



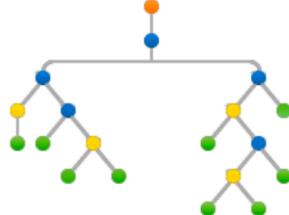
Karar Ağacıları

Hesaplamalı Matematik I - Özgür Martin

Karar Ağaçları (Decision Trees)

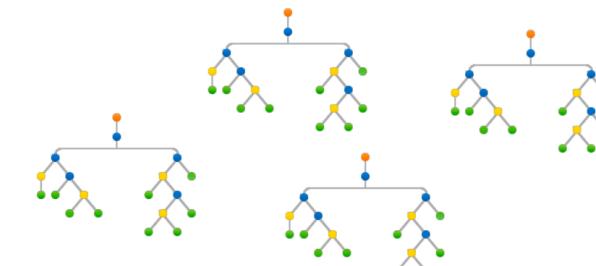
Amaç: Değişkenler uzayını basit alt bölgelere ayırmak

Bağlantım Ağaçları
Sınıflandırma Ağaçları

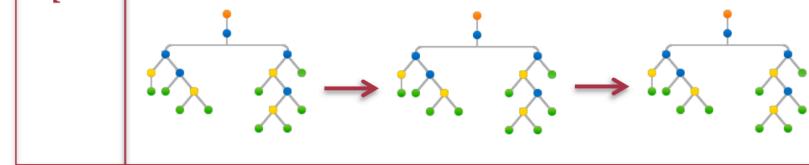


Topluluk Yöntemleri

Torbalama (Bagging) ve
Rasgele Ormanlar (Random Forests)



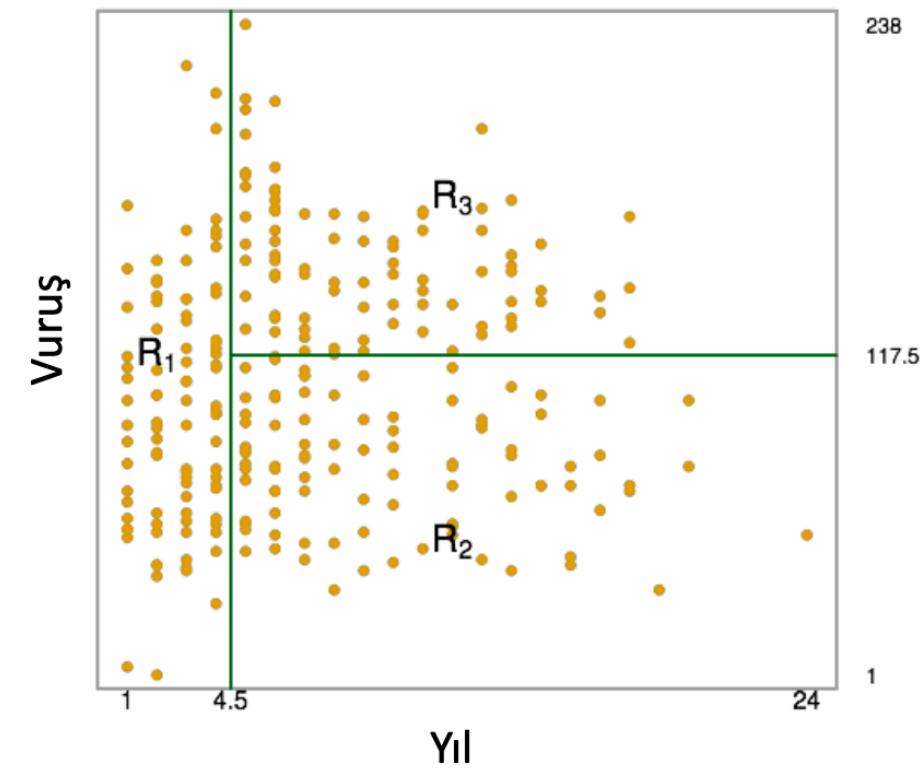
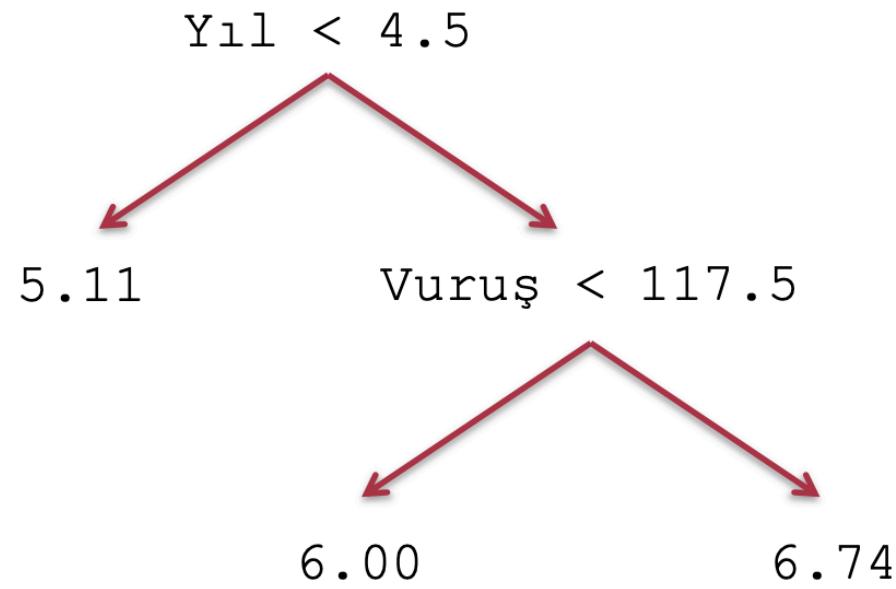
Takviye (Boosting)



- İlker Birbil, IMO2020 ders notları. ([www](#))
- “An Introduction to Statistical Learning – with Applications in R,”
G. James, D. Witten, T. Hastie, R. Tibshirani. 7th Ed., Springer, New York, 2013. ([www](#))

Örnek Ağaç

Bir beyzbol oyuncusunun maaşının yaptığı vuruş ve oynadığı toplam yıla göre tahmin edilmesi



Bağlanım Ağaçları

$$X_1, X_2, \dots, X_p \xrightarrow{\text{Çalışmayan bölgeler}} R_1, R_2, \dots, R_J$$

Tahmin: R_j bölgesindeki eğitim verisinin çıktı değerlerinin (y_i) ortalaması

$$R_1, R_2, \dots, R_J \ ?$$

Amaç: KKT değerinin en küçük olduğu bölgelerin bulunması

$$\sum_{j=1}^J \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$$

\hat{y}_{R_j} : R_j bölgesindeki çıktıların ortalaması

Bağlanım Ağaçları

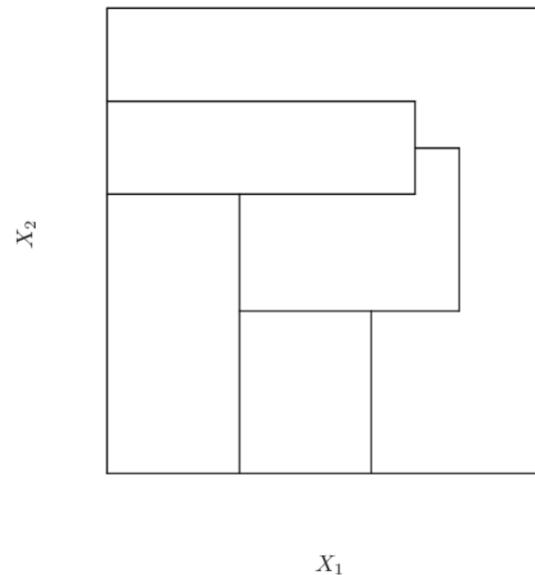
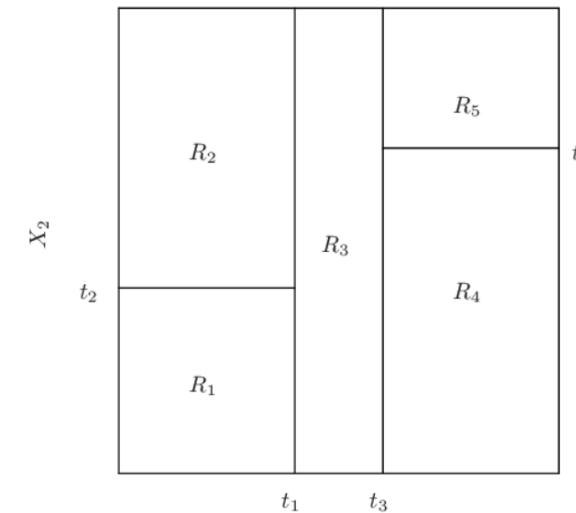
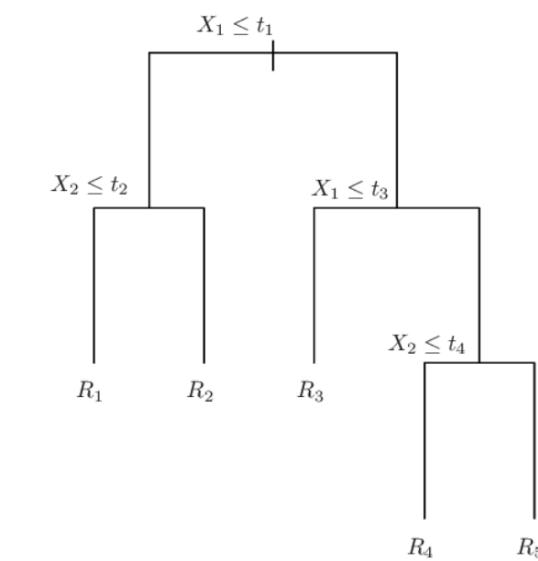
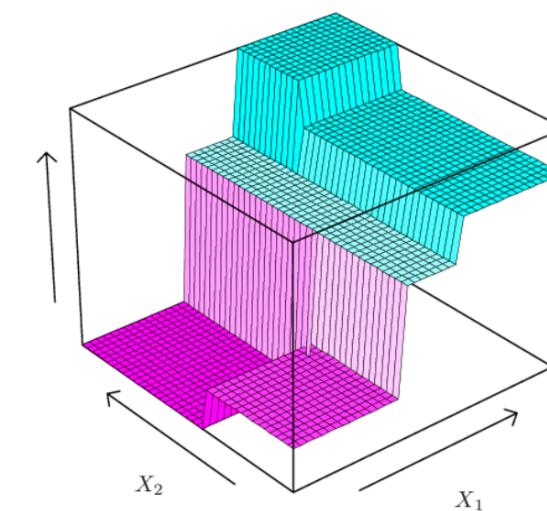
Özyinelemeli İkili Ayırma (Recursive Binary Splitting)

$$R_1(j, s) = \{X | X_j < s\} \quad R_2(j, s) = \{X | X_j \geq s\}$$

Aşağıdaki ifadeyi en küçükleyen j ve s değerlerini bul

$$\sum_{i:x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2$$

Terminal düğümde çok az veri noktası kalınca **dur!**


 X_1

 X_1


Bağlanım Ağaçları

Ağaç Budama (Tree Pruning)

Amaç: Ağacın tamamı ya da büyük kısmı oluşunca ortaya çıkan aşırı öğrenmenin önüne geçmek (düşük *test* hatası)

$$\sum_{m=1}^{|T|} \sum_{i:x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha |T|$$

$T \subset T_0$: alt ağaç (subtree)

$|T|$: T ağacındaki terminal düğüm sayısı

R_m : $m.$ terminal düğüme karşılık
gelen bölge

α : sabit parametre



α parametresini bulmak için k -katlı çapraz geçerlilik sınaması yapılır

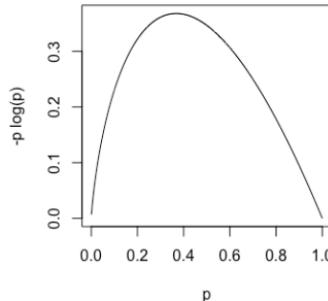
Sınıflandırma Ağaçları

Bağlanım ağaçlarına çok benzer şekilde ilerlenir ancak alt bölgenin *saflik* derecesine bağlı bir hata ölçüsü kullanılır

\hat{p}_{mk} : m . bölgedeki eğitim verisindeki k . sınıfın olanların oranı

Sınıflandırma Hata Oranı

$$E = 1 - \max_k \{\hat{p}_{mk}\}$$

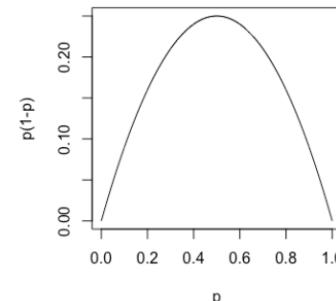


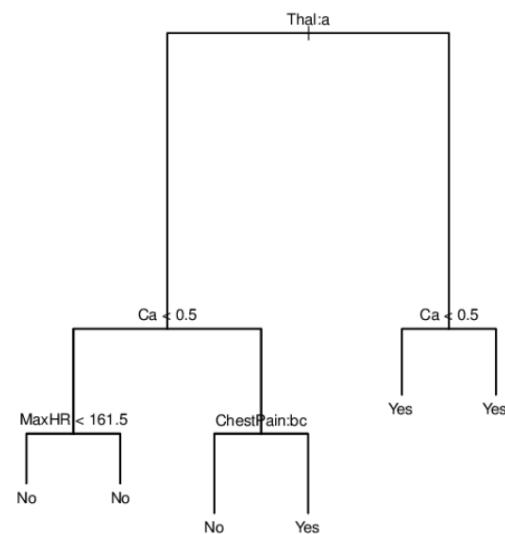
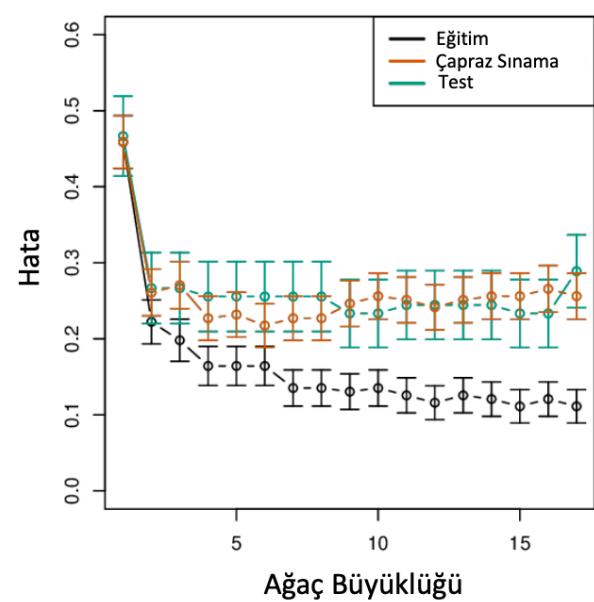
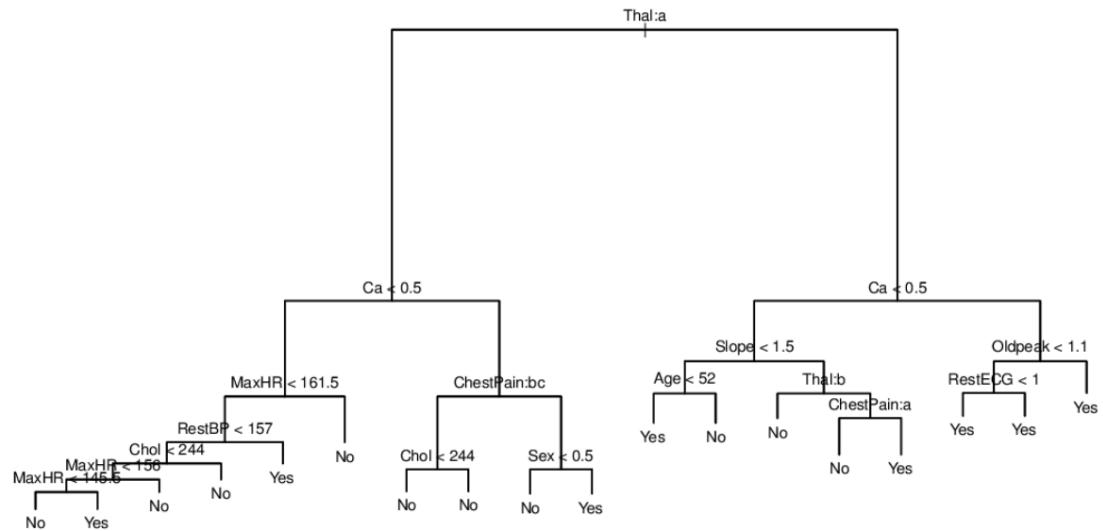
Entropi

$$D = - \sum_{k=1}^K \hat{p}_{mk} \log \hat{p}_{mk}$$

Gini İndeksi

$$G = \sum_{k=1}^K \hat{p}_{mk} (1 - \hat{p}_{mk})$$





Öğrenmeden Beklentimiz: Karar Ağaçları

Kestirim
(Prediction)

$$x_0 = \begin{bmatrix} \circ \\ \vdots \\ \circ \end{bmatrix} \longrightarrow y_0 = ?$$

Ne?

EVET

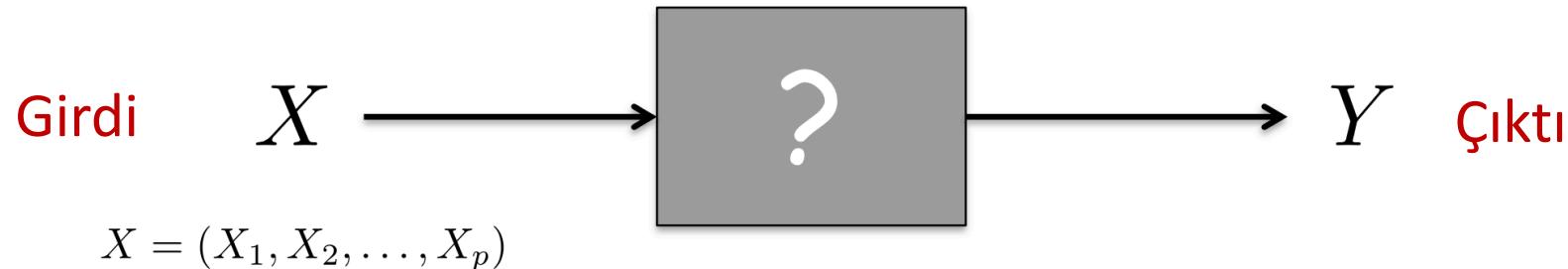
Çıkarım
(Inference)

$$x_0 = \begin{bmatrix} \bullet \\ \vdots \\ \bullet \end{bmatrix} \xrightarrow{\text{?}} y_0$$

Nasıl?

EVET

Öğrenme Problemi

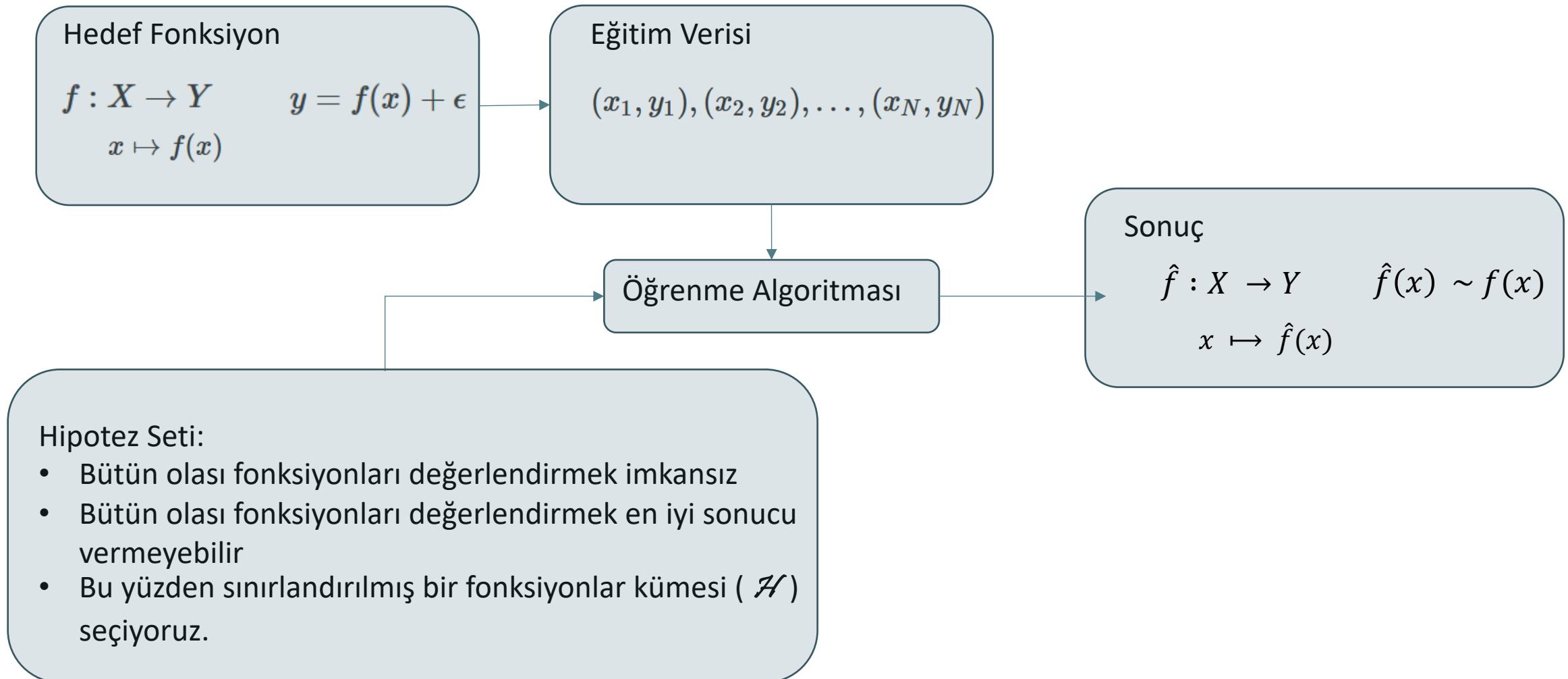


Bilinmeyen Fonksiyon

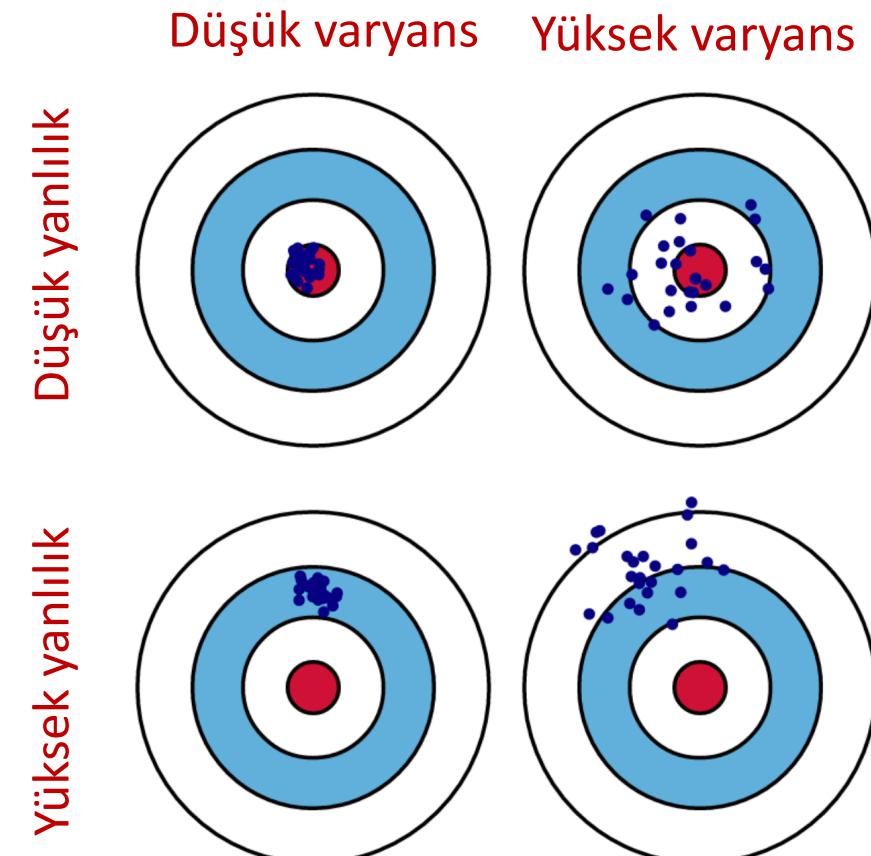
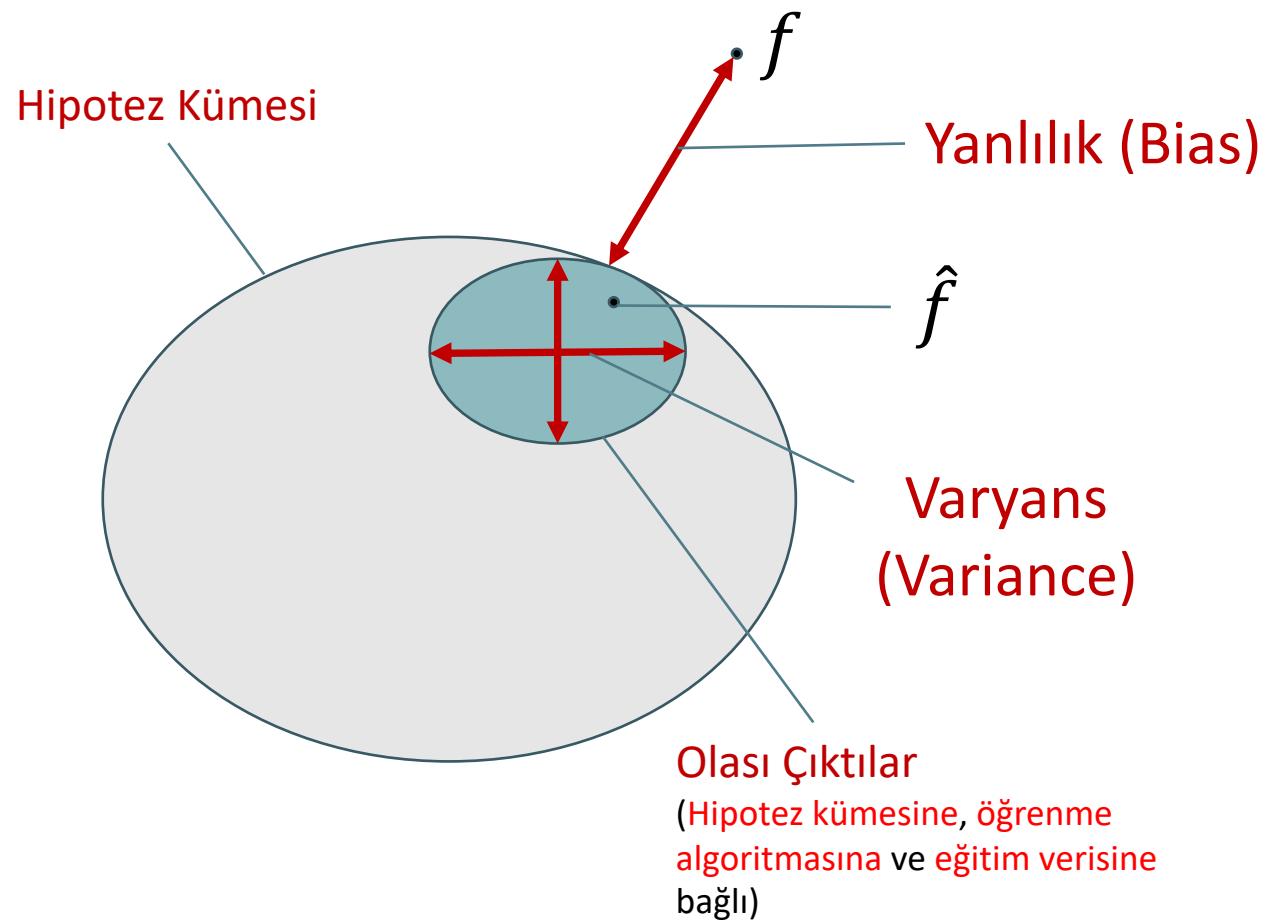
$$Y = f(X) + \epsilon \xrightarrow{\text{Yaklaşık?}} \hat{Y} = \hat{f}(X)$$

↓
Rastgele Hata Terimi
(Girdiden bağımsız)
↑

Öğrenme Problemi



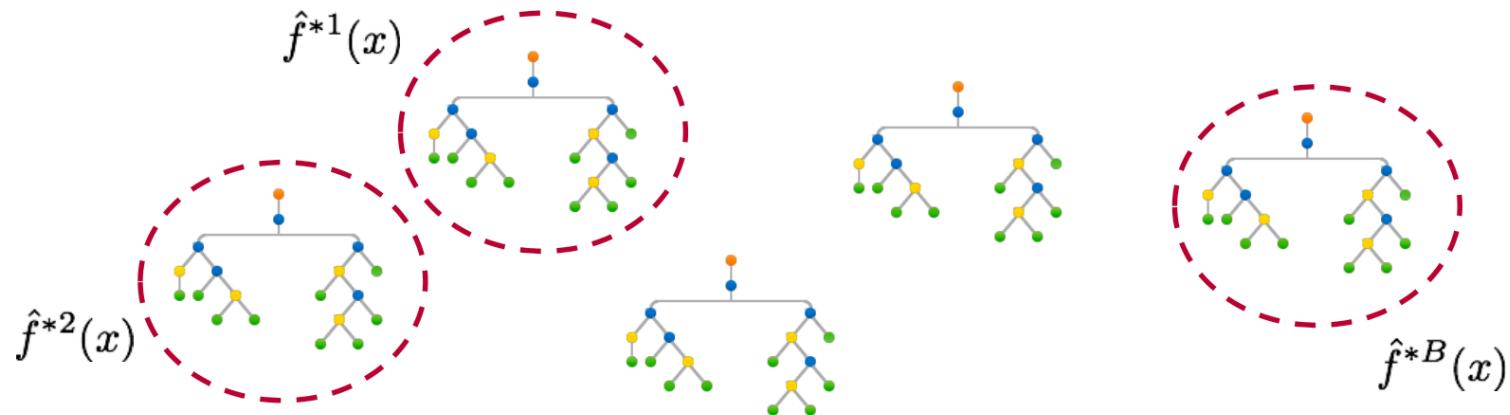
Öğrenme Problemi



Topluluk Yöntemleri

Torbalama

Amaç: Varyansı düşürmek için zorlama tekniğini kullanarak birkaç tane büyük ağaç oluşturular ve onların tahminlerinin ortalaması (bağlanım) ya da çoğunlukta olan sınıf (sınıflandırma) hesaplanır.



Bağlanım

B : farklı eğitim kümesi sayısı

$\hat{f}^{*b}(x)$: b . eğitim kümesi ile
elde edilen tahmin

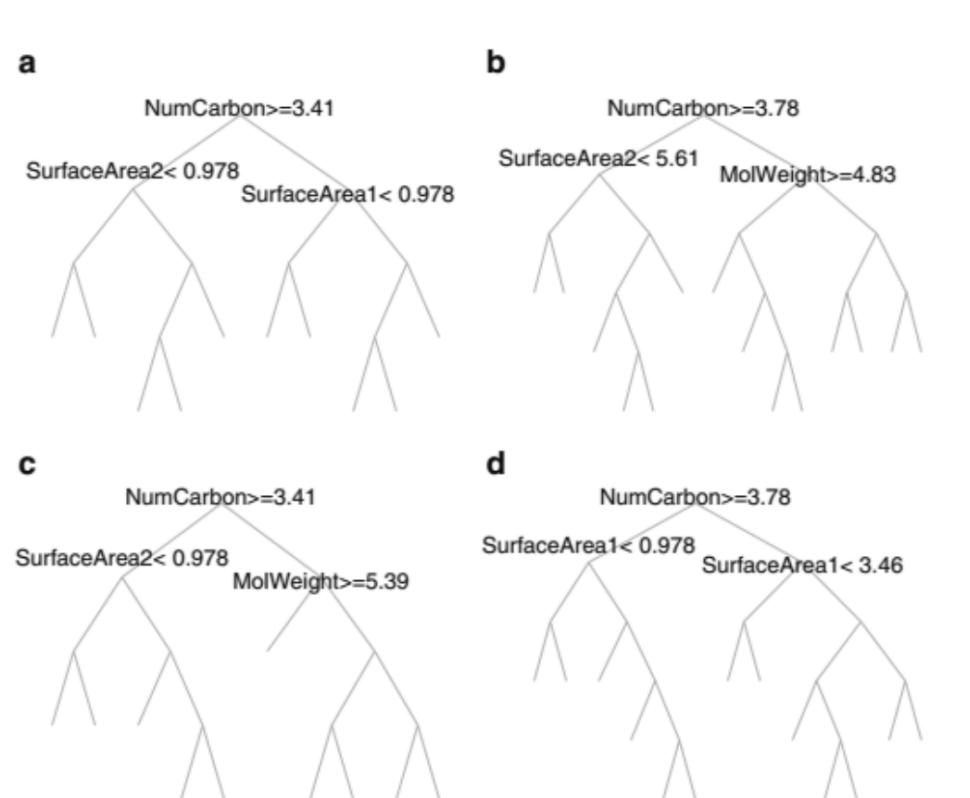
$$\hat{f}_{\text{bag}}(x) = \frac{1}{B} \sum_{b=1}^B \hat{f}^{*b}(x)$$

Topluluk Yöntemleri

Rasgele Ormanlar

Amaç: Ağaçlar arasındaki korelasyonu azaltmak için dalları ayırırken tüm değişkenler yerine sadece rassal sayıda değişkeni kullanmak

Neden?

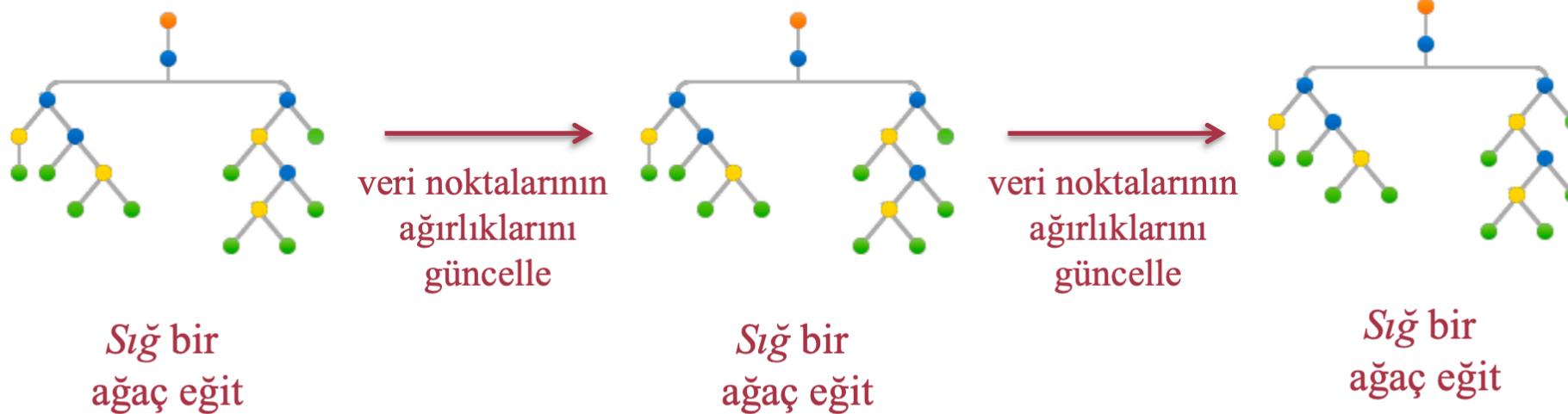


**Applied Predictive Modeling*, M. Kuhn, K. Johnson., Springer, 2013, sf. 195.

Topluluk Yöntemleri

Takviye

Fikir: Fazla etkin olmayan sınıflandırıcıları, ağırlıklı veri örnekleme tekniğini kullanarak bir araya getirerek daha etkin bir sınıflandırıcı elde etmek



Güncelleme Kuralı: Yanlış sınıflandırılan veri noktalarının ağırlıklarını artır

Öğrenmeden Beklentimiz: Topluluk Yöntemleri

Kestirim
(Prediction)

$$x_0 = \begin{bmatrix} \circ \\ \vdots \\ \circ \end{bmatrix} \longrightarrow y_0 = ?$$

Ne?

EVET

Çıkarım
(Inference)

$$x_0 = \begin{bmatrix} \bullet \\ \vdots \\ \bullet \end{bmatrix} \xrightarrow{?} y_0$$

Nasıl?

HAYIR

Ensemble Methods

AdaBoost – İkili Sınıflandırma {+1, -1}

Her bir veri satırı başlangıçta aynı ağırlığa sahip: ($1/n$)

for $k=1$ to K **do**

ağırlıklı verileri kullanarak d dallı bir ağaç eğit ve and yanlış sınıflandırma hatasını hesapla (ϵ_k)

$$\text{Ağırlık değerini hesapla } \ln \frac{1 - \epsilon_k}{\epsilon_k}$$

Ağırlıklı verileri güncelle – yanlış sınıflandırılan verilere daha fazla ağırlık ver

end

Her bir veri için k . aşamadaki değer ile k . model tahminini çarparak takviyeli sınıflandırıcının tahminlerini hesapla ve bu miktarları k 'ye ekle. Eğer toplam pozitif ise veriyi +1 olarak sınıflandır, değilse -1.

Kitaptaki Algoritma 8.2 **bağlanım ağaçları** için takviye örneği veriyor.

Topluluk Yöntemleri

AdaBoost – İkili Sınıflandırma {+1, -1}

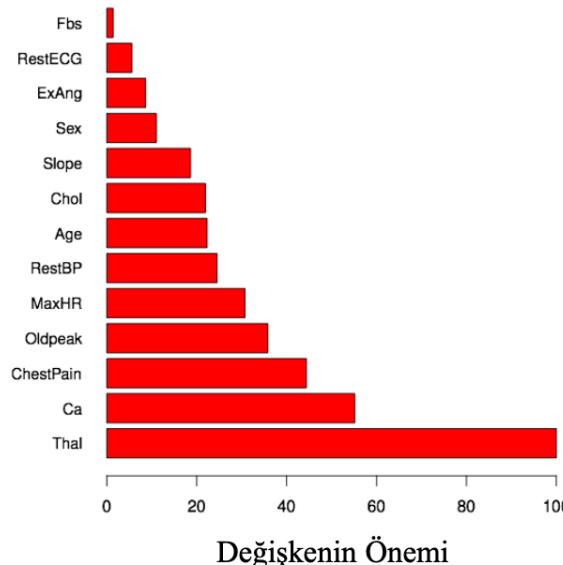
Algoritma 1: AdaBoost

- 1 Başlangıç ağırlıklarını belirle: $w_i^1 = \frac{1}{n}, \quad i = 1, \dots, n$
 - 2 **for** $k = 1, \dots, K$ **do**
 - 3 $y_k(x)$ sınıflandırıcısını şu ağırlıklı sınıflandırma hata fonksiyonu ile eğit:
$$\sum_{i=1}^n w_i^k I(y_k(x_i) \neq y_i)$$
Yanlış sınıflandırılan veri noktalarının oransal ağırlığını hesapla:
$$\varepsilon^k = \frac{\sum_{i=1}^n w_i^k I(y_k(x_i) \neq y_i)}{\sum_{i=1}^n w_i^k}$$
 - 4 Güncelleme parametresini belirle: $\alpha_k = \ln \frac{1-\varepsilon_k}{\varepsilon_k}$
 - 5 Ağırlıkları güncelle: $w_i^{k+1} = w_i^k e^{\alpha_k I(y_k(x_i) \neq y_i)}, \quad i = 1, \dots, n$
 - 6 **Çıktı:** $\sum_{k=1}^K \alpha_k y_k(x)$ değerinin işaretini (+1 ya da -1)
-

[AdaBoost Özeti](#)

Topluluk Yöntemleri

- Takviye ve torbalama yöntemleri başka yöntemler ile de uygulanabilirler
- Torbalama yöntemi paralel uygulama için son derece uygun olmasına rağmen takviye yöntemi sıralı yapısı nedeniyle paralelleştirmeye uygun değildir
- Topluluk yöntemleri ile elde edilen modeli yorumlamak güçtür
- Değişken önemini (variable importance) gösteren grafikler kullanılabilir



Bir **değişkenin önemi** Gini indeksinde ya da entropide elde ettiği ortalama azaltmaya göre belirlenir. Daha sonra değişkenler bu önem sırasına göre oranlanarak sıralanır.