스무고개를 활용한 분류 문제

Classification with Decision Trees

2020학년도 2학기

가슴엔 조국을, 두 눈은 세계로!

Decision Trees 개요

- 배경
  - 아는 정보를 바탕으로 새로운 정보를 평가해야 할 때, 아는 정보를 가장 잘 설명하였던 **순차적인 분류 체계**를 그대로 적용해보자는 아이디어.
- 종류
  - **Classification Trees: 분류 문제 (예: 동부전선을 침투한 인원은 귀순자인가, 간첩인가?)**
  - Regression Trees: 예측 문제 (예: 독재자의 집권 기간은 몇 년인가?)

Classification Trees 개요

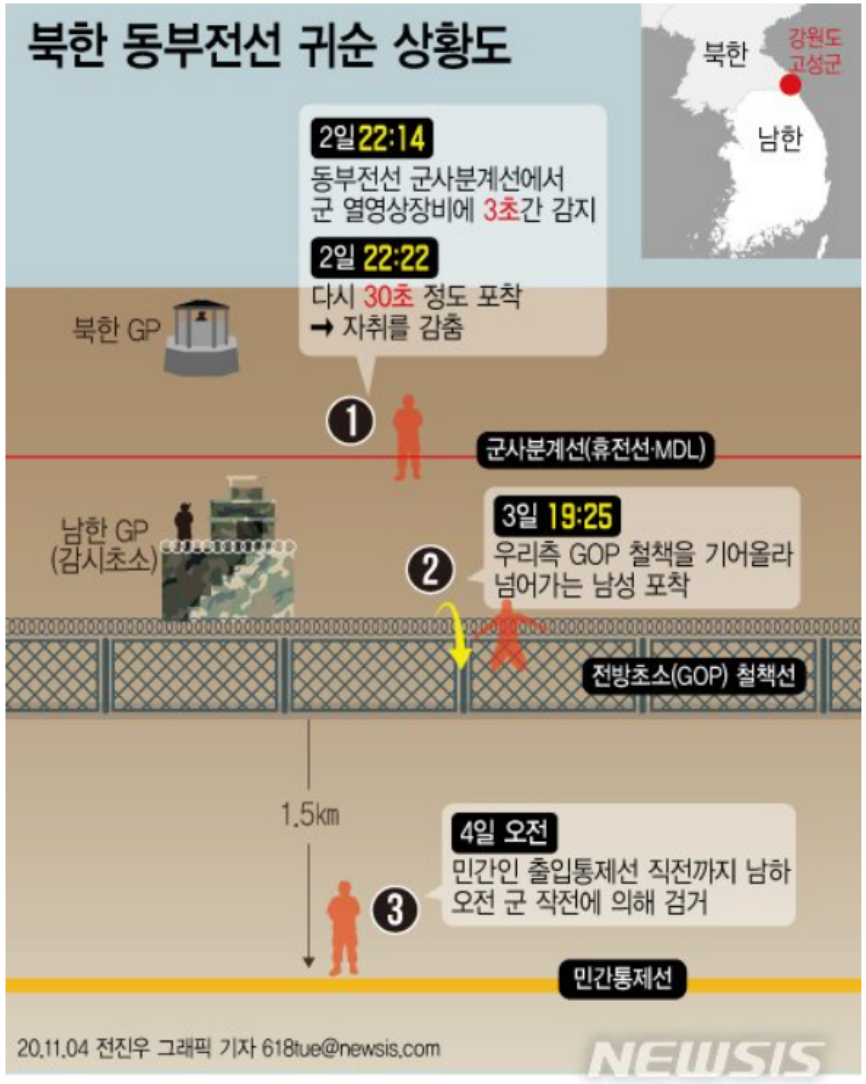
분류 문제 (예: 동부전선을 침투한 인원은 귀순자인가, 간첩인가?)

[1보] 군, 강원도 동부전선에 대침투경계 '진돗개' 발령

송고시간 | 2020-11-04 08:55

김귀근 기자

분류 문제 (예: 동부전선을 침투한 인원은 귀순자인가, 간첩인가?)

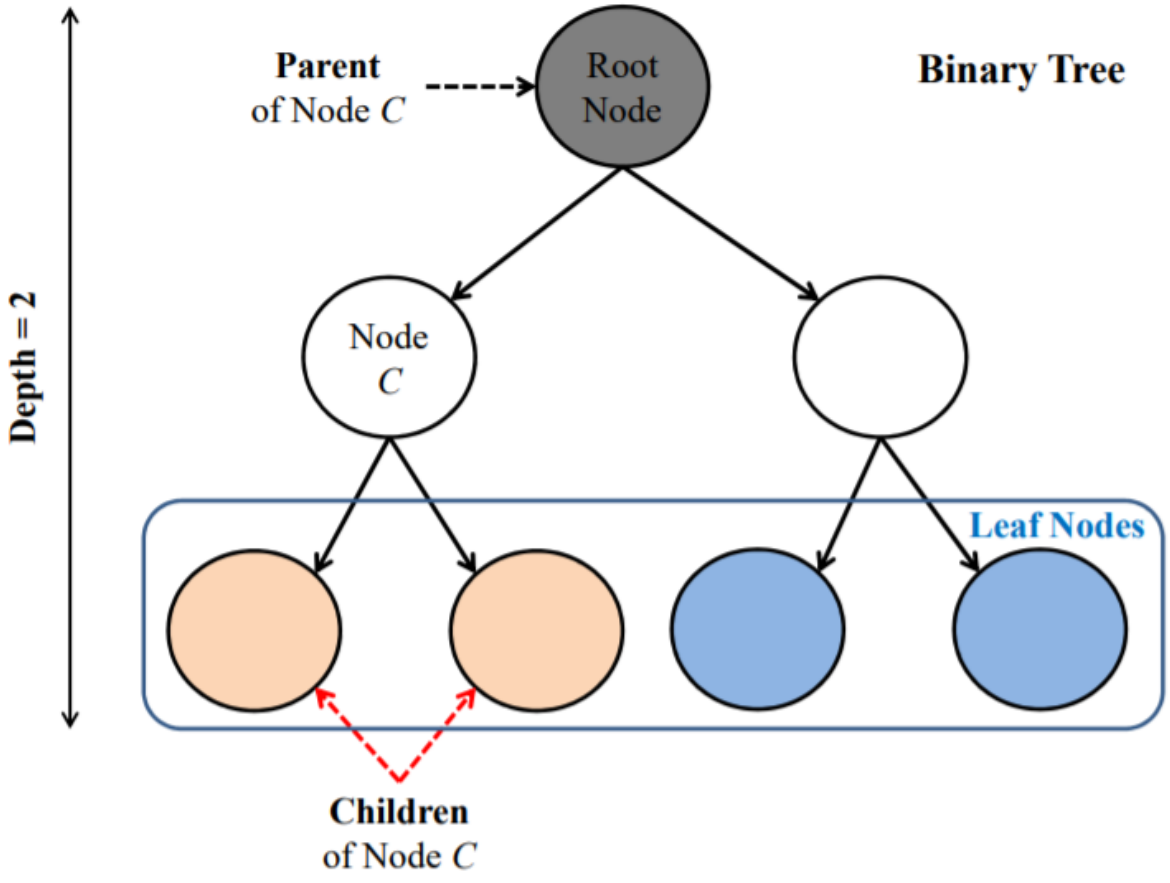


분류 문제 (가상의 데이터)

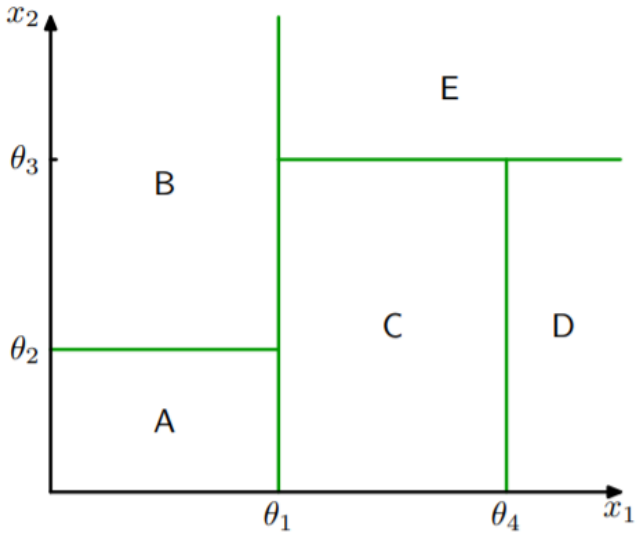
총기 소지 여부	2차 발견 경과 시간	군복 착용 여부	북한군 수색·보고 동향	침투 시각	조사 결과
미소지	3시간	미착용	식별	07:00	귀순자
소지	8시간	미착용	미식별	03:16	공비
미소지	1시간	착용	식별	09:17	귀순자
미소지	20분	착용	식별	08:05	귀순자
미소지	3시간	미착용	식별	15:20	귀순자
미소지	58시간	미착용	미식별	23:00	공비
미소지	5시간	미착용	미식별	21:27	귀순자
미소지	38시간	미착용	미식별	02:16	공비
미소지	2시간	착용	식별	07:27	귀순자
소지	27시간	미착용	미식별	04:15	공비
미소지	10시간	미착용	미식별	08:15	귀순자
미소지	3시간	미착용	미식별	21:35	귀순자
미소지	2시간	착용	미식별	22:36	귀순자
소지	8시간	미착용	미식별	01:08	공비
미소지	6시간	착용	식별	07:04	귀순자
미소지	8시간	미착용	미식별	00:01	귀순자
미소지	1시간	미착용	미식별	05:03	귀순자
미소지	3시간	미착용	미식별	04:00	귀순자
소지	5시간	착용	미식별	06:25	공비
미소지	21시간	미착용	미식별	22:14	?

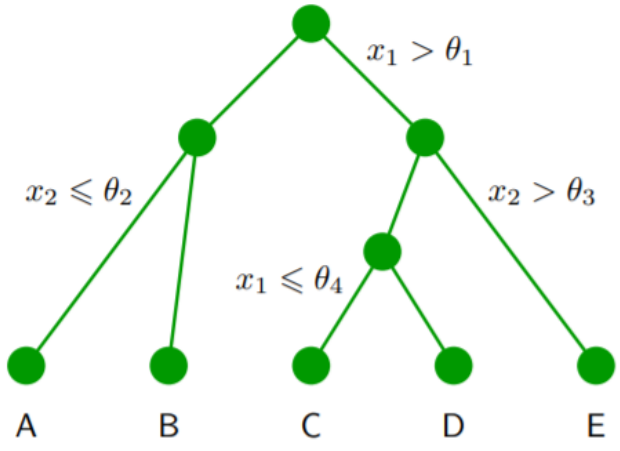
가슴엔 조국을, 두 눈은 세계로!

분류 나무(Classification Trees)의 기본 용어



2개의 연속변수가 있는 분류 나무의 예시

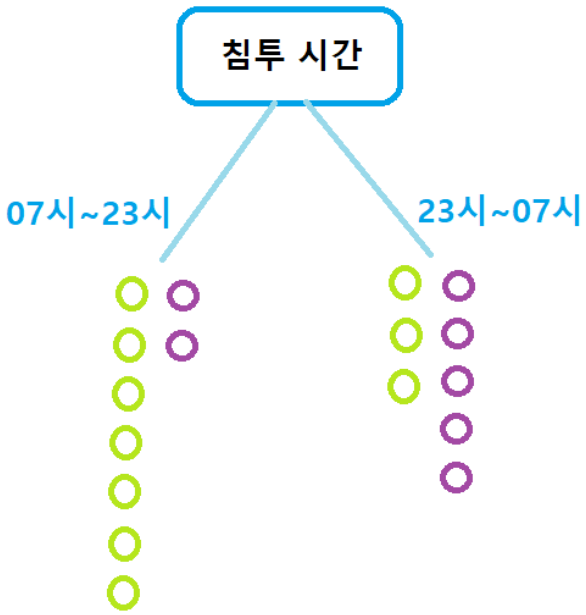


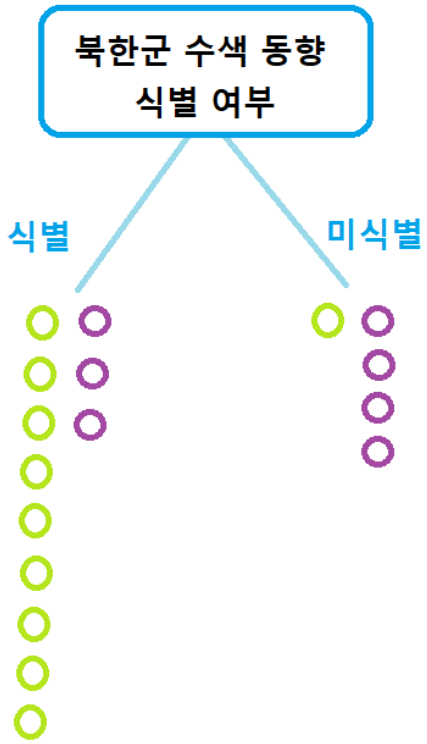


분류 나무의 논점

- 논점 1
  - 어떤 Node부터 위치시켜야 하는가?
- 논점 2
  - 새로운 Node를 추가할지 여부는 어떻게 판단하는가?
- 논점 3
  - 어떻게 가지를 치는 것이 좋은가?

어떤 Node부터 위치시켜야 하는가?





어떤 Node부터 위치시켜야 하는가?

- 연두색은 귀순자(T)를, 보라색은 공비(F)를 나타낸다고 하자.
- '침투 시간'을 A, '북한군 동향'을 B라고 할 때, 각 특성(feature)에 따른 분류 결과는 아래와 같다.
  - **A** [T:F] → (왼쪽 잎) 7:2 / (오른쪽 잎) 3:5
  - **B** [T:F] → (왼쪽 잎) 9:3 / (오른쪽 잎) 1:4
- 이때, A로 먼저 나무를 만드는 것이 좋은가, B로 먼저 분류하는 것이 좋은가?

Gini Impurity

- 일종의 '불순도'를 측정하는 한 가지 방법(Gini impurity)은 아래와 같다.
- 우선, A의 Gini impurity는,
  - 왼쪽 잎: $1 - (\frac{7}{9})^2 - (\frac{2}{9})^2 \approx 0.3456$.
  - 오른쪽 잎: $1 - (\frac{3}{8})^2 - (\frac{5}{8})^2 \approx 0.4687$.
- 여기에서, Gini impurity는 작을수록 좋다.
  - 분류 후 순도가 높으면 Gini impurity는 작아진다.

Gini Impurity

- A의 총 Gini impurity는 왼쪽 잎과 오른쪽 잎의 가중평균으로,
- $(0.3456) \cdot \frac{9}{17} + (0.4687) \cdot \frac{8}{17} \approx 0.4035$ 이다.

Gini Impurity

- 다음으로, B의 Gini impurity는,
  - 왼쪽 잎: $1 - (\frac{9}{12})^2 - (\frac{3}{12})^2 \approx 0.375$.
  - 오른쪽 잎: $1 - (\frac{1}{5})^2 - (\frac{4}{5})^2 \approx 0.32$.

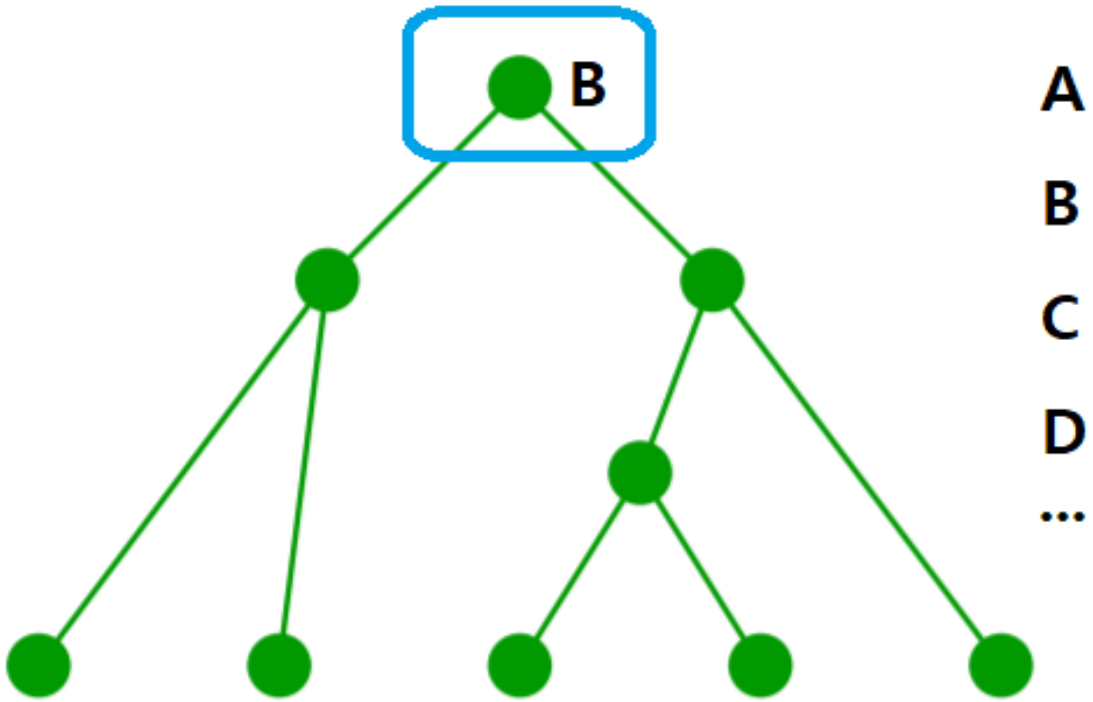
Gini Impurity

- B의 총 Gini impurity는 왼쪽 잎과 오른쪽 잎의 가중평균으로,
- $(0.375) \cdot \frac{12}{17} + (0.32) \cdot \frac{5}{17} \approx 0.3588$ 이다.

어떤 Node부터 위치시켜야 하는가?

- A의 Gini impurity는 0.4035로, B의 Gini impurity는 0.3588로 계산되었으므로,
- 나무를 만들 때, Gini impurity가 더욱 낮아 우수한 분류 기준인 **B를** 먼저 위치시킨다.
- A, B, C, D 등 여러가지 분류 기준이 있을 때는, 그 중에서 가장 Gini impurity가 낮은, 고가치 정보로 먼저 분류하는 나무를 만든다.

어떤 Node부터 위치시켜야 하는가?



분류 나무의 논점 (Reprise)

- 논점 1
  - 어떤 Node부터 위치시켜야 하는가?
- 논점 2
  - 새로운 Node를 추가할지 여부는 어떻게 판단하는가?
- 논점 3
  - 어떻게 가지를 치는 것이 좋은가?

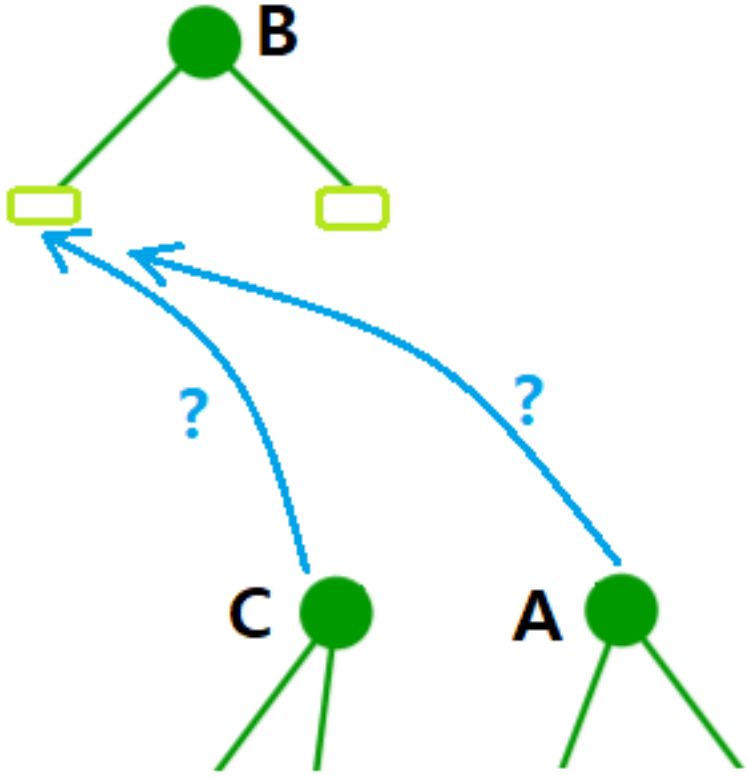
새로운 Node를 추가할지 여부는 어떻게 판단하는가?

- B를 먼저 둔다고 하자.
- 그렇다면 B의 왼쪽 잎과 오른쪽 잎을 다시 node로 전환하여 A, C, D 등 다른 분류 기준을 가져와서 가지치기를 더 해나갈지 판단해보아야 한다.

새로운 Node를 추가할지 여부는 어떻게 판단하는가?

- **B의 왼쪽 잎을 node로 전환할지부터 살펴보자.**
- 우선 앞서 계산한 바와 같이, B의 왼쪽 잎에서 Gini impurity는 0.375였다.

새로운 Node를 추가할지 여부는 어떻게 판단하는가?



새로운 Node를 추가할지 여부는 어떻게 판단하는가?

- 그렇다면 **B의 왼쪽 잎을 출발점으로 삼을 때,**
- 분류 기준 A를 가져왔을 때 A의 왼쪽 잎과 오른쪽 잎의 Gini impurity를 계산하고,
- 다시 가중평균을 하여, A의 총 Gini impurity 값을 구할 수 있다.

새로운 Node를 추가할지 여부는 어떻게 판단하는가?

- 이때, A의 총 Gini impurity 값이 0.375보다 작다면 가는(go) 것이고,
- 0.375보다 크다면(혹은 크거나 같다면) 멈추는(stop) 것이다.
- C, D 등 다른 분류 기준에 대해서도 마찬가지로 계산을 하면 된다.
- 복수의 분류 지표에서 A의 값보다 작은 Gini impurity가 계산된다면, 그 중에서 가장 작은 값의 분류 지표를 선택하여 이어 붙이면 된다.

분류 나무의 논점 (Reprise)

- 논점 1
  - 어떤 Node부터 위치시켜야 하는가?
- 논점 2
  - 새로운 Node를 추가할지 여부는 어떻게 판단하는가?
- 논점 3
  - 어떻게 가지를 치는 것이 좋은가?

어떻게 가지를 치는 것이 좋은가?

- 이진(binary) 분류 지표라면 고민할 필요가 없으나,
- 복수(multiple)의 분류결과(class)를 갖거나,
- 연속변수인 경우에는 어디에서 자를지 고민을 하게 된다.

어떻게 가지를 치는 것이 좋은가?

- 결국, 최대한 많은(가능하다면 모든) 경우의 수를 검토하여,
- 그 중에서 Gini impurity가 가장 작도록 만드는 방식으로
- 가지를 치면 된다. (greedy approach라고도 부른다.)

참고사항

- 나무가 만들어지는 과정에서 Gini impurity가 중요한 역할을 한다는 점이 나타나지만,
- 지난 시간에 살펴본 거리(distance)의 정의가 각양각색이었듯이
- 불순도를 측정하는 지표도 매우 다양하다.
  - Tsallis Entropy, Shannon Entropy 등

토론거리: 나무의 깊이(depth)는 클수록 좋을까?