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 = \text{the probability that 1 appears in } set P_0 = \text{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 0.00 = 0 0.00 = 0
```

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

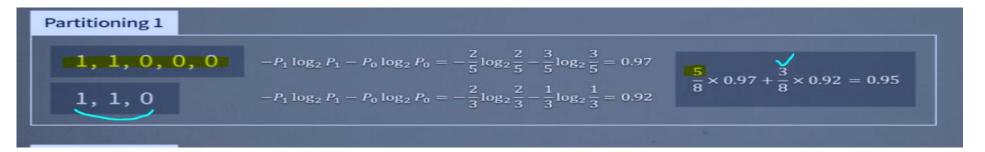
0, 1만 나올 확률이 아니라 0, 1, 2 ... 등이 나올 확률인 경우 확장해서 사용한다.

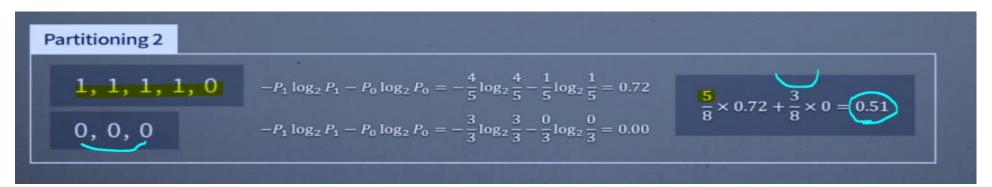
Homogenous Split

• 어떤 집합을 2개로 나눈 경우 얼마나 순수하게 나뉘었는지에 대

한 지표

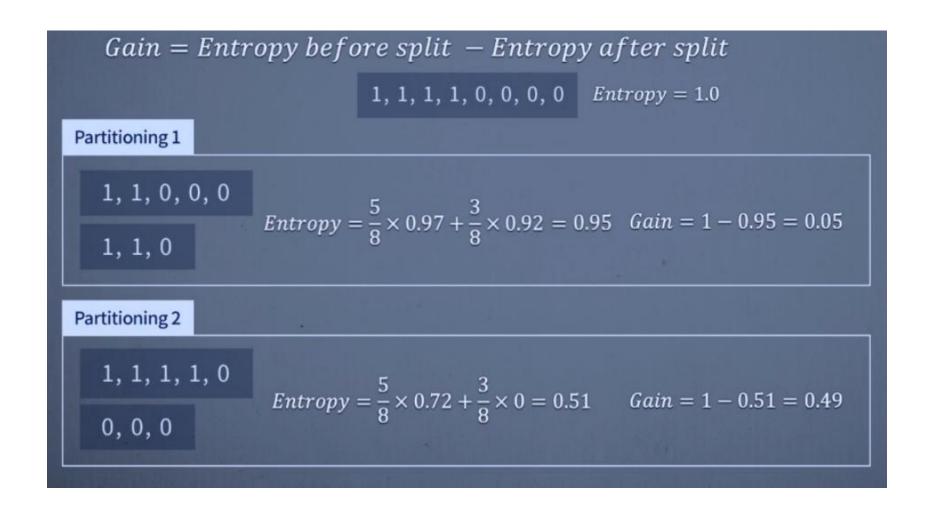
```
1, 1, 1, 1, 0, 0, 0, 0 P_1 = \{\{1,1,0,0,0\}, \{1,1,0\}\} \qquad P_2 = \{\{1,1,1,1,0\}, \{0,0,0\}\}
```





Entropy 값 의 평균을 비교하여 더 작은 값이 더 순수하다.

Gain : 지표



Decision tree-Basic Idea

1 T T T T T T T T T T T T T T T T T T T				Play
	1	T	T	T
3 F T F	2	Т	F	T
	3	F T	F	Т
4 F F T	4	F F	Т	F
5 F F F	5	FF	F	F

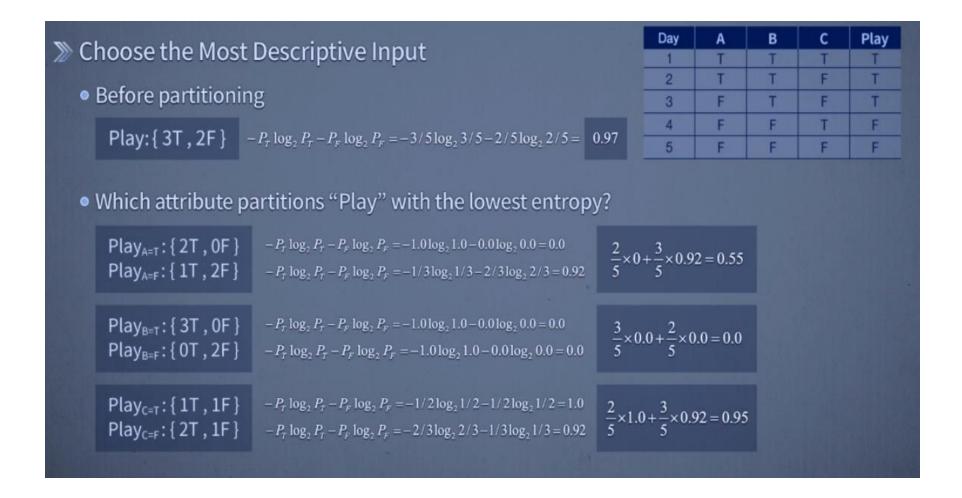
가장 값을 잘 예측할 수 있 는 input을 찾아내는 것이 목표

Decision tree-Basic Idea



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

Decision tree-Basic Idea



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

이를 통해 규칙 생성 가능