

의사결정나무

Decision tree-Basic Idea : Entropy

- 불순도?
 - 서로 다른 data가 얼마나 섞여 있는가?
 - 높을 수록 유사한 비율로 data가 섞여 있음을 의미
- 불순도를 평가하는 함수 : Entropy

» A function measuring the degree of disorder (heterogeneity)

$$Entropy(set) = -P_1 \log_2 P_1 - P_0 \log_2 P_0$$

P_1 = the probability that 1 appears in set

P_0 = the probability that 0 appears in set

$A = \{1, 1, 1, 0, 0, 0, 0, 0\}$

$$Entropy(A) = -3/8 \log_2 3/8 - 5/8 \log_2 5/8$$

$$= 0.954$$

$$P_1 = 3/8$$

$$P_0 = 5/8$$



Decision tree-Basic Idea : Entropy

$$Entropy(set) = - \sum_{i=1}^n P_{c_i} \log_2 P_{c_i}$$

P_{c_i} = the probability that c_i appears in set

0, 1만 나올 확률이 아니라 0, 1, 2 ... 등이 나올 확률인 경우 확장해서 사용한다.

Homogenous Split

- 어떤 집합을 2개로 나눈 경우 얼마나 순수하게 나뉘었는지에 대한 지표

$$1, 1, 1, 1, 0, 0, 0, 0$$
$$P_1 = \{\{1, 1, 0, 0, 0\}, \{1, 1, 0\}\}$$
$$P_2 = \{\{1, 1, 1, 1, 0\}, \{0, 0, 0\}\}$$

Partitioning 1

$$\begin{array}{l} 1, 1, 0, 0, 0 \\ 1, 1, 0 \end{array}$$
$$-P_1 \log_2 P_1 - P_0 \log_2 P_0 = -\frac{2}{5} \log_2 \frac{2}{5} - \frac{3}{5} \log_2 \frac{3}{5} = 0.97$$
$$-P_1 \log_2 P_1 - P_0 \log_2 P_0 = -\frac{2}{3} \log_2 \frac{2}{3} - \frac{1}{3} \log_2 \frac{1}{3} = 0.92$$
$$\frac{5}{8} \times 0.97 + \frac{3}{8} \times 0.92 = 0.95$$

Partitioning 2

$$\begin{array}{l} 1, 1, 1, 1, 0 \\ 0, 0, 0 \end{array}$$
$$-P_1 \log_2 P_1 - P_0 \log_2 P_0 = -\frac{4}{5} \log_2 \frac{4}{5} - \frac{1}{5} \log_2 \frac{1}{5} = 0.72$$
$$-P_1 \log_2 P_1 - P_0 \log_2 P_0 = -\frac{3}{3} \log_2 \frac{3}{3} - \frac{0}{3} \log_2 \frac{0}{3} = 0.00$$
$$\frac{5}{8} \times 0.72 + \frac{3}{8} \times 0 = 0.51$$

Entropy 값
의 평균을
비교하여 더
작은 값이
더 순수하다.

Gain : 지표

$$\text{Gain} = \text{Entropy before split} - \text{Entropy after split}$$

$$1, 1, 1, 1, 0, 0, 0, 0 \quad \text{Entropy} = 1.0$$

Partitioning 1

$$1, 1, 0, 0, 0$$

$$1, 1, 0$$

$$\text{Entropy} = \frac{5}{8} \times 0.97 + \frac{3}{8} \times 0.92 = 0.95 \quad \text{Gain} = 1 - 0.95 = 0.05$$


Partitioning 2

$$1, 1, 1, 1, 0$$

$$0, 0, 0$$

$$\text{Entropy} = \frac{5}{8} \times 0.72 + \frac{3}{8} \times 0 = 0.51 \quad \text{Gain} = 1 - 0.51 = 0.49$$

Decision tree-Basic Idea



Day	A	B	C	Play
1	T	T	T	T
2	T	T	F	T
3	F	T	F	T
4	F	F	T	F
5	F	F	F	F

- Choose the most predictive INPUT, and
- Predict using that INPUT

가장 값을 잘 예측할 수 있는 input을 찾아내는 것이 목표

Decision tree-Basic Idea

Day	A	B	C	Play
1	T			T
2	T			T
3	F			T
4	F			F
5	F			F

If A = T then { T, T }
else { T, F, F }

Day	A	B	C	Play
1		T		T
2		T		T
3		T		T
4		F		F
5		F		F

If B = T then { T, T, T }
else { F, F }

Day	A	B	C	Play
1			T	T
2			F	T
3			F	T
4			T	F
5			F	F

If C = T then { T, F }
else { T, T, F }

각자 A, B, C 기준으로 나눈다.
출력을 가장 homogenous하게 split할 수 있는 입력

Decision tree-Basic Idea

» Choose the Most Descriptive Input

- Before partitioning

$$\text{Play: } \{3T, 2F\} \quad -P_T \log_2 P_T - P_F \log_2 P_F = -3/5 \log_2 3/5 - 2/5 \log_2 2/5 = 0.97$$

Day	A	B	C	Play
1	T	T	T	T
2	T	T	F	T
3	F	T	F	T
4	F	F	T	F
5	F	F	F	F

- Which attribute partitions “Play” with the lowest entropy?

$$\text{Play}_{A=T}: \{2T, 0F\} \quad -P_T \log_2 P_T - P_F \log_2 P_F = -1.0 \log_2 1.0 - 0.0 \log_2 0.0 = 0.0$$

$$\text{Play}_{A=F}: \{1T, 2F\} \quad -P_T \log_2 P_T - P_F \log_2 P_F = -1/3 \log_2 1/3 - 2/3 \log_2 2/3 = 0.92$$

$$\frac{2}{5} \times 0 + \frac{3}{5} \times 0.92 = 0.55$$

$$\text{Play}_{B=T}: \{3T, 0F\} \quad -P_T \log_2 P_T - P_F \log_2 P_F = -1.0 \log_2 1.0 - 0.0 \log_2 0.0 = 0.0$$

$$\text{Play}_{B=F}: \{0T, 2F\} \quad -P_T \log_2 P_T - P_F \log_2 P_F = -1.0 \log_2 1.0 - 0.0 \log_2 0.0 = 0.0$$

$$\frac{3}{5} \times 0.0 + \frac{2}{5} \times 0.0 = 0.0$$

$$\text{Play}_{C=T}: \{1T, 1F\} \quad -P_T \log_2 P_T - P_F \log_2 P_F = -1/2 \log_2 1/2 - 1/2 \log_2 1/2 = 1.0$$

$$\text{Play}_{C=F}: \{2T, 1F\} \quad -P_T \log_2 P_T - P_F \log_2 P_F = -2/3 \log_2 2/3 - 1/3 \log_2 1/3 = 0.92$$

$$\frac{2}{5} \times 1.0 + \frac{3}{5} \times 0.92 = 0.95$$

Gain : 처음 entropy –
나중 entropy
Gain이 가장 클수록
(Entropy가 가장 작을
수록) 좋음

이를 통해 규칙 생성
가능