의사결정 트리

4.1 기본 프로세스
4.2 분할 선택

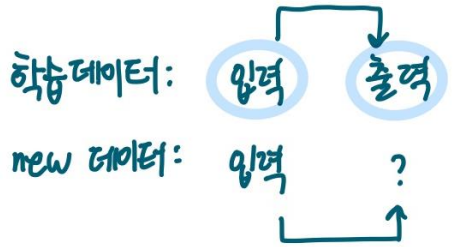
Intro

의사결정 나무는 어떤 종류의 학습법일까요?

4.0.0. 지도학습 vs 비지도학습

지도학습

- 학습 데이터의 입력과 출력 간의 관계를 학습하여 규칙이나 함수로 표현되는 모델을 찾음.



비지도학습

- 출력 정보가 없는 학습 데이터에 대해서 데이터의 패턴을 발견하는 학습

** 입력 데이터 = 독립 변수, 출력 데이터 = 종속 변수

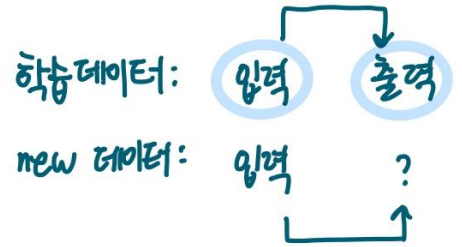
4.0.0. 의사결정 나무는 지도학습

지도학습

- 학습 데이터의 입력과 출력 간의 관계를 학습하여 규칙이나 함수로 표현되는 모델을 찾음.

데이터 유형

분류	회귀
출력 값이 범주형 데이터인 경우	출력 값이 연속형 데이터인 경우



의사결정나무는
분류문제 회귀 문제 둘 다 학습

* 입력값은 범주형이든 연속형이든 상관 없다.

** 입력 데이터 = 독립 변수, 출력 데이터 = 종속 변수

4.1.1. 의사결정 나무 구성요소

◎ root node 루트 노드 : 나무구조가 시작되는 노드. (=모든 샘플의 집합)

- 트리의 최상위에 위치, 가장 중요한 속성을 배치 ← 상위의 노드일수록 중요한 속성을 배치

◎ leaf node or terminal node 단말 노드 : 하위에 다른 노드가 연결되어 있지 않은 노드

◎ branch node 분기 노드 : 트리 구조에서 최소한 한 개의 자식 노드를 갖는 노드

◎ parent node 부모 노드 : 임의의 노드 바로 위에 있는 노드

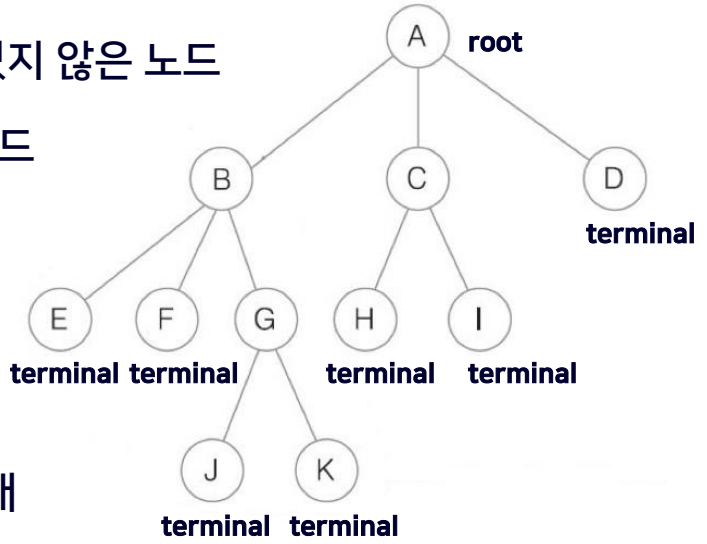
◎ child node 자식 노드 : 임의의 노드 바로 아래에 있는 노드

◎ sibling node 형제 노드 : 동일한 부모 노드를 가지는 노드들

◎ level 레벨 : 루트 노드에서 임의의 노드까지 방문한 노드의 수 ex) 레벨 1, 1세대

◎ depth 깊이 : 트리의 최대 레벨

◎ degree 차수 : 하위 트리의 개수 = 간선 수

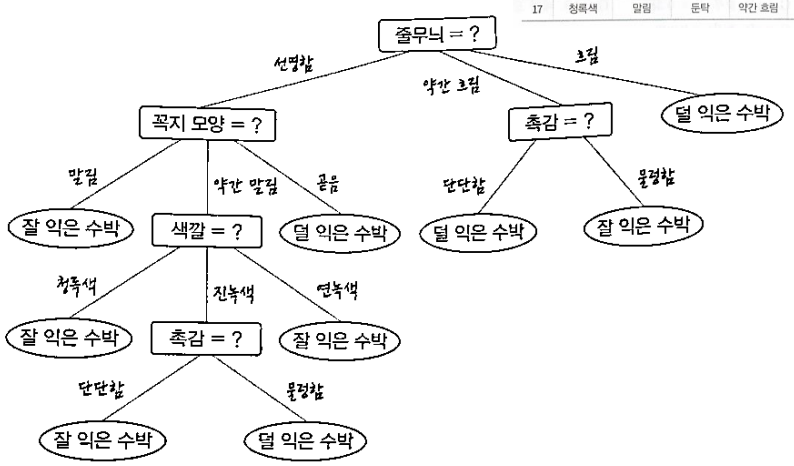


4.1.2. 기본 프로세스

표 4.1 ＼ 수박 데이터 세트 2.0

번호	색깔	꼭지 모양	소리	줄무늬	배꼽 모양	촉감	잘 익은 수박
1	청록색	말림	혼탁	선명함	움푹 패임	단단함	예
2	진녹색	말림	둔탁	선명함	움푹 패임	단단함	예
3	진녹색	말림	혼탁	선명함	움푹 패임	단단함	예
4	청록색	말림	둔탁	선명함	움푹 패임	단단함	예
5	연녹색	말림	혼탁	선명함	움푹 패임	단단함	예
6	청록색	약간 말림	혼탁	선명함	약간 패임	물렁함	예
7	진녹색	약간 말림	혼탁	약간 흐림	약간 패임	물렁함	예
8	진녹색	약간 말림	혼탁	선명함	약간 패임	단단함	예
9	진녹색	약간 말림	둔탁	약간 흐림	약간 패임	단단함	아니오
10	청록색	곧음	맑음	선명함	평평함	물렁함	아니오
11	연녹색	곧음	맑음	흐림	평평함	단단함	아니오
12	연녹색	말림	혼탁	흐림	평평함	단단함	아니오
13	청록색	약간 말림	혼탁	약간 흐림	움푹 패임	단단함	아니오
14	연녹색	약간 말림	둔탁	약간 흐림	움푹 패임	단단함	아니오
15	진녹색	약간 말림	혼탁	선명함	약간 패임	물렁함	아니오
16	연녹색	말림	혼탁	흐림	평평함	단단함	아니오
17	청록색	말림	둔탁	약간 흐림	약간 패임	단단함	아니오

- 개념: 의사결정 규칙 decision rule을 나무구조로 도표화하여 분류와 예측을 수행하는 분석방법
- 직관적, 판정 질문(조건문)을 알 수 있다.
  - → 분류된 이유, 근거가 필요한 분야가 적용됨.
  - → 상위의 노드일수록 중요한 속성.
    - 예) 의료, 은행 대출, 카드 발급 대상 등
- if – then 구조 (If 판정질문, then 최종결론)



한 단계 위의 결정 결과는 정답 범위를 한정시키는 역할을 한다. (p. 90, 5 line)

'어떤 속성이 상위에 있냐', '속성값을 어떤 기준으로 구분하느냐' 에 따라 모델의 성능이 달라진다.

4.1.3. 의사결정 나무의 궁극적인 목표

표 4.1 ＼ 수박 데이터 세트 2.0

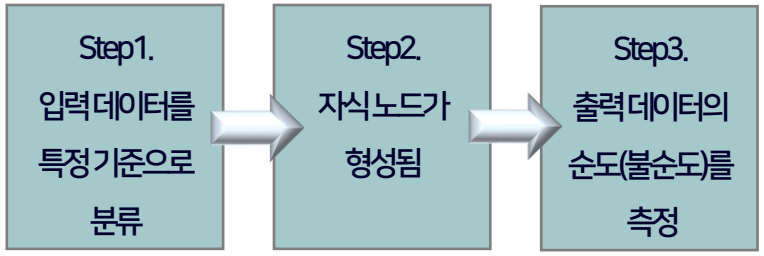
번호	색깔	꼭지 모양	소리	줄무늬	배꼽 모양	촉감	잘 익은 수박
1	청록색	말림	혼탁	선명함	움푹 패임	단단함	예
2	진녹색	말림	둔탁	선명함	움푹 패임	단단함	예
3	진녹색	말림	혼탁	선명함	움푹 패임	단단함	예
4	청록색	말림	둔탁	선명함	움푹 패임	단단함	예
5	연녹색	말림	혼탁	선명함	움푹 패임	단단함	예
6	청록색	약간 말림	혼탁	선명함	약간 패임	물렁함	예
7	진녹색	약간 말림	혼탁	약간 흐림	약간 패임	물렁함	예
8	진녹색	약간 말림	혼탁	선명함	약간 패임	단단함	예
9	진녹색	약간 말림	둔탁	약간 흐림	약간 패임	단단함	아니오
10	청록색	곧음	맑음	선명함	평평함	물렁함	아니오
11	연녹색	곧음	맑음	흐림	평평함	단단함	아니오
12	연녹색	말림	혼탁	흐림	평평함	단단함	아니오
13	청록색	약간 말림	혼탁	약간 흐림	움푹 패임	단단함	아니오
14	연녹색	약간 말림	둔탁	약간 흐림	움푹 패임	단단함	아니오
15	진녹색	약간 말림	혼탁	선명함	약간 패임	물렁함	아니오
16	연녹색	말림	혼탁	흐림	평평함	단단함	아니오
17	청록색	말림	둔탁	약간 흐림	약간 패임	단단함	아니오

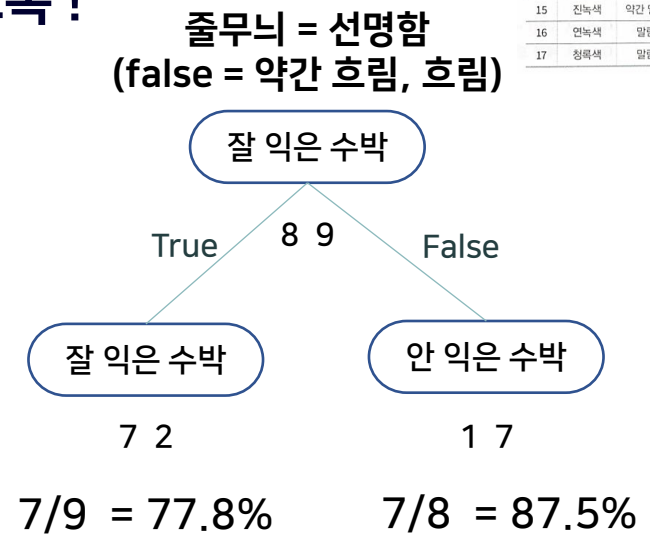
- **분기된 노드(자식 노드)가 최대한 같은 클래스에 속하도록 !**

  <첨부된 그림>에서 이상적인 목표: 잘 익은 수박의 표본이 9:0

  But 현실적으로 불가능 & 과적합 문제

- **분리 기준(속성) 정하기**





- **순도 purity**: 출력 데이터의 분포를 구별하는 정도

  순도↑ → 성능 ↑

  ⇔ 불순도, 불확실성(entrophy)

- **최적의 분할 속성 선택 = 순도가 높은 속성 = 불확실성이 낮은 속성**

4.2 분할 선택

분류 기준(분류 속성) : 순도를 측정해서 구함.

순도(불순도)를 측정하는데 자주 사용하는 지표 3개: 정보 이득(entrophy), 정보 이득율, 지니계수

4.2.1 정보이득

- **정보 이득 Information Gain**: 불확실성의 감소량. 정보이론에서 정보량을 측정하는 것을 기반.

  정보이득 ↑ or 정보량↑ → 정보의 가치 ↑

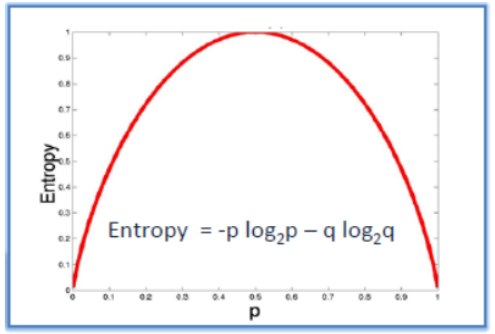
  불확실성 ↑ → 정보량 ↓ (∵ 익숙X)

  엔트로피를 이용하여 계산: Gain = Ent(분류하기 전 노드) – Ent(분류한 후의 노드들)

$$Entropy(A) = -\sum_{k=1}^{m} p_k \log_2 (p_k) \qquad Gain(T, X) = Entropy(T) - Entropy(T, X)$$

  *** log의 밑이 2인 이유:** IG는 컴퓨터의 정보량을 계산하는 것 → 컴퓨터의 정보량은 bit수로 계산
  2진수 : 0과1

  *** '-'를 붙이는 이유:** log의 진수에 확률이 들어감 -> 무조건 값이 음수임

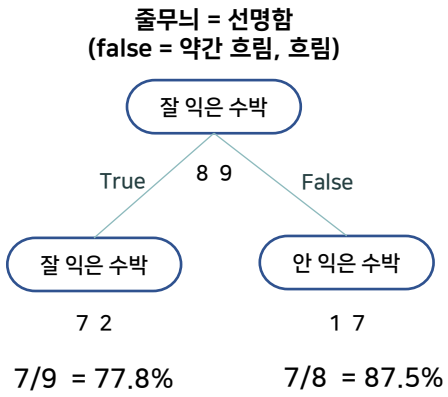


Entropy = -0.5 $\log_2$0.5 – 0.5 $\log_2$0.5 = 1

- **정보이득 = C(임의의 값) → 불확실성이 C만큼 감소했다.**

4.2.1 정보이득 계산 과정: $Gain(T,X) = Entropy(T) - Entropy(T,X)$ $Entropy(A) = -\sum_{k=1}^{m} p_k \log_2 (p_k)$

Case 1) 분류 속성 동일, branch수 다름

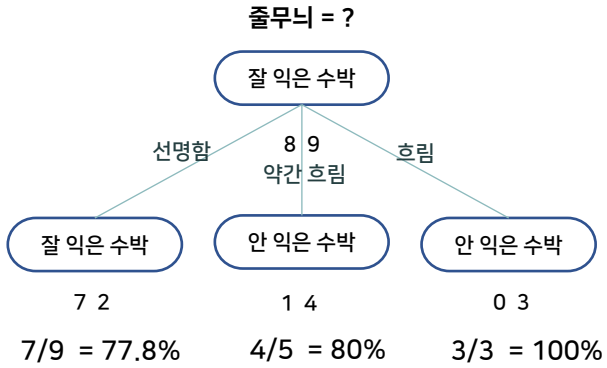


Ent(부모노드) $= -\left(\frac{8}{17}\log_2\frac{8}{17} + \frac{9}{17}\log_2\frac{9}{17}\right) = 0.998$

Ent(자식노드1) $= -\left(\frac{7}{9}\log_2\frac{7}{9} + \frac{2}{9}\log_2\frac{2}{9}\right) = 0.764$

Ent(자식노드2) $= -\left(\frac{1}{8}\log_2\frac{1}{8} + \frac{7}{8}\log_2\frac{7}{8}\right) = 0.544$

* Ent(줄무늬) $= \frac{9}{17}E(\text{자식노드1}) + \frac{8}{17}(\text{자식노드2})$
$= \frac{9}{17} \times 0.764 + \frac{8}{17} \times 0.544 = 0.661$
* Gain(줄무늬) $= 0.998 - 0.661 = 0.337$



Ent(자식노드1) $= -\left(\frac{7}{9}\log_2\frac{7}{9} + \frac{2}{9}\log_2\frac{2}{9}\right) = 0.764$

Ent(자식노드2) $= -\left(\frac{1}{5}\log_2\frac{1}{5} + \frac{4}{5}\log_2\frac{4}{5}\right) = 0.722$

Ent(자식노드3) $= -\left(\frac{0}{3}\log_2\frac{0}{3} + \frac{3}{3}\log_2\frac{3}{3}\right) = 0$
0이라고 가정.

* Ent(줄무늬 V.2) $= \frac{9}{17}(\text{자식노드1}) + \frac{5}{17}(\text{자식노드2}) + \frac{3}{17}(\text{자식노드3})$
$= \frac{9}{17} \times 0.764 + \frac{5}{17} \times 0.722 + \frac{3}{17} \times 0$
$= 0.617$
* Gain(줄무늬 V.2) $= 0.998 - 0.617 = 0.381$

- **가중치G**: 자식노드의 샘플 수 ↑ → 영향력↑

$$\text{가중치 } G = \frac{m\text{번째 자식 노드의 표본의 수}}{\text{부모 노드의 표본의 수}}$$

<첫번째 모형>에서 왼쪽 노드의 엔트로피를 계산할 때, 왼쪽 노드의 가중치는 9/17이고, 오른쪽 노드의 가중치는 8/17이다.

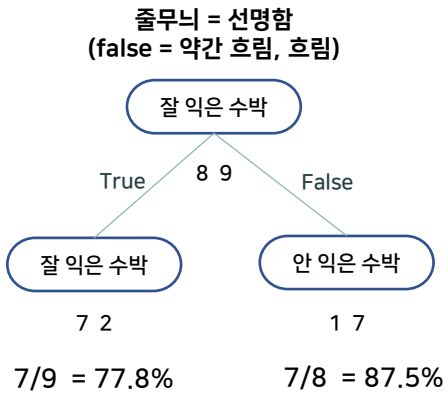
- **가정**: $P_k = 0$, $\log_2 P_k$의 값은 발산하지만,
Entropy 계산할 때는 $\log_2 P_k = 0$ 이라고 가정

★ 가지 branch = 2개 → 불확실성이 0.337만큼 감소.

★ 가지 branch = 3개 → 불확실성이 0.381만큼 감소. **성능 ↑**

4.2.1 정보이득 계산 과정: $Gain(T,X) = Entropy(T) - Entropy(T,X)$ $Entropy(A) = -\sum_{k=1}^{m} p_k \log_2 (p_k)$

Case 1) 분류 속성 동일, branch수 다름

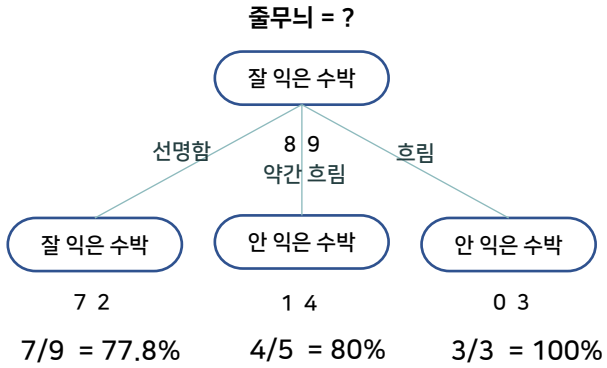


Ent(부모노드) $= -\left(\frac{8}{17}\log_2\frac{8}{17} + \frac{9}{17}\log_2\frac{9}{17}\right) = 0.998$

Ent(자식노드1) $= -\left(\frac{7}{9}\log_2\frac{7}{9} + \frac{2}{9}\log_2\frac{2}{9}\right) = 0.764$

Ent(자식노드2) $= -\left(\frac{1}{8}\log_2\frac{1}{8} + \frac{7}{8}\log_2\frac{7}{8}\right) = 0.544$

* Ent(줄무늬) $= \frac{9}{17}E(\text{자식노드1}) + \frac{8}{17}(\text{자식노드2})$
$= \frac{9}{17} \times 0.764 + \frac{8}{17} \times 0.544 = 0.661$
* Gain(줄무늬) $= 0.998 - 0.661 = 0.337$



Ent(자식노드1) $= -\left(\frac{7}{9}\log_2\frac{7}{9} + \frac{2}{9}\log_2\frac{2}{9}\right) = 0.764$

Ent(자식노드2) $= -\left(\frac{1}{5}\log_2\frac{1}{5} + \frac{4}{5}\log_2\frac{4}{5}\right) = 0.722$

Ent(자식노드3) $= -\left(\frac{0}{3}\log_2\frac{0}{3} + \frac{3}{3}\log_2\frac{3}{3}\right) = 0$
0이라고 가정.

* Ent(줄무늬 V.2) $= \frac{9}{17}(\text{자식노드1}) + \frac{5}{17}(\text{자식노드2}) + \frac{3}{17}(\text{자식노드3})$
$= \frac{9}{17} \times 0.764 + \frac{5}{17} \times 0.722 + \frac{3}{17} \times 0$
$= 0.617$
* Gain(줄무늬 V.2) $= 0.998 - 0.617 = 0.381$

- **가중치G**: 자식노드의 샘플 수 ↑ → 영향력↑

$$\text{가중치 } G = \frac{m\text{번째 자식 노드의 표본의 수}}{\text{부모 노드의 표본의 수}}$$

<첫번째 모형>에서 왼쪽 노드의 엔트로피를 계산할 때, 왼쪽 노드의 가중치는 9/17이고, 오른쪽 노드의 가중치는 8/17이다.

- **가정**: $P_k = 0$, $\log_2 P_k$의 값은 발산하지만,
Entropy 계산할 때는 $\log_2 P_k = 0$ 이라고 가정

★ 가지 branch = 2개 → 불확실성이 0.337만큼 감소.
★ 가지 branch = 3개 → 불확실성이 0.381만큼 감소. **성능 ↑**

가지↑ → Gain↑
가지 너무 많아도 성능 ↓
⇒ 가지의 개수에 따라 penalty 부여

정보이득율 Information Gain Ratio 등장

4.2.2 정보이득율

- **정보 이득율: 정보 이득의 확장된 개념.**

  등장 배경: 가지가 많아질수록 Gain 값이 커지는 경향.

  가지가 너무 많으면 오히려 성능이 떨어지는데, 가지가 많은 모형을 선택할 수 있음.

- **내재 값 Intrinsic value: 가지의 개수로 penalty를 부여**

$$IV(A) = -\sum_{m=1}^{m=n}\left(\frac{m\text{번째 자식 노드의 표본의 수}}{\text{부모 노드의 표본의 수}} * \log_2\left(\frac{m\text{번째 자식 노드의 표본의 수}}{\text{부모 노드의 표본의 수}}\right)\right)$$

$$\text{Gain ratio}(\mathrm{D}, \mathrm{A}) = \frac{Gain(D,A)}{IV(A)}$$

★ 가지 branch = 2개 → Gain = 0.337 → IV? Gain Ratio?

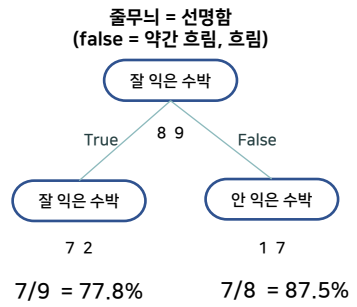
★ 가지 branch = 3개 → Gain = 0.381 → IV? Gain Ratio?

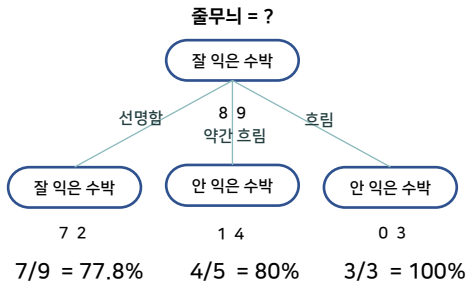
$IV(\text{줄무늬1}) = -\left(\frac{9}{17}\log_2\frac{9}{17} + \frac{8}{17}\log_2\frac{8}{17}\right) = 0.998$

$\text{Gain Ratio}(\cdot, \text{줄무늬1}) = \frac{0.337}{0.998} = 0.338$

$IV(\text{줄무늬2}) = -\left(\frac{9}{17}\log_2\frac{9}{17} + \frac{5}{17}\log_2\frac{5}{17} + \frac{3}{17}\log_2\frac{3}{17}\right) = 1.447$

$\text{Gain Ratio}(\cdot, \text{줄무늬2}) = \frac{0.381}{1.447} = 0.263$





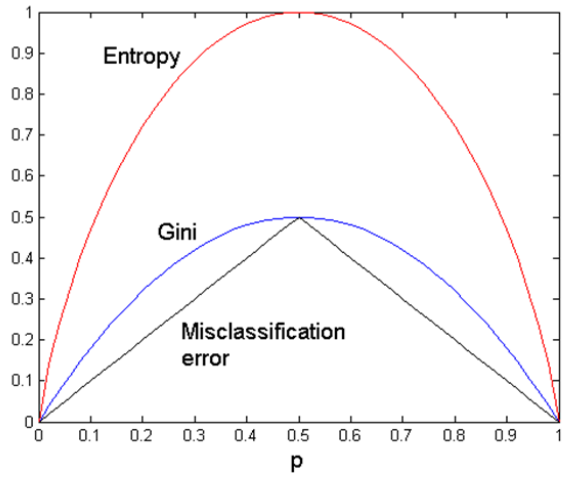
4.2.3 지니계수

- **지니계수: 불확실성(Entropy)을 의미. 정보이득 Information Gain과 같은 개념**

  지니계수↓ = 불확실성 ↓, 지니계수↓ → 성능 ↑

- **Information Gain과 계산 방법이 다름**

$$G.I(A) = \sum_{i=1}^{d} \left(R_i \left(1 - \sum_{k=1}^{m} p_{ik}^2 \right) \right)$$



*복원 추출 개념을 씀 → 두 번 측정하면 더 정확해서

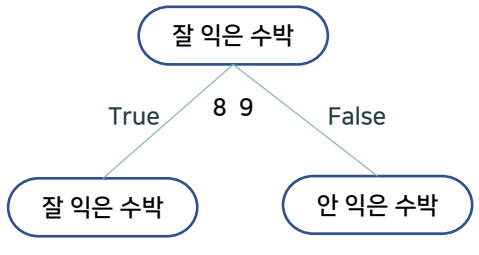
*불확실성을 구하는 것 → 1 – 순도

- **지니계수는 이진 분류만 한다. Information Gain에서 Ratio개념을 도입할 필요가 없다.**
- **이진 분류한 나무 모형 = 이진 트리**: 모든 노드들의 자식 노드가 두 개 이하인 트리

4.2.3 지니계수 계산 과정

Case 2) 분류 속성이 다름

줄무늬 = 선명함
(false = 약간 흐림, 흐림)



7/9 = 77.8%　　7/8 = 87.5%

$G(\text{부모}) = 1-\{(\frac{8}{17})^2+(\frac{9}{17})^2\} = 0.498$

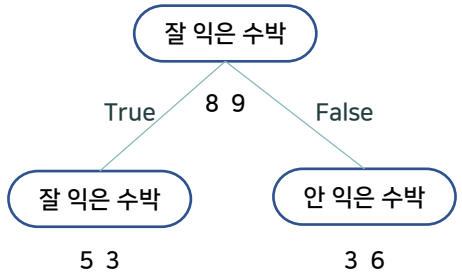
$G(\text{자식1}) = 1-\{(\frac{7}{9})^2+(\frac{2}{9})^2\} = 0.346$

$G(\text{자식2}) = 1-\{(\frac{1}{8})^2+(\frac{7}{8})^2\} = 0.219$

$* \; G(\text{줄무늬}) = \frac{9}{17} \times G(\text{자식1}) + \frac{8}{17} \times G(\text{자식2})$

$= \frac{9}{17} \times 0.346 + \frac{8}{17} \times 0.219 = 0.286$

꼭지모양 = 말림
(false = 약간 말림, 곧음)



5/8 = 62.5%　　6/9 = 66.7%

$G(\text{자식1}) = 1-\{(\frac{5}{8})^2+(\frac{3}{8})^2\} = 0.469$

$G(\text{자식2}) = 1-\{(\frac{3}{9})^2+(\frac{6}{9})^2\} = 0.444$

$* \; G(\text{꼭지모양}) = \frac{8}{17} \times G(\text{자식1}) + \frac{9}{17} \times G(\text{자식2})$

$= \frac{8}{17} \times 0.469 + \frac{9}{17} \times 0.444 = 0.456$

- **가중치G:** 자식노드의 샘플 수 ↑ → 영향력↑

$$\text{가중치 } G = \frac{m\text{번째 자식 노드의 표본의 수}}{\text{부모 노드의 표본의 수}}$$

<첫번째 모형>에서 왼쪽 노드의 엔트로피를 계산할 때, 왼쪽 노드의 가중치는 9/17이고, 오른쪽 노드의 가중치는 8/17이다.

- **가정**

★ 속성: 줄무늬 → 지니계수: 0.286　　**성능 ↑**

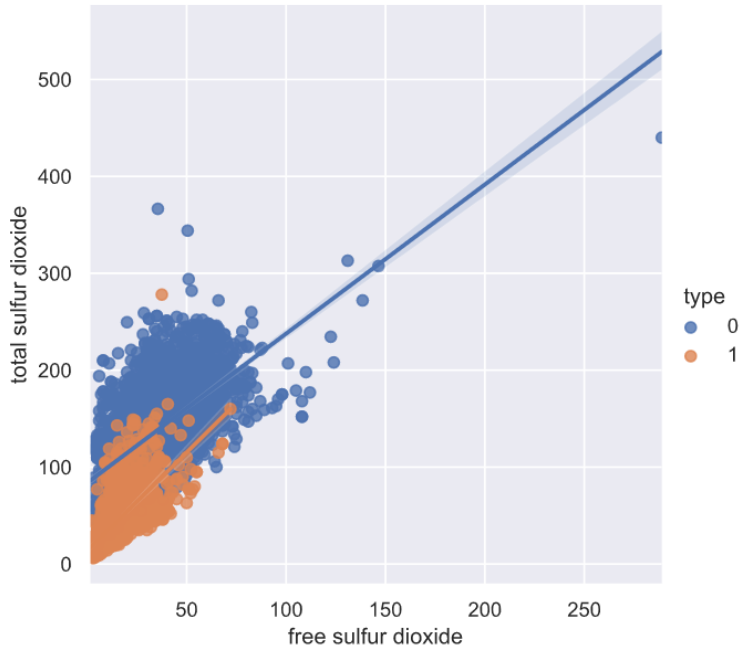
★ 속성: 꼭지모양 → 지니계수: 0.456

4.1.4 기본 알고리즘 : 분할 정보 divede-and-conquer 전략

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	type
0	8.7	0.230	0.32	13.4	0.044	35.0	169.0	0.99975	3.12	0.47	8.8	0
1	5.0	0.270	0.40	1.2	0.076	42.0	124.0	0.99204	3.32	0.47	10.1	0
2	8.0	0.160	0.36	1.5	0.033	14.0	122.0	0.99410	3.20	0.39	10.3	0
3	5.8	0.280	0.35	2.3	0.053	36.0	114.0	0.99240	3.28	0.50	10.2	0
4	7.1	0.140	0.33	1.0	0.104	20.0	54.0	0.99057	3.19	0.64	11.5	0
...	...	...	...	...	...	...	...	...	...	...	...	...
5092	6.0	0.420	0.19	2.0	0.075	22.0	47.0	0.99522	3.39	0.78	10.0	1
5093	6.7	1.040	0.08	2.3	0.067	19.0	32.0	0.99648	3.52	0.57	11.0	1
5094	7.3	0.305	0.39	1.2	0.059	7.0	11.0	0.99331	3.29	0.52	11.5	1
5095	9.0	0.470	0.31	2.7	0.084	24.0	125.0	0.99840	3.31	0.61	9.4	1
5096	13.2	0.460	0.52	2.2	0.071	12.0	35.0	1.00060	3.10	0.56	9.0	1

5097 rows × 12 columns

- 속성 선택 → 불순도가 낮은 분할 값을 선택
  → 분할 값 기준으로 2개의 공간으로 분리 (반복)



이번주 Kaggle 문제 <출처: 지승원님>

4.1.4 기본 알고리즘 : 분할 정보 divede-and-conquer 전략

- **재귀 과정: return 함수(종료 조건)**

1. 해당 노드에 포함된 샘플이 모두 같은 클래스에 속할 경우, 더는 분할을 진행하지 않습니다.
2. 해당 속성 집합이 0일 경우, 혹은 모든 샘플이 모든 속성에서 같은 값을 취할 경우, 더는 분할을 진행할 수 없습니다.
3. 해당 노드가 포함하고 있는 샘플의 집합이 0일 경우, 더는 분할을 진행할 수 없습니다.

※ 참고자료 P.90~91

입력: 훈련 세트 $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$
속성 집합 $A = \{a_1, a_2, \ldots, a_d\}$
과정: 함수 TreeGenerate(D, A)
1: node 생성
2: **if** D의 샘플이 모두 같은 클래스 C에 속하면 **then**
3: 해당 node를 레이블이 C인 터미널 노드로 정한다 **return**
4: **end if**
5: **if** A = ∅ **OR** D의 샘플이 A 속성에 같은 값을 취한다면 **then**
6: 해당 node를 터미널 노드로 정하고, 해당 클래스는 D 샘플 중 가장 많은 샘플의 수가 속한 속성으로 정한다 **return**
7: **end if**
8: A에서 최적의 분할 속성 a_*를 선택한다
9: **for** a_*의 각 값 a_*^v에 대해 다음을 행한다 **do**
10: node에서 하나의 가지를 생성한다. D_v는 D는 a_*^v 속성값을 가지는 샘플의 하위 집합으로 표기한다
11: **if** D_v가 0이면 **then**
12: 해당 가지 node를 터미널 노드로 정하고, 해당 클래스는 D 샘플 중 가장 많은 클래스로 정한다 **return**
13: **else**
14: TreeGenerate(D_v, $A\backslash\{a_*\}$)를 가지 노드로 정한다
15: **end if**
16: **end for**
출력: node를 루트 노드root node로 하는 의사결정 트리

Kaggle 문제 설명