Makine Öğrenmesi

Karar ağaçları, regresyon ağaçları, ensemble learning

Orijinal not: Ders6.pdf

UYARI Bu belge, el yazısı ders notlarından Claude (Anthropic) yardımıyla dijitale aktarılmıştır. Yardımcı kaynak olarak hazırlanmıştır; hata veya eksik içerebilir. Lütfen orijinal notları ve ders kaynaklarını esas alın.

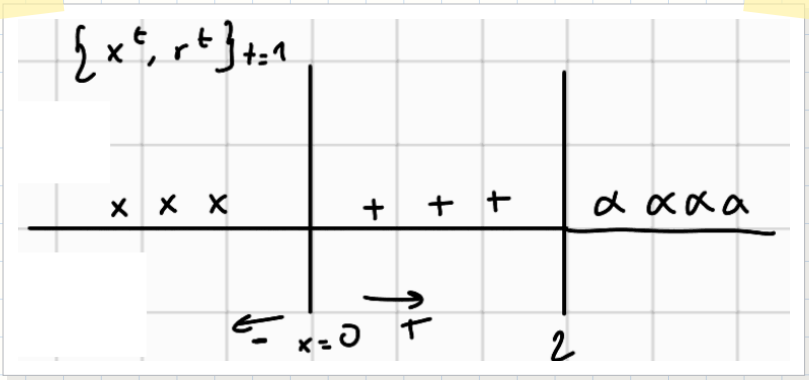
NOTICE *This document was transcribed from handwritten lecture notes with the help of Claude (Anthropic). It is meant as a supplementary resource and may contain errors or omissions. Always check against the original notes and official course material.*

orj. s. 1

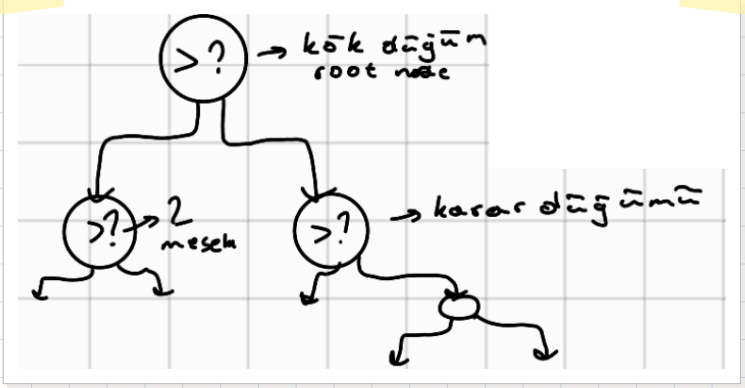
Karar ağaçları , Regresyon ağaçları

Decision tree

ifle yapmaya çalışalım



orijinal çizim, s. 1



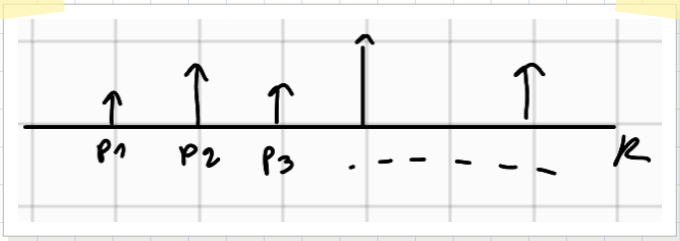
orijinal çizim, s. 1

Sınır değerleri nasıl belirleyeceğimizi bulalım

En sonda amaç homojen olarak verileri ayırmak, olabildiğince tek bir sınıf istiyorum

$$p_i = \frac{N_i}{N}$$

Impurity measures:



orijinal çizim, s. 1

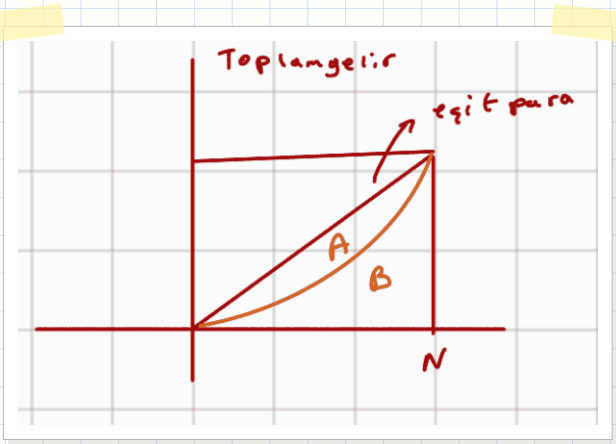
en saf durum tek bir olasılığın 1 kalanın 0 olması

Entropy:

$$I_{\text{entropy}} = -\sum_{i=1}^{K} p_i \log p_i$$

→ tek bir sınıf varsa minimum
çok sınıf varsa maksimum

Gini Sabiti:



orijinal çizim, s. 1

$$\frac{A}{A+B}$$

orj. s. 2

Bizim uygulamada

$$I_{\text{Gini}} = 1 - \sum_{i=1}^{K} p_i^2$$

0 ile $1 - \frac{1}{N^2}$ arasında

dallanırken max entropy nerdeyse ordan devam et
CART

$$\sum_{j=1}^{N} w_j I_j \qquad w_j = \frac{N_j}{N}$$

$$\varepsilon = \varepsilon_k - \varepsilon_{k-1}$$

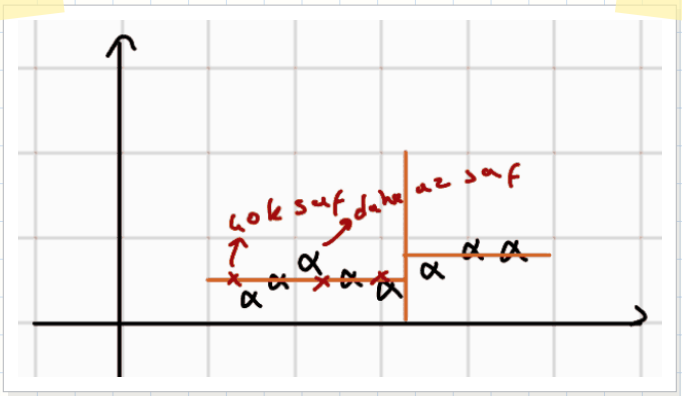
while $\varepsilon >$ threshold δ



orijinal çizim, s. 2

$$\mathcal{L}(\theta \mid X) = \sum_{j=1}^{N} w_j I_j + \underbrace{M \cdot \lambda}_{\substack{\text{yaprak sayısı} \quad \text{kat sayısı} \\ \text{overfittingi engeller}}} \rightarrow \lambda \cdot f(x)$$

Regresyon ağaçları

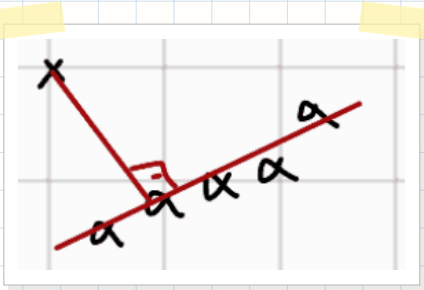


orijinal çizim, s. 2

Metrik olarak varyans ?kullanabiliriz, veya L_1 kullanılabilir.
(Goodness of split)

En büyük sıkıntısı overfitting dalın en alt yaprağına indiğimiz için

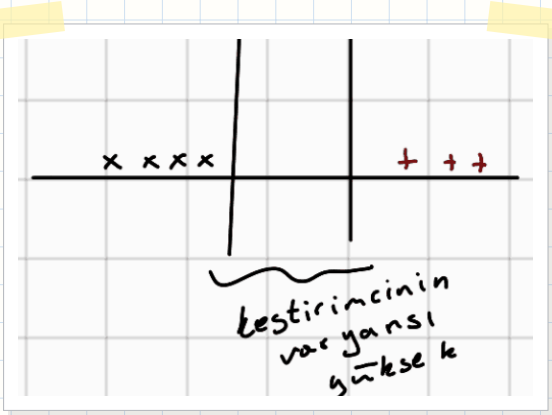
orj. s. 3 L_1 norm aykırı değerlere hassasiyeti az



orijinal çizim, s. 3

Ensemble Learning

Bagging



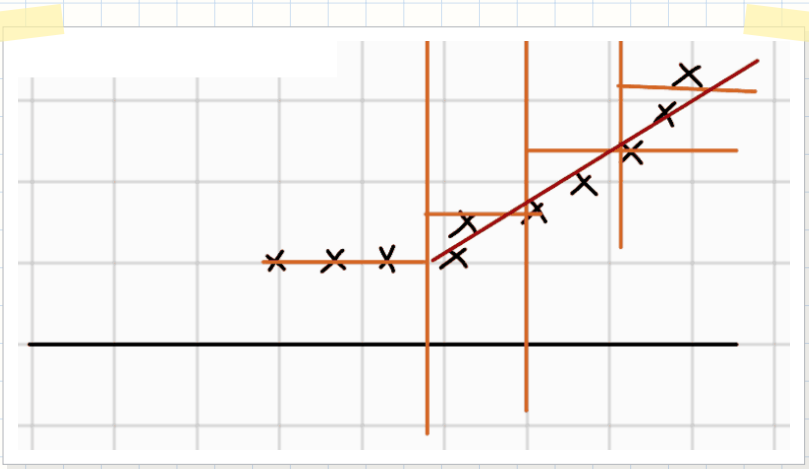
orijinal çizim, s. 3

100 tane karar ağacı alırım, yeni bir örnek geldiği zaman oylama yapıyorum böylece kestirimcinin varyansı küçülür.

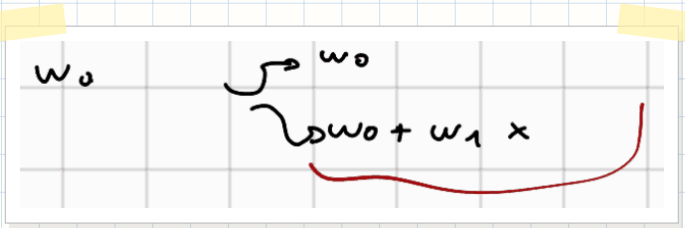
Random forest

Boosting:

Verinin özel bir bölgesine outputta daha iyi bir sınıflandırıcı varsa bu şekilde yaklaşabilir



orijinal çizim, s. 3



orijinal çizim, s. 3

Adaboost

$$f_i(x),\ i \in \{0, \ldots, M\}$$

$$y = \sum_{i=1}^{M} w_i f_i(x)$$

verinin hangi bölgesinde hangi fonksiyon daha iyi

Aynı modelin parametrelerini değiştirip lineer kombinasyonlarını kullanıyoruz

orj. s. 4

$$\mathcal{L}(\theta \mid X) = \sum_{t=1}^{N} w_t \underbrace{(\hat{r}^t - r^t)^2}_{\text{risk gibi}}$$

haftaya feature selection ve dimension reduction işleyeceğiz