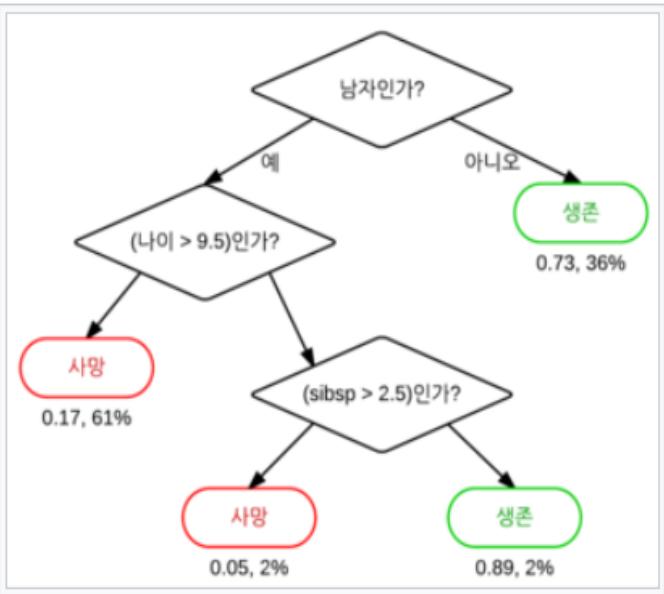
1. Decision Tree(결정트리) 학습법

1) 정의

- 어떤 항목에 대한 관측값과 목표값을 연결시켜주는 예측모델로써 결정트리를 사용하는 회귀 분류모두 가능한 지도학습
  - 결정트리: 의사결정 규칙과 그 결과들을 트리 구조로 도식화한 의사결정 지원 도구의 일종
  - 통계학, 데이터 마이닝, 기계학습에서 사용하는 예측 모델링 방법중 하나임
- 지도 분류 학습에서 가장 유용하게 사용되고 있음.
  - 학습: 학습에 사용되는 자료집합을 적절한 분할 기준 또는 분할 테스트에 따라 부분 집합들로 나누는 과정

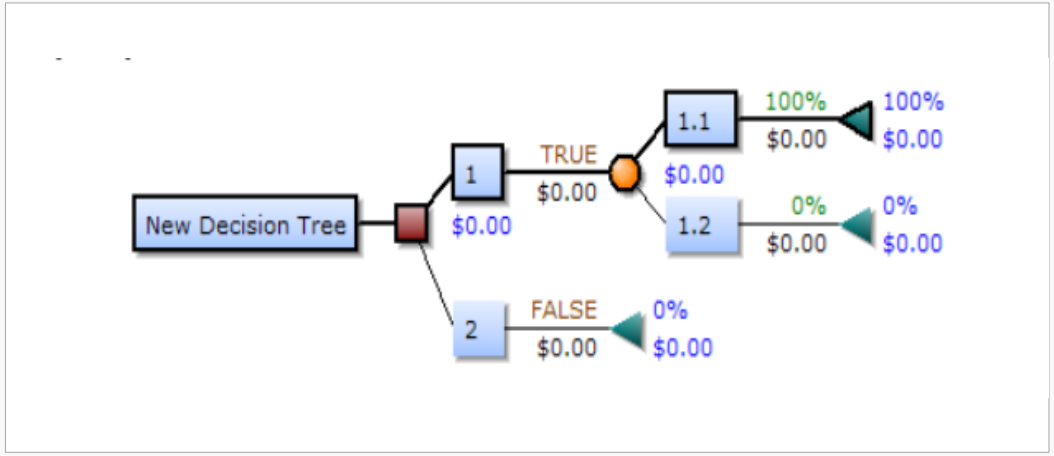


타이타닉호 탑승객의 생존 여부를 나타내는 결정 트리

2) 결정트리

- 데이터를 분석하여 이들 사이에 존재하는 패턴을 예측 가능한 규칙들의 조합으로 나타내며 그 모양이 '나무' 와 같다고 해서 붙여진 이름임
  - 운용과학, 그 중에서도 의사 결정 분석에서 목표에 가장 가까운 결과를 낼 수 있는 전략을 찾기 위해 주로 사용됨

- 결정트리 노드
  - 결정 노드(Decision node): 사각형으로 표시
  - 기회 노트(chance node): 원으로 표시
  - 종단 노드(end node): 삼각형으로 표시

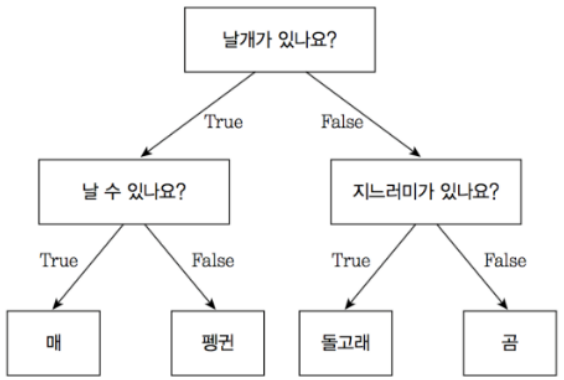


1) 정의

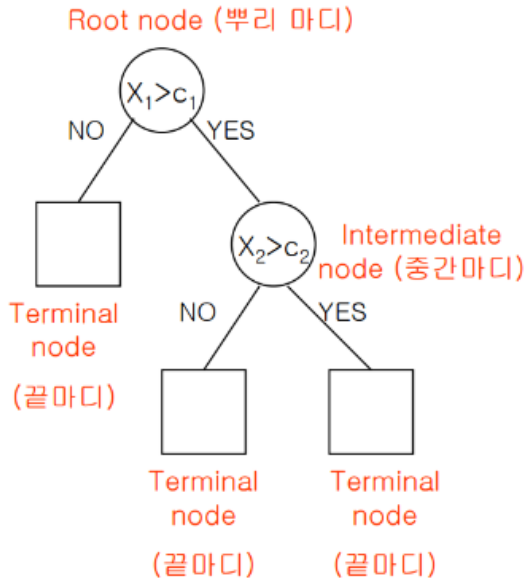
- 입력변수 영역을 두개로 구분하는 재귀적분기와 너무 자세하게 구분된 영역을 통합하는 가지치키(pruning) 두가지 과정으로 나누어짐
- 학습에 사용되는 자료를 순환분할로 나누어서 더 이상 새로운 예측값이 추가되지 않거나 부분 집합의 노드가 목표 변수와 같은 값을 지닐때까지 계속됨

결정트리 분석법	정의
분류 트리	• 예측된 결과로 입력 데이터가 분류되는 클래스를 출력
회귀 트리분석	• 예측된 결과로 특정 의미를 지니는 실수값을 출력

2) 예시



출처: 텐서 플로우 블로그

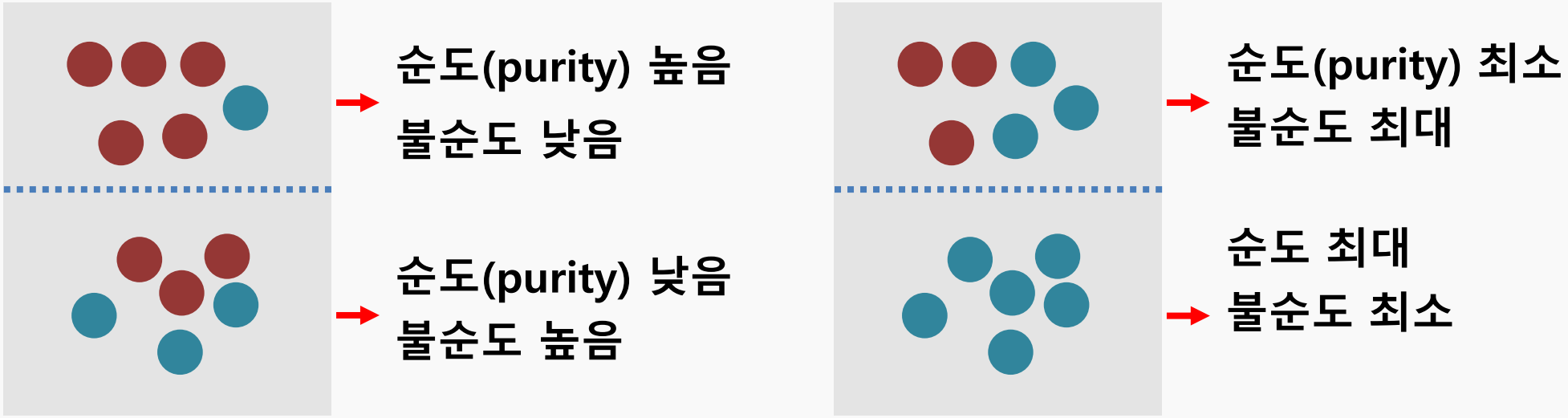


Terminal Node 또는Leaf Node
-> 분리된 집합의 개수(부분집합개수)

회귀의경우: 해당 터미널노드의 종속변수(y)의 평균을 예측값으로 반환, 예측값은 터미널노드의 개수와 일치함.

3) 불순도(impurity) 또는 불확실성(uncertainty)

- 불순도: 해당 범주안에 서로 다른 데이터가 얼마나 섞여 있는지를 뜻함.
- 결정트리는 불순도를 최소화(순도를 최대화) 하는 방향으로 진행함



- 불순도 지표는 엔트로피 또는 지니계수, 오분류오차가 있음.

4) 엔트로피(Entropy)

- 불순도를 수치적으로 나타낸 척도, 엔트로피가 높으면 불순도가 높다는 뜻임

  - 엔트로피가 1이면 불순도 최대 (한 범주안에 데이터가 반반씩 있음을 뜻함)
  - 엔트로피가 0이면 불순도 최소(한 범주안에 같은종류의 데이터만 있음)

$$\text{Entropy} = -\sum_i (p_i) \log_2 (p_i)$$

엔트로피 공식

(Pi = 한 영역 안에 존재하는 데이터 가운데 범주 i에 속하는 데이터의 비율)

경사	표면	속도제한	속도
가파름	울퉁불퉁	Yes	Slow
가파름	완만	Yes	Slow
내리막	울퉁불퉁	No	fast
가파름	완만	No	fast

=-P_slow*log2(P_slow)-P_fast*log2(P_fast)

=-0.5*log2(0.5)-0.5*log2(0.5)=-0.5*-1-0.5*-1=1

5) 정보획득(Information gain)

- 분기이전의 엔트로피에서 분기이후의 엔트로피를 뺀 수치가 정보획득량임

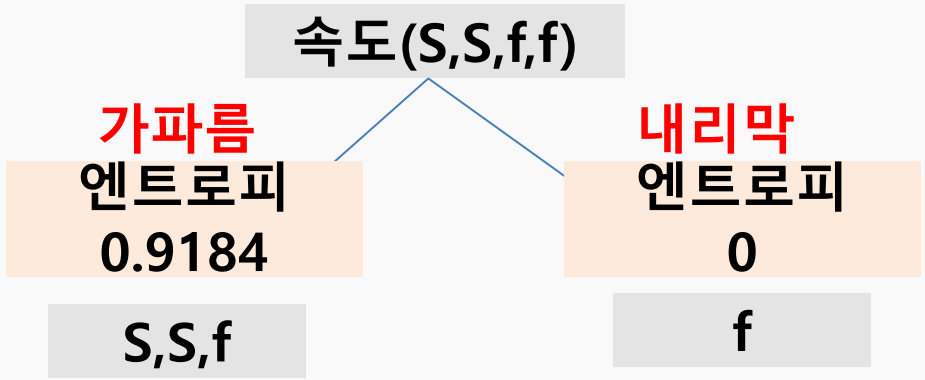
  - 엔트로피가 1인 상태에서 0.7인 상태로 바뀌었을때 정보획득은 0.3임

  =entropy(parent)-[weighted average]*entropy(children)

- 결정트리 알고리즘은 정보획득을 최대화하는 방향으로 학습이 진행됨

  - 어느 feature의 어느 분기점에서 정보획득이 최대화되는지를 판단해서 분기가 진행됨

5) 정보획득 - 경사기준분기



경사	표면	속도제한	속도
가파름	울퉁불퉁	Yes	Slow
가파름	완만	Yes	Slow
내리막	울퉁불퉁	No	fast
가파름	완만	No	fast

[경사 속성을 가파름으로 분기하였을때 엔트로피]

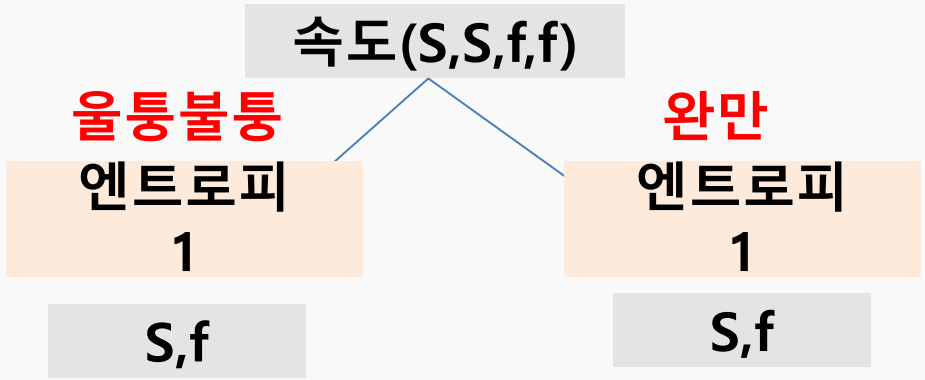
Slow=2/3, fast=1/3

=-P_slow*log2(P_slow)-P_fast*log2(P_fast) =-0.666*log2(0.666)-0.333*log2(0.333)=0.9184

[경사를 기준으로 분기한 후의 엔트로피값] =3/4*0.9184+1/4*0=0.6888

[정보획득] 1 - 0.6888 = 0.3112

5) 정보획득 - 표면기준분기



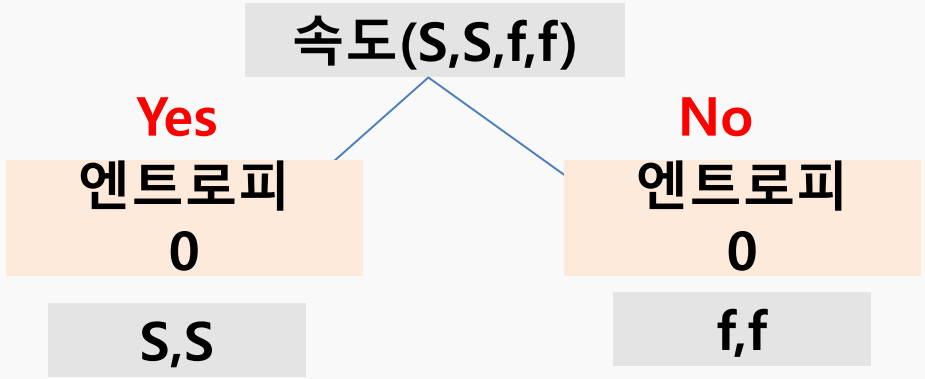
경사	표면	속도제한	속도
가파름	울퉁불퉁	Yes	Slow
가파름	완만	Yes	Slow
내리막	울퉁불퉁	No	fast
가파름	완만	No	fast

[표면속성으로 분기하였을때 울퉁불퉁과, 완만 모두 엔트로피 1나옴

=-P_Slow*log2(P_Slow)-P_fast*log2(P_fast)=1

[정보획득] 1 - 1 = 0

5) 정보획득 - 속도제한



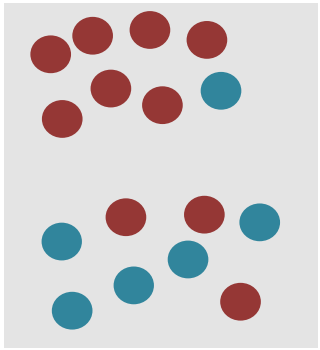
경사	표면	속도제한	속도
가파름	울퉁불퉁	Yes	Slow
가파름	완만	Yes	Slow
내리막	울퉁불퉁	No	fast
가파름	완만	No	fast

[속도제한으로 분기하였을때 Yes와 No 모두 엔트로피 0임]

=-P_Slow*log2(P_Slow)-P_fast*log2(P_fast)= -1*0-1*0 = 0

[정보획득] 1 - 0 = 1

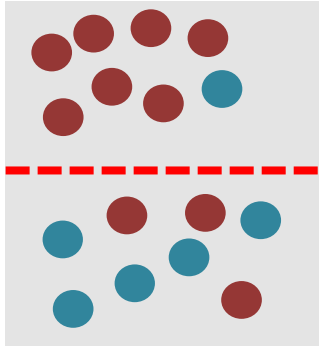
5) 정보획득 - 연속형 데이터



$$Entropy(A) = -\sum_{k=1}^{m} p_k \log_2 (p_k)$$

-10/16 * log(10/16) - 6/16 * log(6/16) =0.95

엔트로피
0.95



$$Entropy(A) = \sum_{i=1}^{d} R_i \left(-\sum_{k=1}^{m} p_k \log_2 (p_k) \right)$$

8/16 * (-7/8 * log(7/8) - 1/8 * log(1/8))
+
8/16 * (-3/8 * log(/8) - 5/8 * log(5/8))

엔트로피
0.75

정보획득
0.95-0.75=0.2

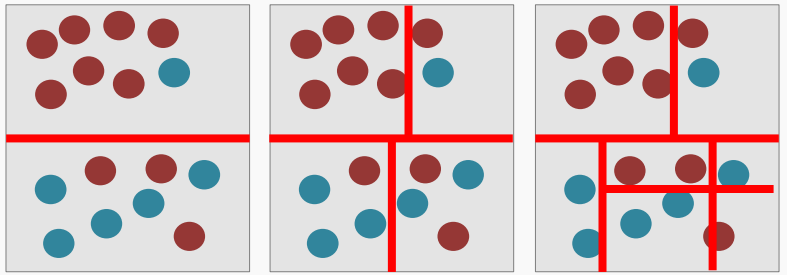
6) 가지치기(Pruning)

- 오버피팅을 막기위한 전략임. 트리에 가지가 너무 많다면 오버피팅으로 볼수 있음.
- 최대 깊이나 터미널 노드의 최대 개수, 혹은 한 노드가 분할하기 위한 최소 데이터수를 제한

- min_sample_split 파라미터를 조정하여 한 노드에 들어 있는 최소 데이터 수를 정해줄 수 있음.

  예) min_sample_split=10 이면 한 노드에 10개의 데이터가 있다면 그 노드는 더이상 분기 하지 않음

- max_depth를 통해서 최대 깊이를 지정해줄수도 있음.



- 의사결정나무는 가지치기의 비용함수를 최소화하는 분기를 찾아내도록 학습됨

의사결정나무의 비용 복잡도(오류가 적으면서 터미널 노드가 적은 단순한 모델일수록 작은값)
= 검증데이터에 대한 오분류율 + 알파값(0.01~0.1) 터미널 노드의 수(구조의 복잡도)

3. 의사결정트리 학습의 장정과 한계

장점

- 결과를 해석하고 이해하기 쉽다.
- 자료를 가공할 필요가 거의 없다.
- 수치자료와 범주 자료 모두 적용가능하다.
- 화이트박스 모델을 사용한다.
- 안정적이며 대규모의 데이터 셋에서도 잘 동작한다.

한계

- 결정트리 학습자가 훈련 데이터를 제대로 일반화하지 못할 경우 너무 복잡한 결정트리를 만들수 있다. (과적합)
- 가지치기 작업을 해야한다.
- 배타적논리합(XOR), 패리티, 멀티플렉서와 같은 문제 학습하기 어렵다.