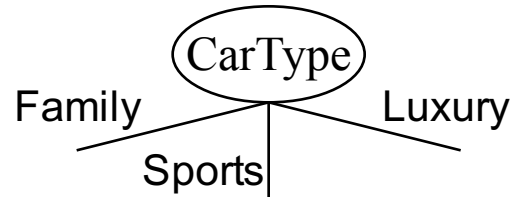
Uzayı bölme

- Öznitelik tiplerine göre değişir
 - Kategorik
 - Sıralı
 - Sürekli
- Bölme tiplerine göre değişir
 - İkili
 - Çoklu

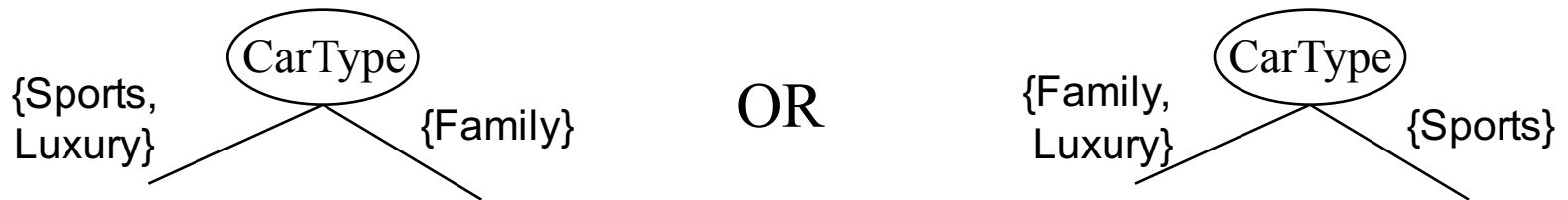
Ağaç-tabanlı yaklaşımlar

Uzayı bölme

- Kategorik değerleri bölme
 - Çoklu bölme



- İkili bölme

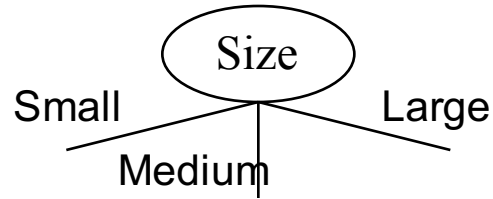


Ağaç-tabanlı yaklaşımlar

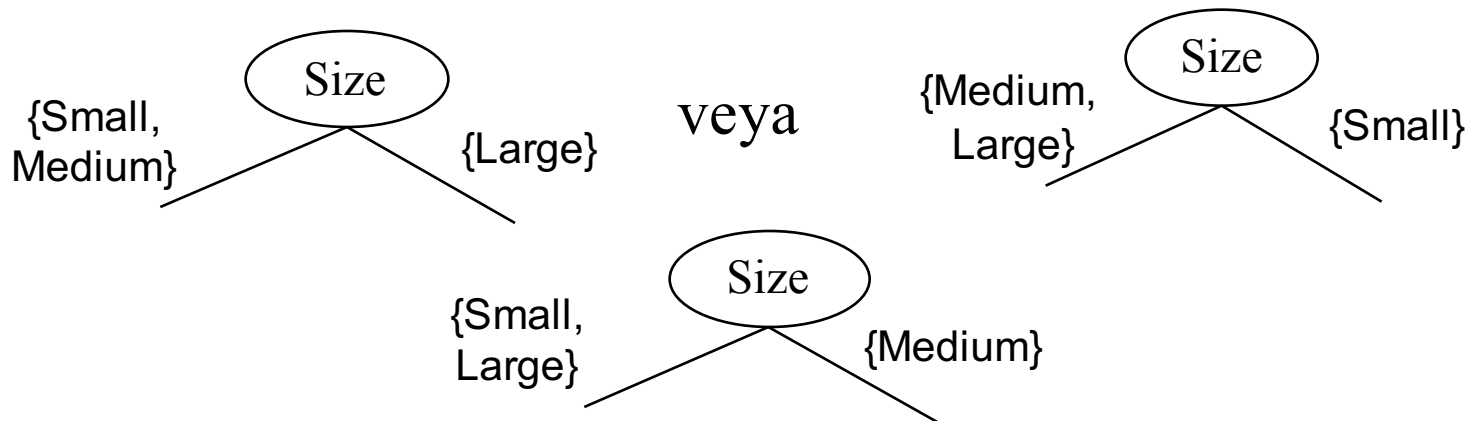
Uzayı bölme

□ Sıralı değerleri bölme

■ Çoklu bölme



■ İkili bölme



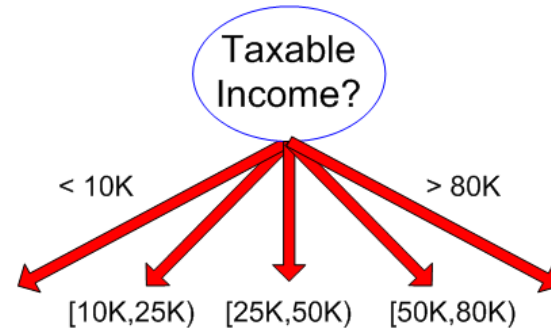
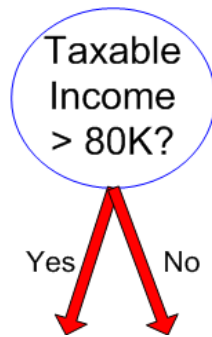
Ağaç-tabanlı yaklaşımlar

Uzayı bölme

□ Sürekli değerleri bölme

■ Birden fazla yöntem vardır

- Kesikli hale getirerek sıralı değişken elde etme
- İkili kararlar: $(A < v)$ or $(A \geq v)$
 - Olası her sınır arasından en iyisini seç
 - Hesaplama gerektirir.

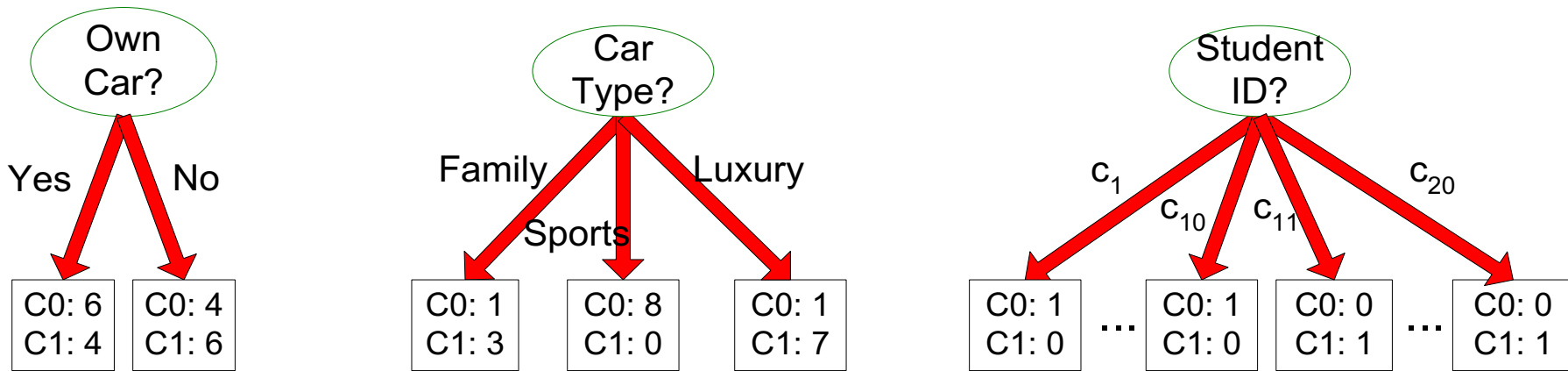


Ağaç-tabanlı yaklaşımlar

Uzayı bölme

□ En iyi karar sınırının bulma

Bölmeden önce: Sınıf 0 için 10 örnek,
Sınıf 1 için 10 örnek,



Hangisi en iyisi?

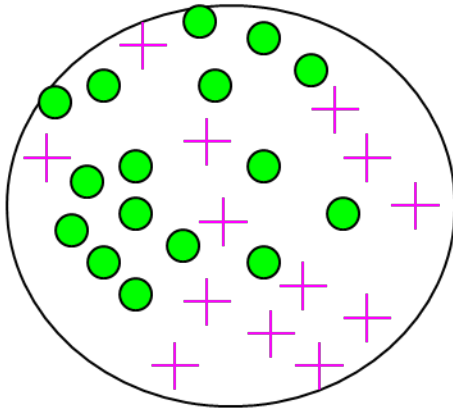
Ağaç-tabanlı yaklaşımlar

Uzayı bölme

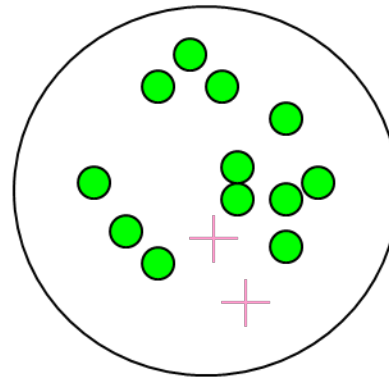
□ Impurity

- Bir grup gözlemin sınıf cinsinden homojen olma durumunun ölçütü

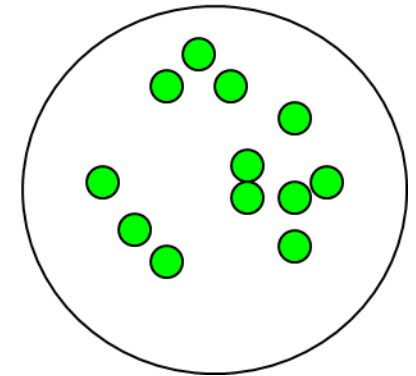
Very impure group



Less impure



Minimum impurity



Ağaç-tabanlı yaklaşımlar

Uzayı bölme

- Aç gözlü yaklaşım:
 - Homojen sınıf dağılımı içeren düğümler (node) bulmak
- Saflığın bir ölçütü gerekmektedir
 - Gini Index
 - Entropy
 - Tahmin hatası

C0: 5
C1: 5

Homojen değil

C0: 9
C1: 1

Homojen

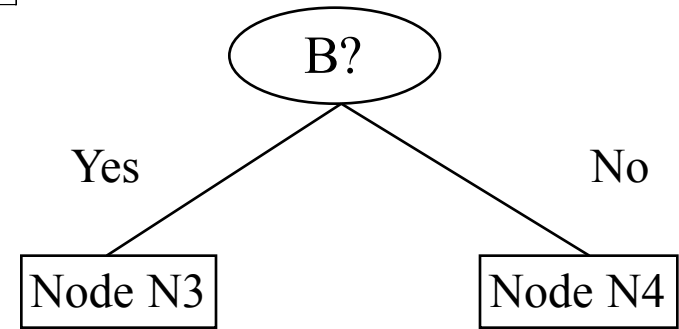
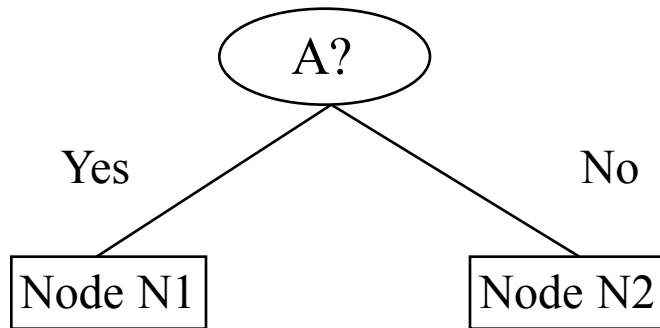
Ağaç-tabanlı yaklaşımlar

Uzayı bölme

Bölmeden önce

C0	N00
C1	N01

→ M0



C0	N10
C1	N11

C0	N20
C1	N21

C0	N30
C1	N31

C0	N40
C1	N41

↓
M1

↓
M2

↓
M3

↓
M4



M12

Kazanç = $M0 - M12$ vs $M0 - M34$

M34

Ağaç-tabanlı yaklaşımlar

Uzayı bölme

□ Homojenlik ölçütü

Düğüm t için Gini Index:

$$GINI(t) = 1 - \sum_j [p(j|t)]^2$$

(Not: $p(j|t)$ t düğümündeki j sınıfından gözlem oranı)

- Maksimum ($1 - 1/n_c$) eğer farklı sınıftaki gözlemler eşit olarak dağılmışsa
- Minimum (0.0) eğer tüm gözlemler bir sınıftansa

C1	0
C2	6
Gini=0.000	

C1	1
C2	5
Gini=0.278	

C1	2
C2	4
Gini=0.444	

C1	3
C2	3
Gini=0.500	

Uzayı bölme

Örnek Gini hesapları

$$GINI(t) = 1 - \sum_j [p(j | t)]^2$$

C1	0
C2	6

$$P(C1) = 0/6 = 0 \quad P(C2) = 6/6 = 1$$

$$Gini = 1 - P(C1)^2 - P(C2)^2 = 1 - 0 - 1 = 0$$

C1	1
C2	5

$$P(C1) = 1/6 \quad P(C2) = 5/6$$

$$Gini = 1 - (1/6)^2 - (5/6)^2 = 0.278$$

C1	2
C2	4

$$P(C1) = 2/6 \quad P(C2) = 4/6$$

$$Gini = 1 - (2/6)^2 - (4/6)^2 = 0.444$$

Uzayı bölme

Gini'ye göre bölüm bulma

- Bölümleme kalitesi aşağıdaki gibi bulunur

$$GINI_{split} = \sum_{i=1}^k \frac{n_i}{n} GINI(i)$$

n_i = çocuk node i deki gözlem sayısı,

n = başlangıç gözlem sayısı

Uzayı bölme

Gini'ye göre bölüm bulma - Kategorik

Çoklu bölüm

	CarType		
	Family	Sports	Luxury
C1	1	2	1
C2	4	1	1
Gini	0.393		

İkili bölüm

	CarType	
	{Sports, Luxury}	{Family}
C1	3	1
C2	2	4
Gini	0.400	

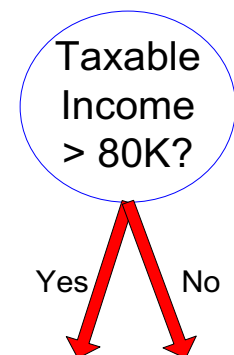
	CarType	
	{Sports}	{Family, Luxury}
C1	2	2
C2	1	5
Gini	0.419	

Uzayı bölme

Gini'ye göre bölüm bulma - Sürekli

- Bir değere göre ikili kararlar verilir
- Olası birden çok bölme değeri opsiyonu var
 - Olası bölme değerleri = Birbirinden farklı olan tüm değerler
- Her bölme değeri sınıfları sayar
 - Her bölüm için sınıfları say, $A < v$ and $A \geq v$
- En iyi v seçme yöntemi
 - Her v için sınıfları say
 - Hesaplama açısından etkin değil.

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



Uzayı bölme

Gini'ye göre bölüm bulma - Sürekli

- Etkin hesaplama için, her öznitelik için
 - Değerleri sırala
 - Sıralı değerlerin üstünden git ve sınıf sayılarını güncelle
 - En az gini değeri var noktayı seç

Cheat	No	No	No	Yes	Yes	Yes	No	No	No	No										
Taxable Income																				
Sıralanmış değerler →	60	70	75	85	90	95	100	120	125	220										
Bölüm opsiyonları →	55	65	72	80	87	92	97	110	122	172	230									
	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>				
Yes	0	3	0	3	0	3	1	2	2	1	3	0	3	0	3	0	3	0		
No	0	7	1	6	2	5	3	4	3	4	3	4	4	3	5	2	6	1	7	0
Gini	0.420	0.400	0.375	0.343	0.417	0.400	<u>0.300</u>	0.343	0.375	0.400	0.420									

Uzayı bölme

Alternatif bölme kriteri

- T düğümü (node) için entropy:

$$Entropy(t) = -\sum_j p(j | t) \log p(j | t)$$

- Düğüm homojenliğini ölçer
 - Maksimum ($\log n_c$) eğer sınıflar eşit dağılmışsa
 - Minimum 0 eğer düğüm homojen ise
- Entropy hesaplamaları gini ile benzerdir

Uzayı bölme

Alternatif bölme kriteri - Entropy

$$Entropy(t) = -\sum_j p(j | t) \log_2 p(j | t)$$

C1	0
C2	6

$$P(C1) = 0/6 = 0 \quad P(C2) = 6/6 = 1$$

$$Entropy = -0 \log 0 - 1 \log 1 = -0 - 0 = 0$$

C1	1
C2	5

$$P(C1) = 1/6 \quad P(C2) = 5/6$$

$$Entropy = - (1/6) \log_2 (1/6) - (5/6) \log_2 (1/6) = 0.65$$

C1	2
C2	4

$$P(C1) = 2/6 \quad P(C2) = 4/6$$

$$Entropy = - (2/6) \log_2 (2/6) - (4/6) \log_2 (4/6) = 0.92$$

Uzayı bölme

Bilgi kazanımı

□ Bilgi kazanımı (Information Gain)

$$GAIN_{split} = Entropy(p) - \left(\sum_{i=1}^k \frac{n_i}{n} Entropy(i) \right)$$

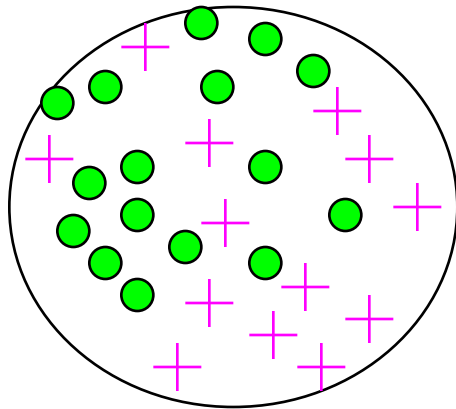
Ebeveyn node, p k bölüme ayrılırsa;
 n_i i bölümündeki gözlem sayısı

- Bilgi kazanımını en fazlayan bölümü seç
- Dezavantaj: Çok yüksek sayıda küçük homojen bölmelere neden olurç

Bilgi Kazanımı ölçümü

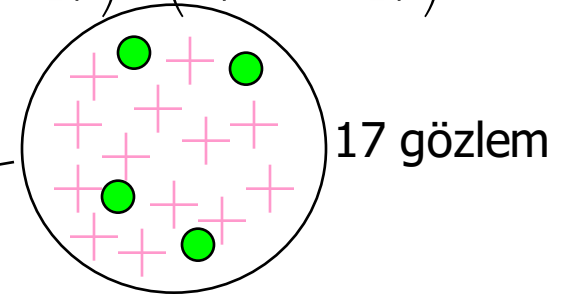
$$\text{Information Gain} = \text{entropy}(\text{parent}) - [\text{average entropy}(\text{children})]$$

Tüm veri (30 gözlem)

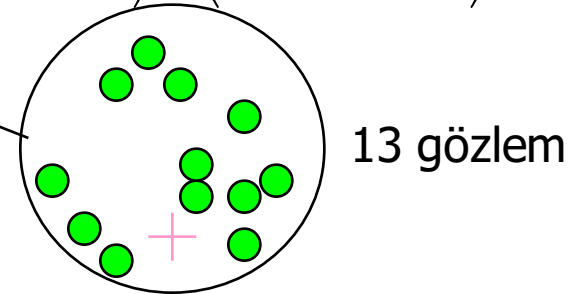


$$\text{parent entropy} = -\left(\frac{14}{30} \cdot \log_2 \frac{14}{30}\right) - \left(\frac{16}{30} \cdot \log_2 \frac{16}{30}\right) = 0.996$$

$$\text{child entropy} = -\left(\frac{13}{17} \cdot \log_2 \frac{13}{17}\right) - \left(\frac{4}{17} \cdot \log_2 \frac{4}{17}\right) = 0.787$$



$$\text{child entropy} = -\left(\frac{1}{13} \cdot \log_2 \frac{1}{13}\right) - \left(\frac{12}{13} \cdot \log_2 \frac{12}{13}\right) = 0.391$$



$$\text{Ortalama Entropy} = \left(\frac{17}{30} \cdot 0.787\right) + \left(\frac{13}{30} \cdot 0.391\right) = 0.615$$

$$\text{Bilgi kazanımı} = 0.996 - 0.615 = 0.38$$

Uzayı bölme

Hatadaki iyileşmeye göre

$$Error(t) = 1 - \max_i P(i | t)$$

C1	0
C2	6

$$P(C1) = 0/6 = 0 \quad P(C2) = 6/6 = 1$$

$$Error = 1 - \max(0, 1) = 1 - 1 = 0$$

C1	1
C2	5

$$P(C1) = 1/6 \quad P(C2) = 5/6$$

$$Error = 1 - \max(1/6, 5/6) = 1 - 5/6 = 1/6$$

C1	2
C2	4

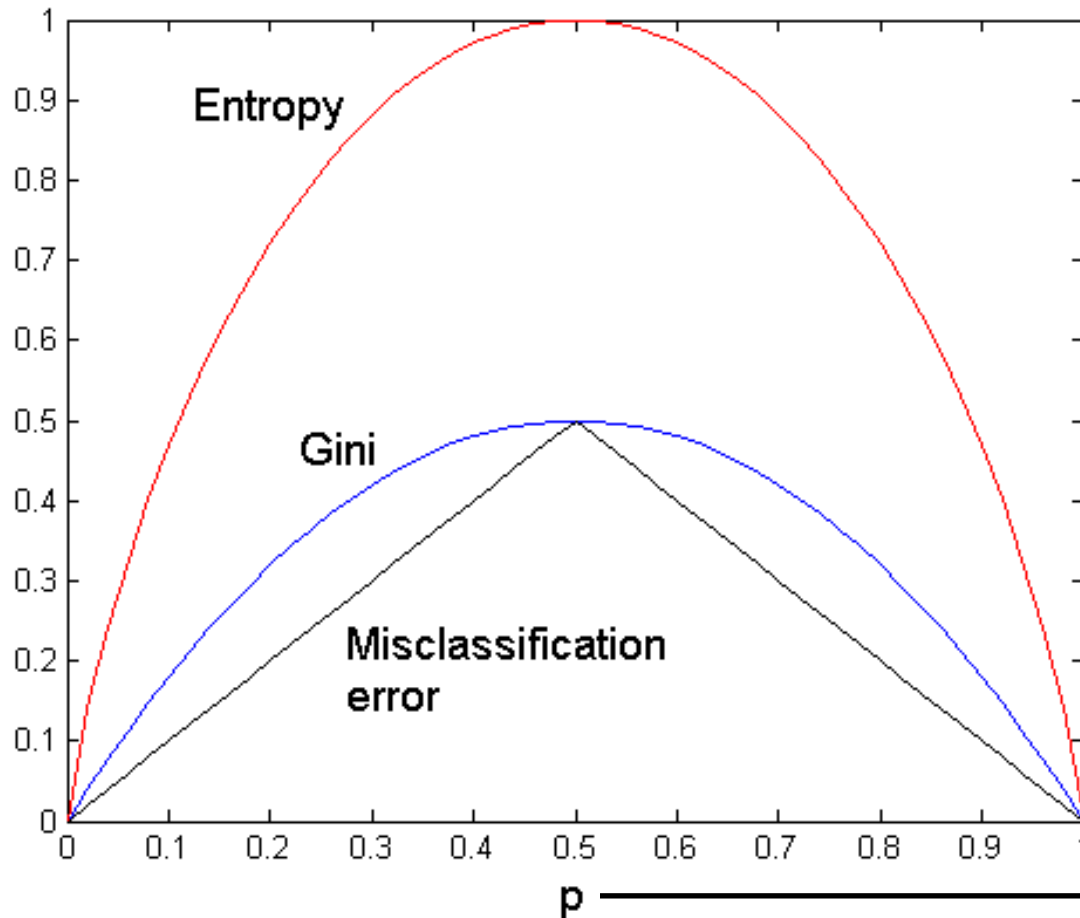
$$P(C1) = 2/6 \quad P(C2) = 4/6$$

$$Error = 1 - \max(2/6, 4/6) = 1 - 4/6 = 1/3$$

Uzayı bölme

Ölçütler arası karşılaştırma

İki sınıflı problem için



Bir sınıftan
gözlem oranı

Uzayı bölme

Bölmeyi ne zaman durdurmalıyız?

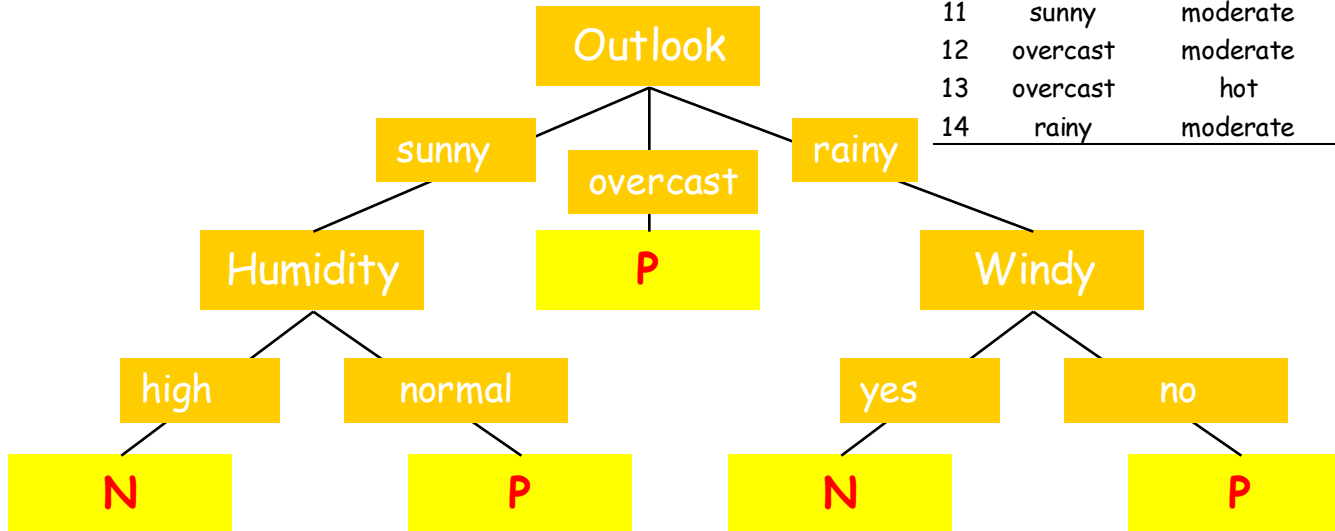
#	Attribute				Class Play
	Outlook	Temperature	Humidity	Windy	
1	sunny	hot	high	no	N
2	sunny	hot	high	yes	N
3	overcast	hot	high	no	P
4	rainy	moderate	high	no	P
5	rainy	cold	normal	no	P
6	rainy	cold	normal	yes	N
7	overcast	cold	normal	yes	P
8	sunny	moderate	high	no	N
9	sunny	cold	normal	no	P
10	rainy	moderate	normal	no	P
11	sunny	moderate	normal	yes	P
12	overcast	moderate	high	yes	P
13	overcast	hot	normal	no	P
14	rainy	moderate	high	yes	N

Uzayı bölme

Bölmeyi ne zaman durdurmalıyız?

Basit bir ağaç

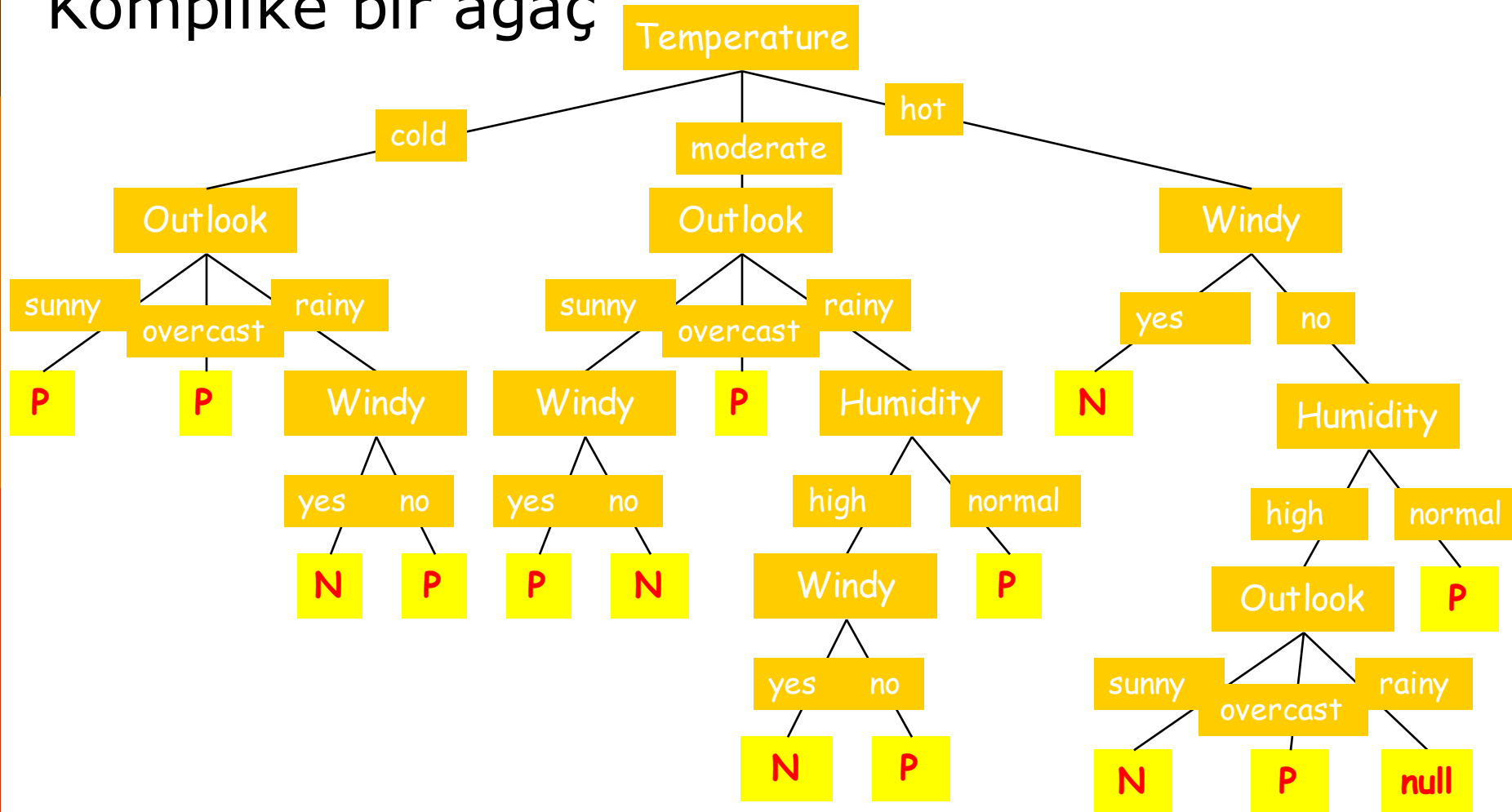
#	Attribute				Class Play
	Outlook	Temperature	Humidity	Windy	
1	sunny	hot	high	no	N
2	sunny	hot	high	yes	N
3	overcast	hot	high	no	P
4	rainy	moderate	high	no	P
5	rainy	cold	normal	no	P
6	rainy	cold	normal	yes	N
7	overcast	cold	normal	yes	P
8	sunny	moderate	high	no	N
9	sunny	cold	normal	no	P
10	rainy	moderate	normal	no	P
11	sunny	moderate	normal	yes	P
12	overcast	moderate	high	yes	P
13	overcast	hot	normal	no	P
14	rainy	moderate	high	yes	N



Uzayı bölme

Bölmeyi ne zaman durdurmalıyız?

Komplike bir ağaç



Uzayı bölme

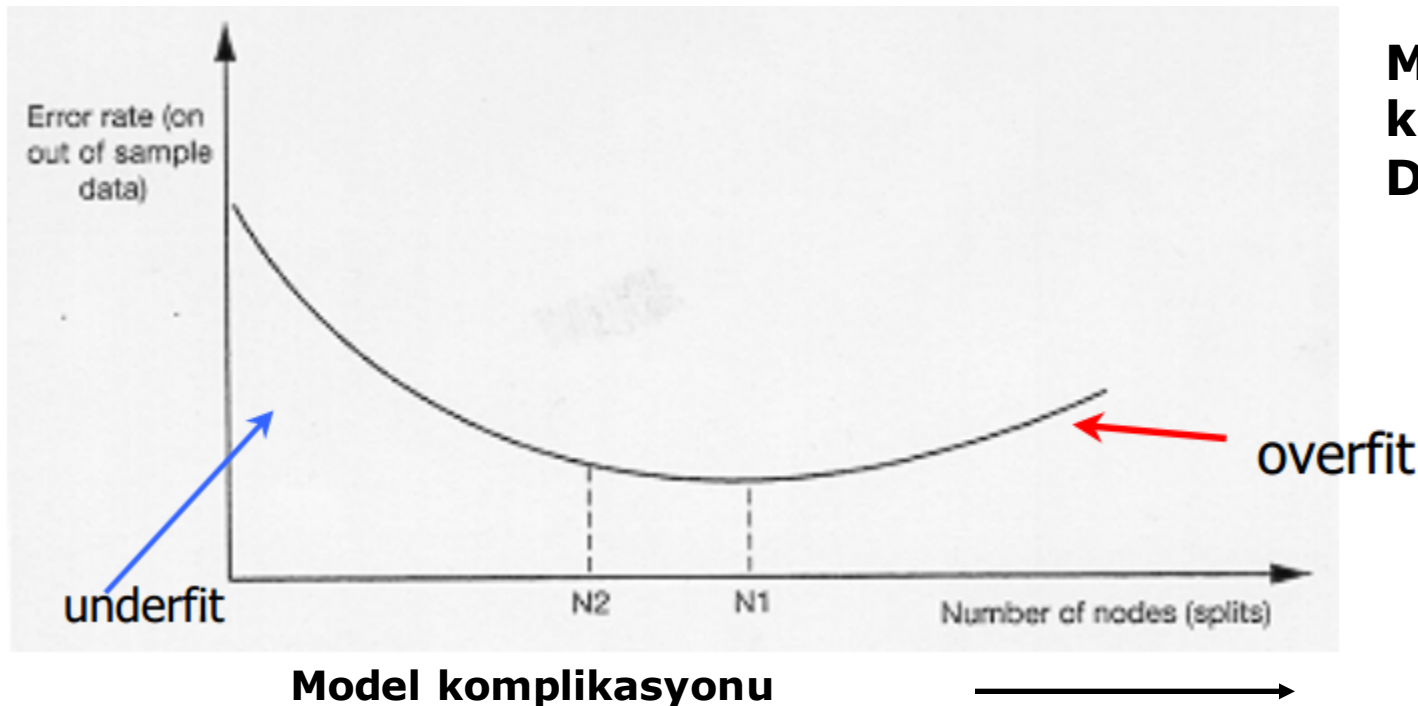
Bölmeyi ne zaman durdurmalıyız?

- Bölüm yapmayı aşağıdaki koşullar sağlandığında bırakabiliriz
 - Belli bir saflık elde edildiğinde
 - Tüm gözlemler aynı sınıftaysa (sınıflandırma ağacı)
 - Tüm gözlemler benzer değerlere sahipse (regresyon ağacı)
 - Belli bir derinliğe ulaşıldığında
 - Belli bir düğüm sayısına ulaşıldığında
 - Saflık iyileştirilemiyorsa

Sınıflandırma ile ilgili önemli meseleler

Underfitting (yetersiz uyum) ve Overfitting (aşırı uyma)

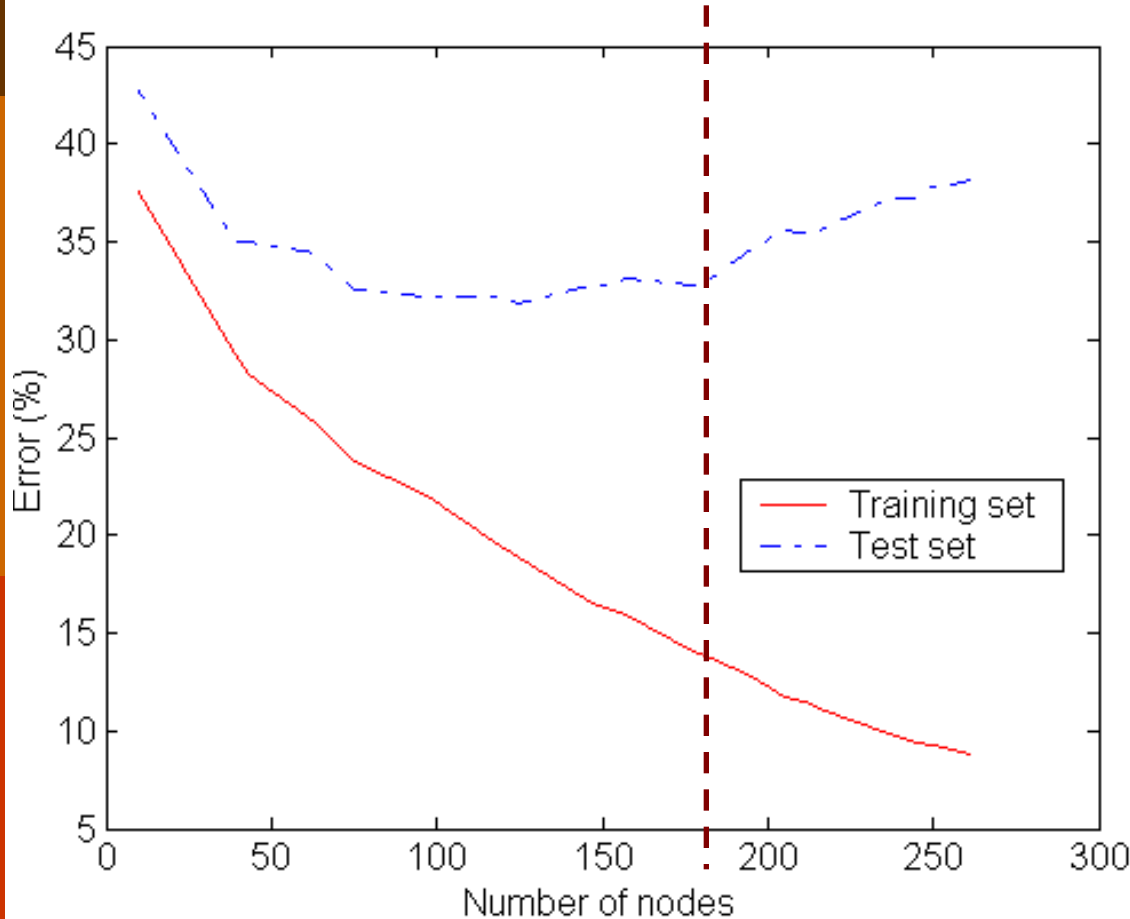
- Test verisindeki performans nasıl olacak?



**Model
komplikasyonu
Düğüm sayısı**

Sınıflandırma ile ilgili önemli meseleler

Underfitting (yetersiz uyum) ve Overfitting (aşırı uyma)

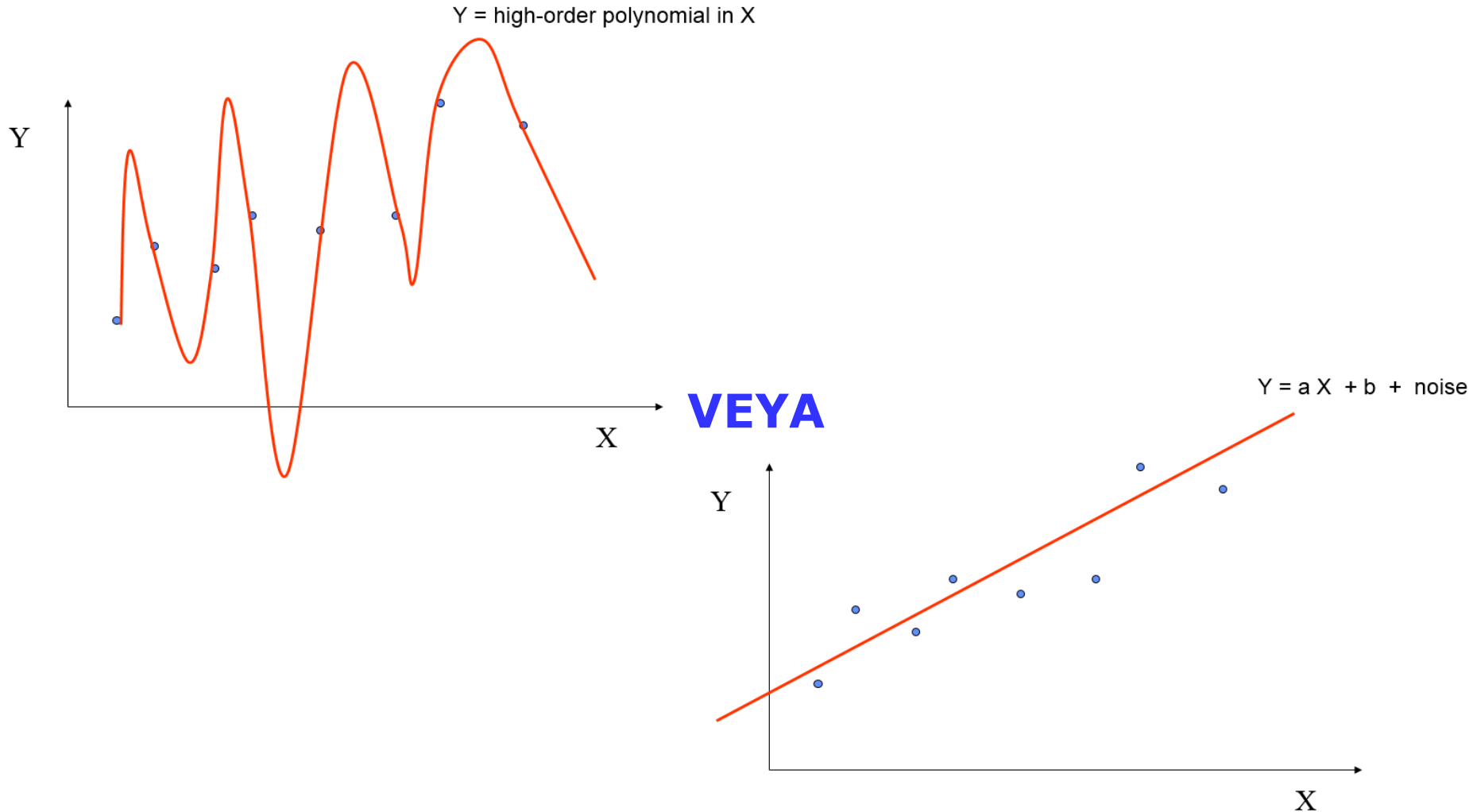


Overfitting: öğrenme verisini o kadar detaylı öğrenir ki, test verilerini tahmin etmede sorun yaşar

Underfitting: model o kadar basit kalır ki, test verisini tahmin etmek zorlaşır

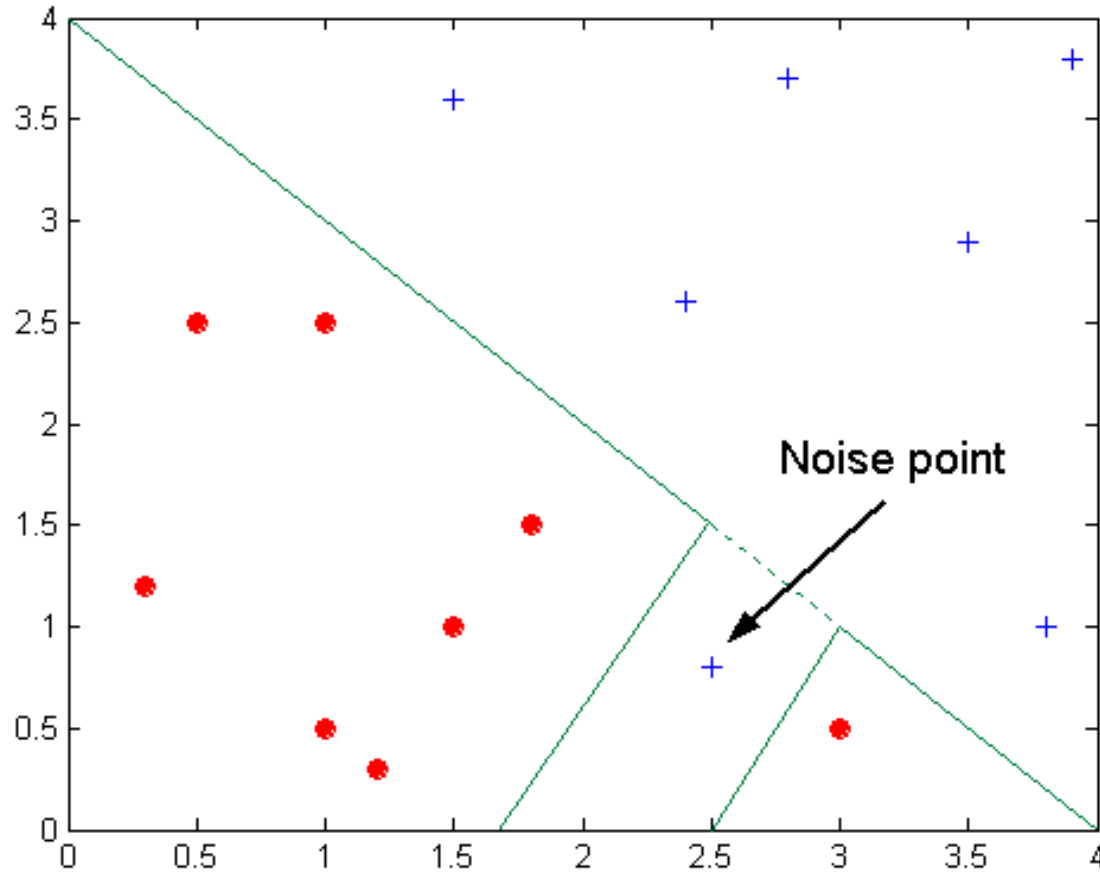
Sınıflandırma ile ilgili önemli meseleler

Underfitting (yetersiz uyum) ve Overfitting (aşırı uyma)



Sınıflandırma ile ilgili önemli meseleler

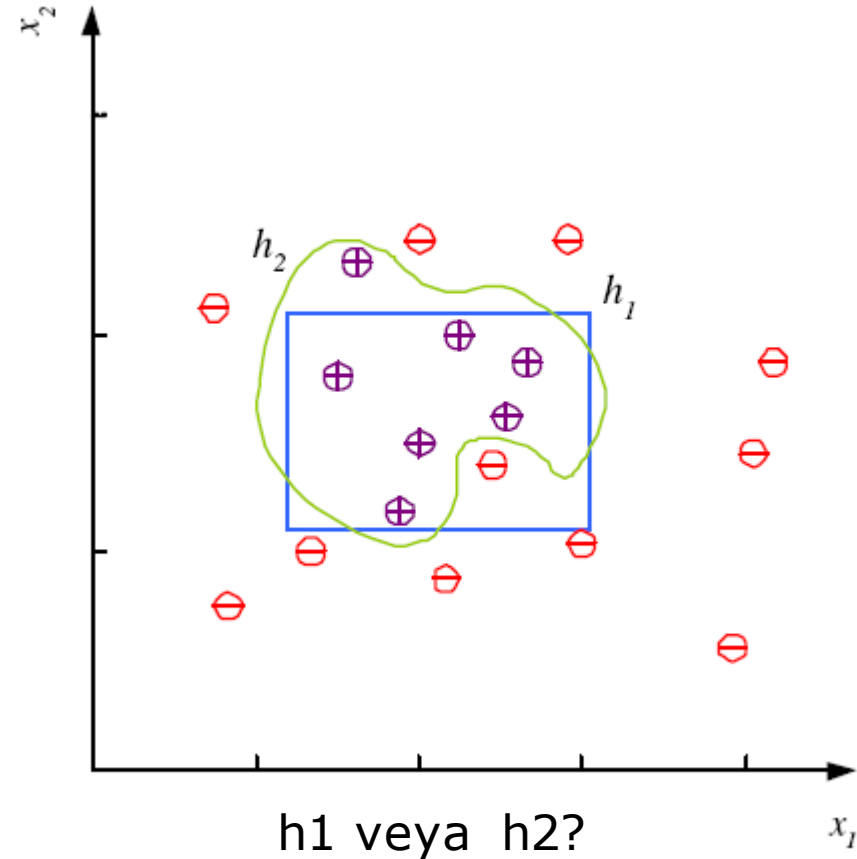
Overfitting (aşırı uyma) sebebiyle gürültü öğrenme



Karar verme sınırı gürültü sebebiyle bozulur

Tavsiyeler

- Farklı modelleri karşılaştırırken
 - Kullanımı kolay olanları
 - Model öğrenimi kolay
 - Gürültüye duyarlı olmayan
 - Açıklanabilir
 - Genelleştirmesi kolay (Occam's razor)



Ağaçların avantajları

- Yorumlanması ve anlaşılması kolaydır.
- Minimal ön işleme gerektirir. Diğer yöntemler:
 - Normalizasyon, kayıp değer tahmini, kategorik değerlerin sayısal değerlere çevrilmesi vb. gerektirir.
- Ağaç öğrenmesi ve ağaçtan tahmin yapması hızlıdır
- Sürekli, kategorik ve ordinal değişkenleri yapısını bozmadan ele alır.
- Veri dağılımı ile ilgili varsayımlardan etkilenmez
- Kendi içinde değişken seçimi yapar
 - Alakasız veya gürültülü veriden etkilenmesi minimaldir

Ağaçların dezavantajları

- Aşırı öğrenmeye meyillidir.
- Verideki ufak değişiklikler bambaşka bir ağaca sebep olabilir.
- Bazı tür karar sınırlarını bulmakta zorlanabilir
 - XOR
- Eğer belli bir sınıf çoğunluk ise, o sınıfı öğrenmeye meyillidir.
 - Öğrenimden önce sınıflar arası denge sağlanması gerekir
- Sürekli değişkenler üzerinden kural bulmaya meyillidir