# 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

- Twoing Algoritması
- Gini Algoritması

## Twoing Algoritması

- **Twoing algoritmasında eğitim kümesi her adımda iki parçaya ayrılarak bölümleme yapılır.**
- Aday bölünmelerin sağ ve sol kısımlarının her birisi için tekrar oranı alınır.
- Aday bölünmelerin sağ ve sol kısımlarındaki her bir nitelik değeri için sınıf değerlerinin her birisinin olma olasılığı hesaplanır.
- Her bölünme için **uygunluk değeri en yüksek olan alınır.**

$$\Phi(B \mid d) = 2\frac{|B_{sol}|}{|T|}\frac{|B_{sag}|}{|T|}\sum_{j=1}^{n} abs\left(\frac{|\text{Tsinif}_j|}{|B_{sol}|} - \frac{|\text{Tsinif}_j|}{|B_{sag}|}\right)$$

- Burada, $T$ eğitim kümesindeki kayıt sayısını, $B$ aday bölünmeyi, $d$ düğümü, $Tsinif_j$ ise $j$.sınıf değerini gösterir.

3

## Twoing Algoritması

**Örnek**

| PERSONEL | MAAŞ | DENEYİM | GÖREV | MEMNUN |
|---|---|---|---|---|
| 1 | NORMAL | ORTA | UZMAN | EVET |
| 2 | YÜKSEK | YOK | UZMAN | EVET |
| 3 | DÜŞÜK | YOK | YÖNETİCİ | EVET |
| 4 | YÜKSEK | ORTA | YÖNETİCİ | EVET |
| 5 | DÜŞÜK | ORTA | YÖNETİCİ | EVET |
| 6 | YÜKSEK | İYİ | YÖNETİCİ | EVET |
| 7 | DÜŞÜK | İYİ | YÖNETİCİ | EVET |
| 8 | YÜKSEK | ORTA | UZMAN | HAYIR |
| 9 | DÜŞÜK | ORTA | UZMAN | HAYIR |
| 10 | YÜKSEK | İYİ | UZMAN | HAYIR |
| 11 | DÜŞÜK | İYİ | UZMAN | HAYIR |

4

## Twoing Algoritması

**Örnek – devam**

- Aday bölünmeler aşağıdaki gibidir.

| BÖLÜNME | SOL | SAĞ |
|---|---|---|
| 1 | MAAŞ = NORMAL | MAAŞ = {DÜŞÜK, YÜKSEK} |
| 2 | MAAŞ = YÜKSEK | MAAŞ = {DÜŞÜK, NORMAL} |
| 3 | MAAŞ = DÜŞÜK | MAAŞ = {NORMAL, YÜKSEK} |
| 4 | DENEYİM = YOK | DENEYİM = {ORTA, İYİ} |
| 5 | DENEYİM = ORTA | DENEYİM = {YOK, İYİ} |
| 6 | DENEYİM = İYİ | DENEYİM = {YOK, ORTA} |
| 7 | GÖREV = UZMAN | GÖREV = YÖNETİCİ |
| 8 | GÖREV = YÖNETİCİ | GÖRE = UZMAN |

5

## Twoing Algoritması

**Örnek – devam**

- MAAŞ = {NORMAL} için

| PERSONEL | MAAŞ | DENEYİM | GÖREV | MEMNUN |
|---|---|---|---|---|
| 1 | NORMAL | ORTA | UZMAN | EVET |
| 2 | YÜKSEK | YOK | UZMAN | EVET |
| 3 | DÜŞÜK | YOK | YÖNETİCİ | EVET |
| 4 | YÜKSEK | ORTA | YÖNETİCİ | EVET |
| 5 | DÜŞÜK | ORTA | YÖNETİCİ | EVET |
| 6 | YÜKSEK | İYİ | YÖNETİCİ | EVET |
| 7 | DÜŞÜK | İYİ | YÖNETİCİ | EVET |
| 8 | YÜKSEK | ORTA | UZMAN | HAYIR |
| 9 | DÜŞÜK | ORTA | UZMAN | HAYIR |
| 10 | YÜKSEK | İYİ | UZMAN | HAYIR |
| 11 | DÜŞÜK | İYİ | UZMAN | HAYIR |

$$P_{sol} = \frac{|B_{sol}|}{|T|} = \frac{1}{11} = 0{,}09$$

$$P_{(EVET/t_{sol})} = \frac{|\text{Tsinif}_{EVET}|}{|B_{sol}|} = \frac{1}{1} = 1$$

$$P_{(HAYIR/t_{sol})} = \frac{|\text{Tsinif}_{HAYIR}|}{|B_{sol}|} = \frac{0}{1} = 0$$

| BÖLÜNME | $\|B_{sol}\|$ | $P_{Sol}$ | $\|sinif_{EVET}\|$ | $\|sinif_{HAYIR}\|$ | $P(EVET\|t_{Sol})$ | $P(HAYIR\|t_{Sol})$ |
|---|---|---|---|---|---|---|
| 1 | 1 | 0,09 | 1 | 0 | 1 | 0 |
| 2 | 5 | 0,45 | 3 | 2 | 0,6 | 0,4 |
| 3 | 5 | 0,45 | 3 | 2 | 0,6 | 0,4 |
| 4 | 2 | 0,18 | 2 | 0 | 1 | 0 |
| 5 | 5 | 0,45 | 3 | 2 | 0,6 | 0,4 |
| 6 | 4 | 0,36 | 2 | 2 | 0,5 | 0,5 |
| 7 | 6 | 0,55 | 2 | 4 | 0,33 | 0,67 |
| 8 | 5 | 0,45 | 5 | 0 | 1 | 0 |

6

## Twoing Algoritması

### Örnek – devam

- MAAŞ = {DÜŞÜK, YÜKSEK}

| PERSONEL | MAAŞ | DENEYİM | GÖREV | MEMNUN |
|---|---|---|---|---|
| 1 | NORMAL | ORTA | UZMAN | EVET |
| 2 | YÜKSEK | YOK | UZMAN | EVET |
| 3 | DÜŞÜK | YOK | YÖNETİCİ | EVET |
| 4 | YÜKSEK | ORTA | YÖNETİCİ | EVET |
| 5 | DÜŞÜK | ORTA | YÖNETİCİ | EVET |
| 6 | YÜKSEK | İYİ | YÖNETİCİ | EVET |
| 7 | DÜŞÜK | İYİ | YÖNETİCİ | EVET |
| 8 | YÜKSEK | ORTA | UZMAN | HAYIR |
| 9 | DÜŞÜK | ORTA | UZMAN | HAYIR |
| 10 | YÜKSEK | İYİ | UZMAN | HAYIR |
| 11 | DÜŞÜK | İYİ | UZMAN | HAYIR |

$$P_{sag} = \frac{|B_{sag}|}{|T|} = \frac{10}{11} = 0{,}91$$

$$P_{(EVET/t_{sag})} = \frac{|\text{Tsinif}_{EVET}|}{|B_{sag}|} = \frac{6}{10} = 0{,}6$$

$$P_{(HAYIR/t_{sag})} = \frac{|\text{Tsinif}_{HAYIR}|}{|B_{sag}|} = \frac{4}{10} = 0{,}4$$

| BÖLÜNME | $\|B_{sag}\|$ | $P_{sag}$ | $\|sinif_{EVET}\|$ | $\|sinif_{HAYIR}\|$ | $P(EVET\|t_{Sag})$ | $P(HAYIR\|t_{Sag})$ |
|---|---|---|---|---|---|---|
| 1 | 10 | 0,91 | 6 | 4 | 0,6 | 0,4 |
| 2 | 6 | 0,55 | 4 | 2 | 0,67 | 0,33 |
| 3 | 6 | 0,55 | 4 | 2 | 0,67 | 0,33 |
| 4 | 9 | 0,82 | 5 | 4 | 0,56 | 0,44 |
| 5 | 6 | 0,55 | 4 | 2 | 0,67 | 0,33 |
| 6 | 7 | 0,64 | 5 | 2 | 0,71 | 0,29 |
| 7 | 5 | 0,45 | 5 | 0 | 1 | 0 |
| 8 | 6 | 0,55 | 2 | 4 | 0,33 | 0,67 |

## Twoing Algoritması

### Örnek – devam

| BÖLÜNME | SOL | SAĞ |
|---|---|---|
| 1 | MAAŞ = NORMAL | MAAŞ = {DÜŞÜK, YÜKSEK} |
| 2 | MAAŞ = YÜKSEK | MAAŞ = {DÜŞÜK, NORMAL} |
| 3 | MAAŞ = DÜŞÜK | MAAŞ = {NORMAL, YÜKSEK} |
| 4 | DENEYİM = YOK | DENEYİM = {ORTA, İYİ} |
| 5 | DENEYİM = ORTA | DENEYİM = {YOK, İYİ} |
| 6 | DENEYİM = İYİ | DENEYİM = {YOK, ORTA} |
| 7 | GÖREV = UZMAN | GÖREV = YÖNETİCİ |
| 8 | GÖREV = YÖNETİCİ | GÖRE = UZMAN |

- Uygunluk değeri (1. aday bölünme için)

$$\Phi(1|d) = 2\frac{|B_{sol}|}{|T|}\frac{|B_{sag}|}{|T|}\sum_{j=1}^{n} abs\left(\frac{|\text{Tsinif}_j|}{|B_{sol}|} - \frac{|\text{Tsinif}_j|}{|B_{sag}|}\right)$$

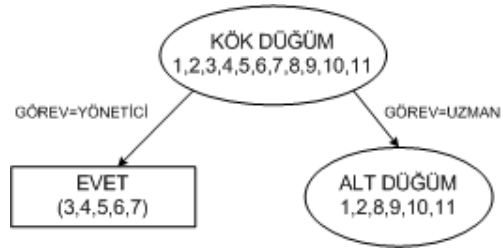
$$= 2(0{,}09)(0{,}91)[|1-0{,}6|+|0-0{,}4|] = 0{,}13$$

| BÖLÜNME | $P_{Sol}$ | $P_{Sağ}$ | $2P_{Sol}P_{Sağ}$ | $\Phi(B\|d)$ |
|---|---|---|---|---|
| 1 | 0,09 | 0,91 | 0,17 | 0,13 |
| 2 | 0,45 | 0,55 | 0,5 | 0,07 |
| 3 | 0,45 | 0,55 | 0,5 | 0,07 |
| 4 | 0,18 | 0,82 | 0,3 | 0,26 |
| 5 | 0,45 | 0,55 | 0,5 | 0,07 |
| 6 | 0,36 | 0,64 | 0,46 | 0,2 |
| 7 | 0,55 | 0,45 | 0,5 | 0,66 |
| 8 | 0,45 | 0,55 | 0,5 | 0,66 |

## Twoing Algoritması

**Örnek – devam**

- Aynı işlemler ALT DÜĞÜM için tekrarlanır.

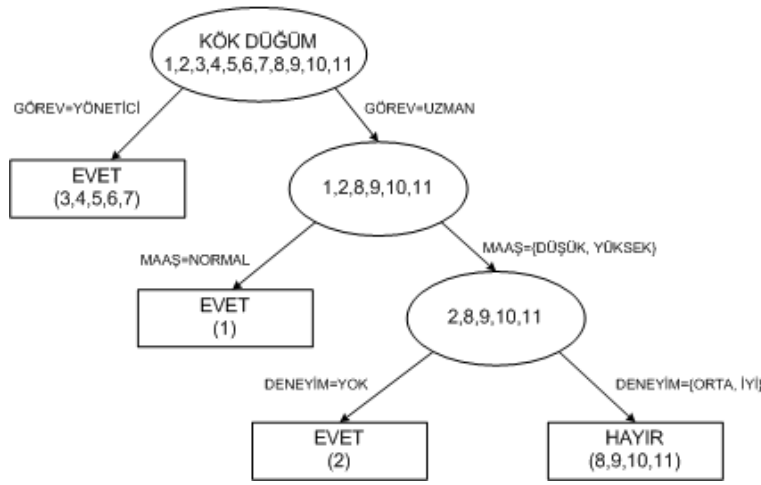



9

## Twoing Algoritması

**Örnek – devam**

- Sonuç karar ağacı.




10

## Twoing Algoritması

**Örnek – devam**

Karar ağacından elde edilen kurallar

1. **EĞER** (GÖREV = YÖNETİCİ) **İSE** (MEMNUN = EVET)
2. **EĞER** (GÖREV = UZMAN) **VE** (MAAŞ = NORMAL) **İSE** (MEMNUN = EVET)
3. **EĞER** (GÖREV = UZMAN) **VE** (MAAŞ = DÜŞÜK **VEYA** MAAŞ = YÜKSEK) **VE** (DENEYİM=YOK) **İSE** (MEMNUN = EVET)
4. **EĞER** (GÖREV = UZMAN) **VE** (MAAŞ = DÜŞÜK **VEYA** MAAŞ = YÜKSEK) **VE** (DENEYİM = ORTA **VEYA** DENEYİM = İYİ) **İSE** (MEMNUN = HAYIR)

11

## Konular

- Twoing Algoritması
- Gini Algoritması

12

## Gini Algoritması

- **Gini algoritmasında nitelik değerleri iki parçaya ayrılarak bölümleme yapılır.**
- Her bölünme için $Gini_{sol}$ ve $Gini_{sağ}$ değerleri hesaplanır.

$$Gini_{sol} = 1 - \sum_{i=1}^{k} \left( \frac{|\text{Tsinif}_i|}{|B_{sol}|} \right)^2 \qquad Gini_{sag} = 1 - \sum_{i=1}^{k} \left( \frac{|\text{Tsinif}_i|}{|B_{sag}|} \right)^2$$

- Burada, $Tsinif_i$ soldaki bölümdeki her bir sınıf değerini, $Tsinif_i$ sağdaki bölümdeki her bir sınıf değerini, $|B_{sol}|$ sol bölümdeki tüm değer sayısını, $|B_{sağ}|$ sağ bölümdeki tüm değer sayısını gösterir.

$$Gini_j = \frac{1}{n}\left( |B_{sol}| Gini_{sol} + |B_{sag}| Gini_{sag} \right)$$

- Her bölümlemeden sonra **Gini değeri en küçük olan seçilir.**

13

## Gini Algoritması

**Örnek**

| SIRA | EĞİTİM | YAŞ | CİNSİYET | SONUÇ |
|---|---|---|---|---|
| 1 | ORTA | YAŞLI | ERKEK | EVET |
| 2 | İLK | GENÇ | ERKEK | HAYIR |
| 3 | YÜKSEK | ORTA | KADIN | HAYIR |
| 4 | ORTA | ORTA | ERKEK | EVET |
| 5 | İLK | ORTA | ERKEK | EVET |
| 6 | YÜKSEK | YAŞLI | KADIN | EVET |
| 7 | İLK | GENÇ | KADIN | HAYIR |
| 8 | ORTA | ORTA | ERKEK | EVET |

| | EĞİTİM | | YAŞ | | CİNSİYET | |
|---|---|---|---|---|---|---|
| SONUÇ | İLK | ORTA,YÜKSEK | GENÇ | ORTA,YAŞLI | KADIN | ERKEK |
| EVET | 1 | 4 | 0 | 5 | 1 | 4 |
| HAYIR | 2 | 1 | 2 | 1 | 2 | 1 |

14

## Gini Algoritması

**Örnek**

| | EĞİTİM | | YAŞ | | CİNSİYET | |
|---|---|---|---|---|---|---|
| SONUÇ | İLK | ORTA,YÜKSEK | GENÇ | ORTA,YAŞLI | KADIN | ERKEK |
| EVET | 1 | 4 | 0 | 5 | 1 | 4 |
| HAYIR | 2 | 1 | 2 | 1 | 2 | 1 |

**EĞİTİM için**

$$Gini_{sol} = 1 - \left[\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right] = 0{,}444$$

$$Gini_{sag} = 1 - \left[\left(\frac{4}{5}\right)^2 + \left(\frac{1}{5}\right)^2\right] = 0{,}320$$

15

## Gini Algoritması

**Örnek**

| | EĞİTİM | | YAŞ | | CİNSİYET | |
|---|---|---|---|---|---|---|
| SONUÇ | İLK | ORTA,YÜKSEK | GENÇ | ORTA,YAŞLI | KADIN | ERKEK |
| EVET | 1 | 4 | 0 | 5 | 1 | 4 |
| HAYIR | 2 | 1 | 2 | 1 | 2 | 1 |

**YAŞ için**

$$Gini_{sol} = 1 - \left[\left(\frac{0}{2}\right)^2 + \left(\frac{2}{2}\right)^2\right] = 0$$

$$Gini_{sag} = 1 - \left[\left(\frac{5}{6}\right)^2 + \left(\frac{1}{6}\right)^2\right] = 0{,}278$$

16

## Gini Algoritması

**Örnek**

| | EĞİTİM | | YAŞ | | CİNSİYET | |
|---|---|---|---|---|---|---|
| SONUÇ | İLK | ORTA,YÜKSEK | GENÇ | ORTA,YAŞLI | KADIN | ERKEK |
| EVET | 1 | 4 | 0 | 5 | 1 | 4 |
| HAYIR | 2 | 1 | 2 | 1 | 2 | 1 |

**CİNSİYET için**

$$Gini_{sol} = 1 - \left[\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right] = 0{,}444$$

$$Gini_{sag} = 1 - \left[\left(\frac{4}{5}\right)^2 + \left(\frac{1}{5}\right)^2\right] = 0{,}320$$

17

## Gini Algoritması

**Örnek**

| | EĞİTİM | | YAŞ | | CİNSİYET | |
|---|---|---|---|---|---|---|
| SONUÇ | İLK | ORTA,YÜKSEK | GENÇ | ORTA,YAŞLI | KADIN | ERKEK |
| EVET | 1 | 4 | 0 | 5 | 1 | 4 |
| HAYIR | 2 | 1 | 2 | 1 | 2 | 1 |

**Gini değerleri**

$$Gini_{EGITIM} = \frac{3(0{,}444) + 5(0{,}320)}{8} = 0{,}367$$

$$Gini_{YAS} = \frac{2(0) + 6(0{,}278)}{8} = 0{,}209$$

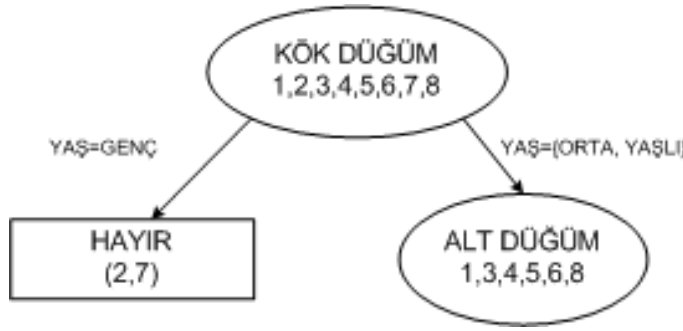
$$Gini_{CINSIYET} = \frac{3(0{,}444) + 5(0{,}320)}{8} = 0{,}367$$

**İlk bölünme YAŞ niteliğine göre yapılacaktır.**

18

## Gini Algoritması

**Örnek**

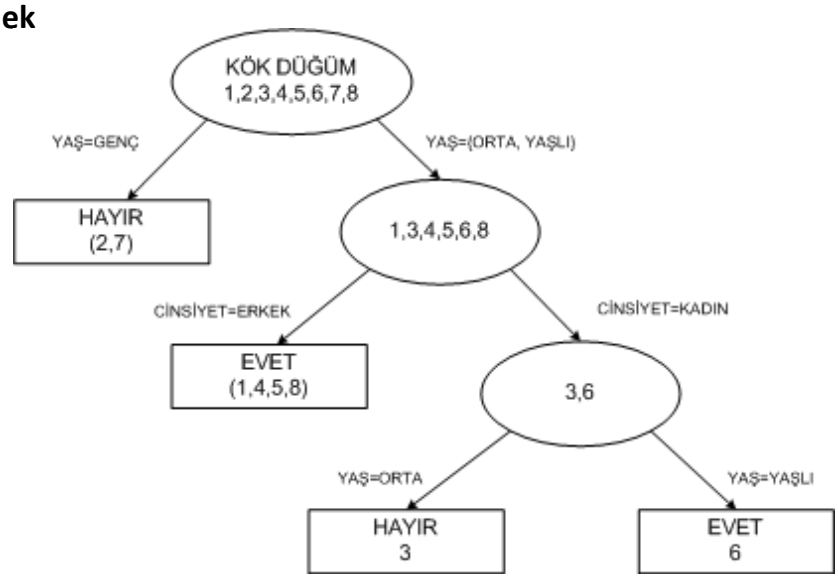



**Aynı işlemler ALT DÜĞÜM için tekrarlanır.**

19

## Gini Algoritması

**Örnek**




20

## Gini Algoritması

**Örnek – devam**

Karar ağacından elde edilen kurallar

1. **EĞER** (YAŞ = GENÇ) **İSE** (SONUÇ = HAYIR)
2. **EĞER** (YAŞ = ORTA **VEYA** YAŞ = YAŞLI) **VE** (CİNSİYET = ERKEK) **İSE** (SONUÇ = EVET)
3. **EĞER** (YAŞ = ORTA **VEYA** YAŞ = YAŞLI) **VE** (CİNSİYET = KADIN) **VE** (YAŞ = YAŞLI) **İSE** (SONUÇ = EVET)
4. **EĞER** (YAŞ = ORTA **VEYA** YAŞ = YAŞLI) **VE** (CİNSİYET = KADIN) **VE** (YAŞ = ORTA) **İSE** (SONUÇ = HAYIR)