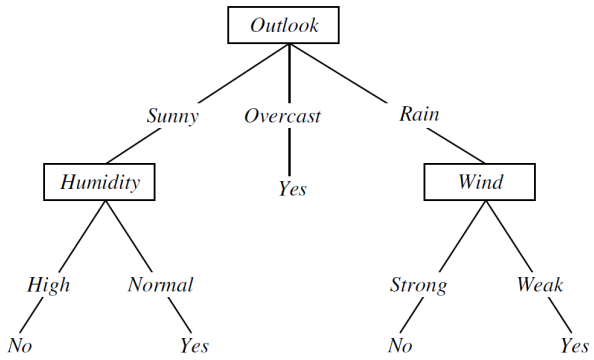
KU KONKUK UNIVERSITY

Decision Tree

Decision Tree

- 결정 트리
  - 자질들의 정보 획득량(information gain)에 따라 트리 형태의 규칙을 자동 생성하는 기계학습 모델
- 예제: Play Tennis
  - 테니스 시합을 할 수 있을지 판단하는 문제



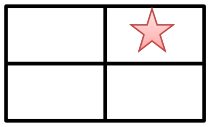
자질 (Feature)

- 자질 (feature)
  - 문제 해결에 영향을 미치는, 판단 근거가 되는 요소
  - 관측 및 측정 가능한 요소
- 예제: Play Tennis
  - 날씨 (outlook)
    - Sunny, Overcast, Rain
  - 온도 (temperature)
    - Hot, Mild, Cool
  - 습도 (humidity)
    - High, Normal, Low
  - 바람 (wind)
    - Strong, Weak

정보 획득량 (Information Gain)

- 정보 획득량
  - 자질의 값을 알게 됨으로써 얻어지는 문제 복잡도(전체 엔트로피)에 대한 감소 기대치
- 엔트로피 (entropy)
  - 문제의 복잡도를 측정하는 척도
  - 정보 이론 (information theory): 특정 확률 p를 가진 메시지를 상대방에게 전달하는데 필요한 비트 수에 대한 기댓값
  - 예제
    - 상자의 상단에 별이 위치한다는 정보를 상대방에게 전달하려면 몇 비트가 필요할까? 우측 상단에 별이 있다면?

$-log_2\,(1/2)$
$= -log_2\,(2^{-1})$
$= 1$



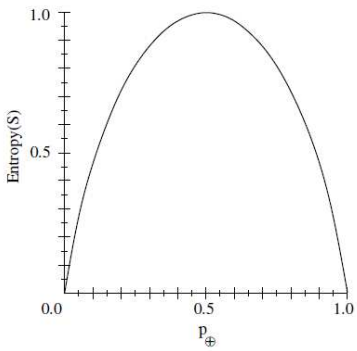
$-log_2\,(1/4)$
$= -log_2\,(2^{-2})$
$= 2$

엔트로피 (Entropy)

- 엔트로피의 최대값
  - 가장 애매한 확률(=1/2)을 가질 때 판단을 내리기 가장 힘듦 → 문제가 복잡함 → 최대 엔트로피를 가짐
- 기댓값
  - 각 사건이 벌어졌을 때의 이득과 그 사건이 벌어질 확률을 곱한 것을 전체 사건에 대해 합한 값

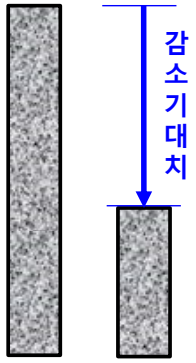
- S is a sample of training examples
- $p_{\oplus}$ is the proportion of positive examples in S
- $p_{\ominus}$ is the proportion of negative examples in S
- Entropy measures the impurity of S

$$Entropy(S) \equiv -p_{\oplus} \log_2 p_{\oplus} - p_{\ominus} \log_2 p_{\ominus}$$



정보 획득량 (Information Gain)

- 정보 획득량
  - 자질의 값을 알게 됨으로써 얻어지는 문제 복잡도(전체 엔트로피)에 대한 감소 기대치



$$Gain(S,A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Training Examples for Play Tennis

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Values(Wind) = { Weak, Strong }

S = [9+, 5-]
S_{Weak} = [6+, 2-]
S_{Strong} = [3+, 3-]

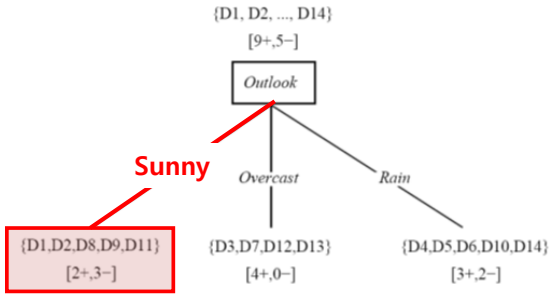
IG(S,Wind) = Entropy(S) - 8/14Entropy(S_{Weak})
- 6/14Entropy(S_{Strong})
= 0.940 - 8/14*0.811
- 6/14*1.00
= 0.048

IG(S,Outlook) = 0.246
IG(S,Humidity) = 0.151
IG(S,Temperature) = 0.029

∴ 정보획득량이 가장 큰 'Outlook'을 DT의 최우선 노드로 결정하는 것이 최적임

Training Examples for Play Tennis

Day		Temperature	Humidity	Wind	PlayTennis
D1		Hot	High	Weak	No
D2		Hot	High	Strong	No
D3		Hot	High	Weak	Yes
D4		Mild	High	Weak	Yes
D5		Cool	Normal	Weak	Yes
D6		Cool	Normal	Strong	No
D7		Cool	Normal	Strong	Yes
D8		Mild	High	Weak	No
D9		Cool	Normal	Weak	Yes
D10		Mild	Normal	Weak	Yes
D11		Mild	Normal	Strong	Yes
D12		Mild	High	Strong	Yes
D13		Hot	Normal	Weak	Yes
D14		Mild	High	Strong	No



S = [2+, 3-]
S_{Hot} = [0+, 2-]
S_{Mild} = [1+, 1-]
S_{Cool} = [1+, 0-]

S = [2+, 3-]
S_{High} = [0+, 3-]
S_{Normal} = [2+, 0-]

S = [2+, 3-]
S_{Strong} = [1+, 1-]
S_{Weak} = [1+, 2-]

IG(S,Temp)=Entropy(S)-2/5Entropy(S_{Hot})-2/5Entropy(S_{Mild})-1/5Entropy(S_{Cool})=0.971-2/5*0-2/5*1-1/5*0=**0.571**

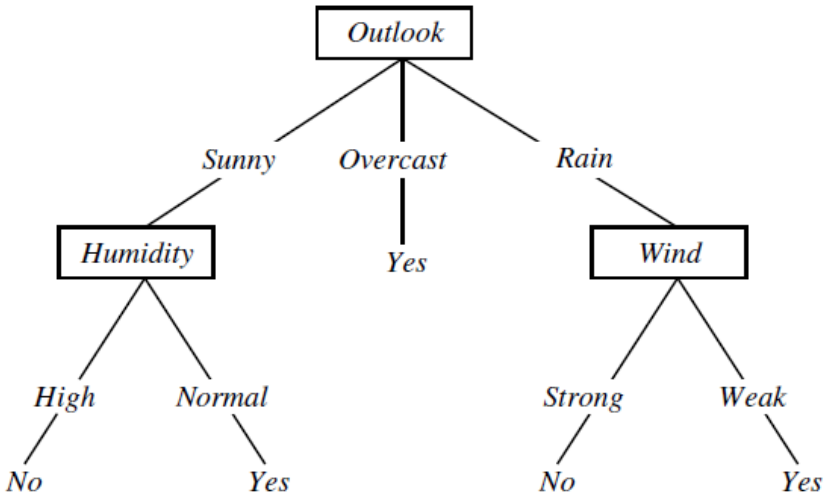
IG(S,Hum)=

IG(S,Wind)=

?

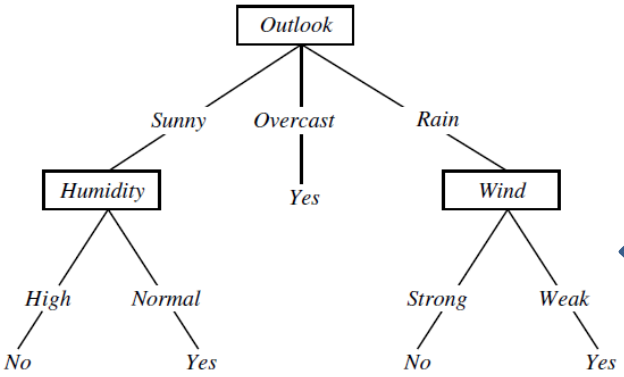

Edited by Harksoo Kim

Decision Tree for Play Tennis




Edited by Harksoo Kim

Converting a Tree to Rules



IF $(Outlook = Sunny) \wedge (Humidity = High)$
THEN $PlayTennis = No$

IF $(Outlook = Sunny) \wedge (Humidity = Normal)$
THEN $PlayTennis = Yes$

...

자질 구성 시 고려사항

- 연속 자질 (continuous feature)
  - 엔트로피 측정을 위해서는 이산 자질(discrete feature)로 변환
  - 예제
    - 기온<15 → Cool, 15<=기온<25 → Mild, 기온>=25 → Hot
- 복잡한 자질
  - 빈도가 낮아서 엔트로피가 높게 측정 됨
  - 문제 해결에 독립적으로 작용 가능하다면 분할하는 것이 좋음
  - 예제
    - Date: 2021년-5월-10일 → Year: 2021년, Month: 5월, Day: 10일
- 비싼 자질
  - 측정 및 관측 비용이 많이 드는 자질은 비용을 엔트로피에 반영하는 것을 고려할 수 있음

Advantages and Disadvantages

- Advantages
  - Does feature selection
  - Handles features of different types
  - Very fast prediction
  - Interpretable decision rules
- Disadvantages
  - Does not combine feature values, difficulty with dependent features
  - Finding optimal trees is intractable
- Implementations
  - C4.5: Free
  - CART: Commercial

질의응답

Q&A

Homepage: http://nlp.konkuk.ac.kr
E-mail: nlpdrkim@konkuk.ac.kr