



부산대학교
PUSAN NATIONAL UNIVERSITY

Decision Trees

Splitting the Space

Rules That Cut Data into Regions

BX3600087-140 / DS3600087-140 / AB3600087-140

기계학습개론

2026학년도 2학기

오용경

데이터사이언스전공

yongkyung.oh@pusan.ac.kr

From Rules to Regions

조건문의 연쇄를 특성 공간에서 본다

조건을 따라 범주를 예측하는 의사결정나무

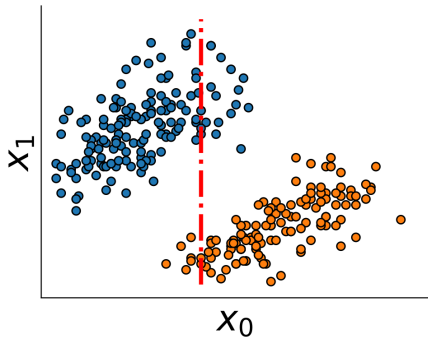
DEFINITION

의사결정나무(decision tree)

특성에 대한 조건으로 관측을 나누고, 도달한 끝점의 예측값을 반환하는 모델

결정트리라고도 부르며, 분류 트리(classification tree)는 끝점에 도달한 학습 관측의 다수 범주를 예측한다

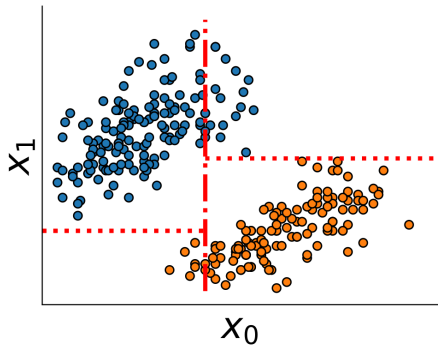
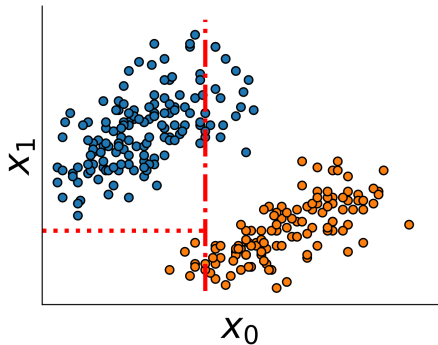
특성 하나와 기준값 하나로 자른다



빨간 점선 하나가 X_0 의 한 기준값에서 자료를 두 구역으로 나눈다

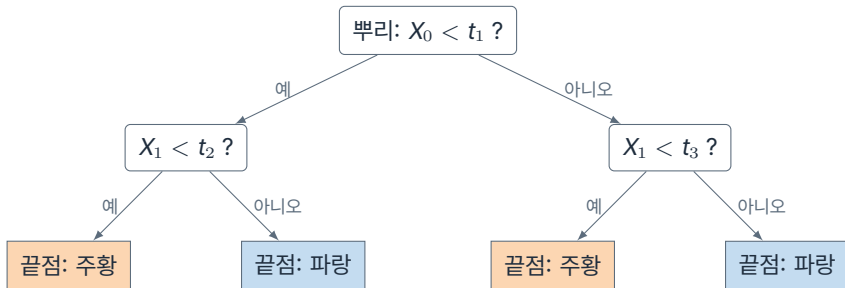
Source: scikit-learn MOOC (2022), Decision Tree Boundaries.

분할을 되풀이하면 직사각형 구역이 생긴다



왼쪽은 두 번, 오른쪽은 세 번 분할한 결과. 새 분할은 앞서 나뉜 한 구역 안에서만 이어진다.

세 번 나눈 구역을 나무로 그리면



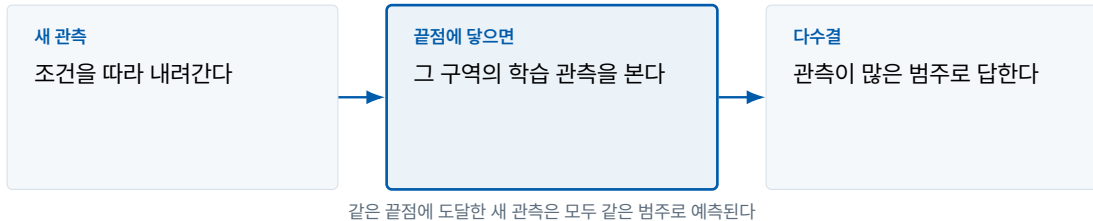
앞 그림 오른쪽(세 번 분할)을 나무로 옮긴 것이다. 세로 점선이 뿌리의 조건 $X_0 < t_1$ 이고, 왼쪽·오른쪽 가로 점선이 두 분기점의 조건이다. 끝점 네 개가 직사각형 구역 네 개에 대응한다

공간의 분할과 나무 구조의 대응

나무의 구성	공간에서의 의미
분기점(internal node)	현재 구역을 둘로 나누는 조건
끝점(terminal node), 잎	더 나누지 않는 구역과 그 예측값
뿌리(root)	모든 학습 관측에서 시작하는 위치
뿌리에서 끝점까지의 경로	한 구역에 도달하기 위한 조건의 연쇄

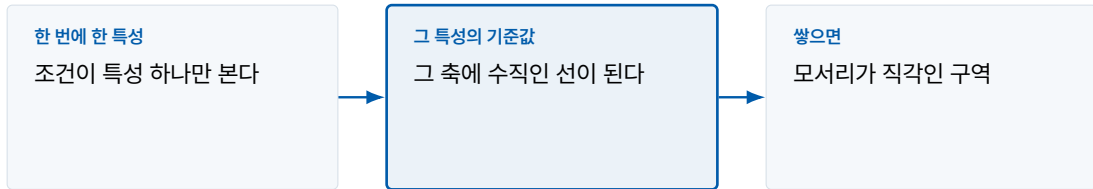
node는 나무의 한 위치로, 분기점과 끝점을 모두 가리킨다

구역 안에서 답하기



Source: James et al. (2023), Section 8.1.

구역이 직사각형인 까닭



한 분할은 특성 하나만 보므로 경계가 각 축에 수직이다

Checkpoint

From Rules to Regions

- 한 번 자를 때 무엇을 고르는가
- 구역 그림의 선 하나는 나무에서 무엇에 해당하는가
- 새 관측의 답은 어떻게 정해지는가

Growing a Tree Top-Down

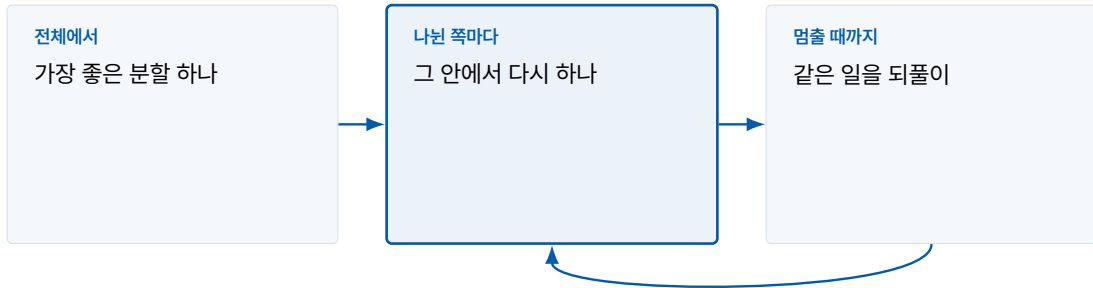
위에서부터 한 번에 하나씩 고른다

분할을 탐욕적으로 선택한다

탐색 방식	계산과 보장
가능한 전체 나무 비교	특성과 분할이 늘면 후보 구조가 매우 많아짐
현재 구역에서 분할 하나 선택	계산을 줄이지만 전체 최적 나무를 보장하지 않음

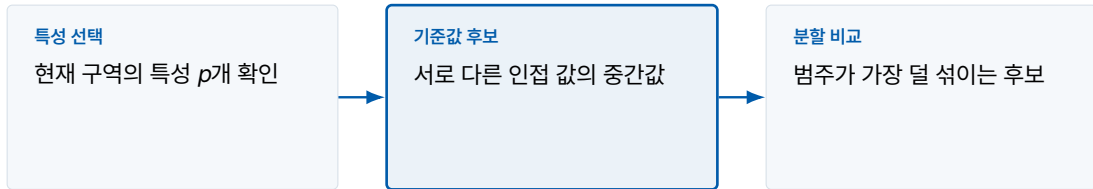
현재 구역에서 분할을 하나씩 골라 되풀이하는 방식을 재귀적 이진 분할(recursive binary splitting)이라 한다

위에서부터 분할을 되풀이한다



Source: James et al. (2023), Section 8.1.1.

수치 특성의 분할 후보를 비교한다



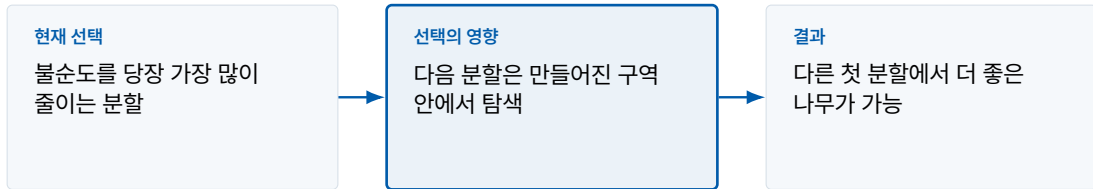
특성마다 서로 다른 값의 개수가 달라 후보 기준값의 수도 다를 수 있다

범주형 특성의 처리는 구현을 확인한다

특성	가능한 분할 방식	주의
수치형	기준값보다 작은지 비교	값 사이의 순서 사용
범주형	범주를 두 집합으로 구분	이를 직접 지원하는 구현에서 사용

Caution scikit-learn의 의사결정나무는 문자열 범주를 그대로 받지 않으므로 수치로 부호화(encoding)한 뒤 입력하며, 이때 분할은 부호화한 수치의 기준값 비교가 된다

현재의 최선이 전체의 최선은 아니다



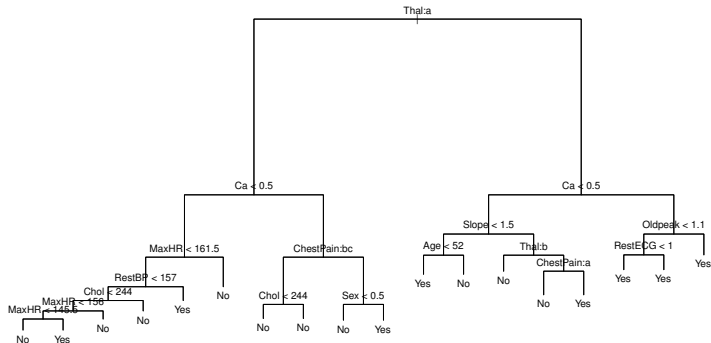
Caution 이후 분할의 모든 조합을 비교하지 않으므로 전체 최적 나무를 보장하지 않는다

나무의 크기를 제한하는 설정

설정	의미	값을 크게 하면
<code>max_depth</code>	뿌리부터 허용하는 최대 분할 횟수	더 깊은 나무를 허용
<code>min_samples_leaf</code>	분할 후 양쪽 끝점에 남아야 할 최소 관측 수	작은 구역을 제한
최소 불순도 감소량	충분히 개선되는 분할만 허용	더 적은 분할만 허용

이 설정은 학습 자료에서 계수를 구하듯 정하지 않고 검증 성능을 비교해 선택한다

멈추지 않은 나무



Heart 자료의 가지치기 전 나무다. 조건 글자를 하나씩 읽기보다 분기가 많고 가지마다 깊이가 다른 구조의 복잡함을 본다. Thal:a는 범주 a와 나머지 범주로 나누는 분할이다

Source: James et al. (2023), Figure 8.6.

분류 트리를 학습하고 그린다

criterion은 다음 절에서 정의할 불순도 기준, max_depth는 멈춤 규칙에 대응한다

```
1 from sklearn.tree import DecisionTreeClassifier, plot_tree
2
3 tree = DecisionTreeClassifier(
4     criterion="gini", max_depth=3, random_state=0
5 )
6 tree.fit(X_train, y_train)
7 plot_tree(tree, feature_names=X_train.columns)
```

Adapted from James et al. (2023), Ch. 8 lab, and scikit-learn MOOC (2022).

Checkpoint

Growing a Tree Top-Down

- 가능한 나무를 모두 만들어 비교하지 않는 까닭은 무엇인가
- 한 번에 하나씩 고르는 방식의 대가는 무엇인가
- `criterion`과 `max_depth`는 무엇을 정하는가

Split Criteria

분할 전후의 불순도를 계산한다

분할 후 범주 구성이 얼마나 순수해지는지 비교한다

부모: 두 범주가 반반	왼쪽	오른쪽
구성이 그대로인 분할	반반	반반
범주를 구분하는 분할	대부분 범주 1	대부분 범주 0

한 구역 안에 범주가 섞인 정도를 불순도(impurity)라 한다. 양쪽의 불순도를 관측 수로 가중하여 비교한다

불순도를 계산하는 두 기준

$$G = \sum_{k=1}^K \hat{p}_k(1 - \hat{p}_k), \quad D = - \sum_{k=1}^K \hat{p}_k \ln \hat{p}_k$$

기준	값의 의미
지니 지수(Gini index)	한 범주만 있으면 0
엔트로피(entropy)	한 범주만 있으면 0, 균등하게 섞이면 최대

$\hat{p}_k$ 는 현재 구역의 범주 k 비율이며, 로그는 자연로그, $0 \ln 0$ 은 0으로 둔다. scikit-learn의 `criterion="entropy"`는 밑이 2인 로그라 값이 $1 / \ln 2 \approx 1.44$ 배지만, 모든 후보에 같은 배수가 곱해지므로 같은 분할을 고른다

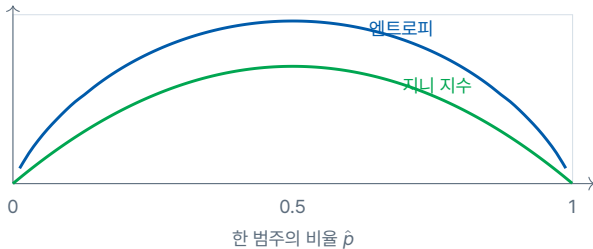
Source: James et al. (2023), Section 8.1.2.

두 식이 주는 값

한 구역의 구성	$\hat{p}$	지니	엔트로피
전부 한 범주	1.00	0.00	0.00
9 대 1	0.90	0.18	0.33
반반	0.50	0.50	0.69

두 척도는 모두 한 범주만 있을 때 0이지만 후보 분할의 순위가 언제나 같은 것은 아니다

두 식의 모양



두 곡선 모두 양 끝에서 0이고 비율 0.5에서 최대. 지니 지수의 최대값은 0.5, 엔트로피는 $\ln 2 \approx 0.693$ 이다. scikit-learn의 엔트로피는 밑이 2인 로그를 쓰므로 반반일 때 최대값 1.0을 출력한다

같은 범주를 예측해도 불순도는 달라진다

범주 1의 비율 $\hat{p}$	예측 범주	정확도	지니
0.90	1	0.90	0.18
0.70	1	0.70	0.42
0.51	1	0.51	0.4998

다수 범주는 그대로지만 정확도와 불순도는 모두 변한다. 분할 후보의 정확도가 같아지는 경우는 별도로 비교해야 한다

분할 하나를 재 보기

두 범주가 50개씩 섞인 구역을 두 후보로 나눠 본다

후보	왼쪽	오른쪽	나눈 뒤 지니·엔트로피
후보 A	40대 10	10대 40	$0.32 \cdot 0.50$
후보 B	30대 25	20대 25	$0.495 \cdot 0.69$

나눈 뒤 값은 양쪽 값을 관측 수로 가중평균한 것이다. 예: B의 지니 $0.55 \times 0.496 + 0.45 \times 0.494 \approx 0.495$

나누기 전 지니가 0.50, 엔트로피가 0.69이므로 후보 A가 훨씬 많이 줄인다

분할 전후의 가중 불순도 감소

$$\Delta G = G_{\text{parent}} - \left[\frac{n_L}{n} G_L + \frac{n_R}{n} G_R \right], \quad n = n_L + n_R$$

부모 구역의 불순도에서 두 자식 구역 불순도의 관측 수 가중평균을 뺀다

엔트로피로 같은 감소량을 계산한 값이 정보이득(information gain)이다. 앞의 예에서 A는 0.19, B는 0.005다

오분류 수가 같은 두 분할 후보

구성은 범주 1 : 범주 0의 관측 수이며 분할마다 총 800개로 같다

후보	왼쪽 / 오른쪽 구성	오분류 수	가중 지니
분할 전	400:400	400	0.500
A	300:100 / 100:300	200	0.375
B	200:400 / 200:0	200	0.333

Heart 나무의 RestECG<1 분할도 양쪽 끝점의 예측이 모두 Yes지만, 한쪽이 9개 모두 Yes가 되어 순도가 높아지므로 선택된다

두 후보의 정확도는 같지만 지니 지수는 B에서 더 작다

분할 기준과 평가지표를 구분한다

역할	기준	이유
분할 선택	지니 지수 또는 엔트로피	범주 구성의 변화를 반영
분류 결과 평가	정확도·재현율 등	문제의 평가 목적에 맞춰 선택

코드의 `criterion`은 분할 기준이며, 모델 비교용 평가지표와 반드시 같을 필요는 없다

비율은 그 위치에 도달한 학습 관측으로 센다

자리	환자	정상	환자 비율
나누기 전	50	50	0.50
왼쪽	40	10	0.80
오른쪽	10	40	0.20

왼쪽과 오른쪽 비율은 각각 그쪽에 도달한 학습 관측 50개로 센다

Checkpoint

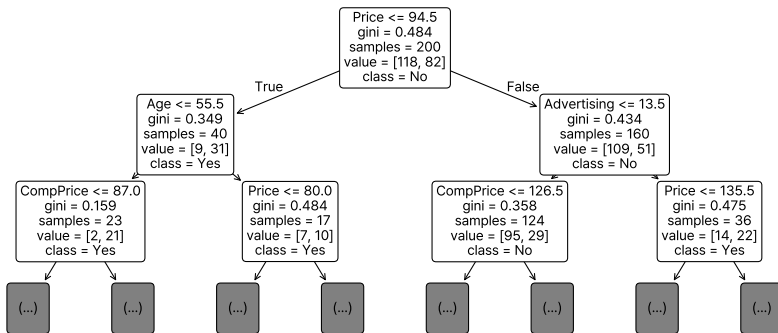
Split Criteria

- 두 범주가 반반 섞여 있으면 지니 지수는 얼마인가
- 분할 전후의 불순도 감소량은 어떻게 구하는가
- 정확도로 분할을 고르면 무엇을 놓치는가

Reading and Sizing a Tree

그려진 나무를 읽고 크기를 정하는 설정을 본다

깊이 3 분류 트리를 그린 결과



Carseats 자료(Sales가 8보다 크면 Yes)의 학습 절반 200개로 앞의 코드를 실행한 결과 중 깊이 2까지(위의 세 층)다(아래는 (...)로 줄임). 상자마다 분할 조건, gini, samples(관측 수), value(No·Yes 수), class(예측 범주)가 적힌다

그려진 나무에서 읽을 값

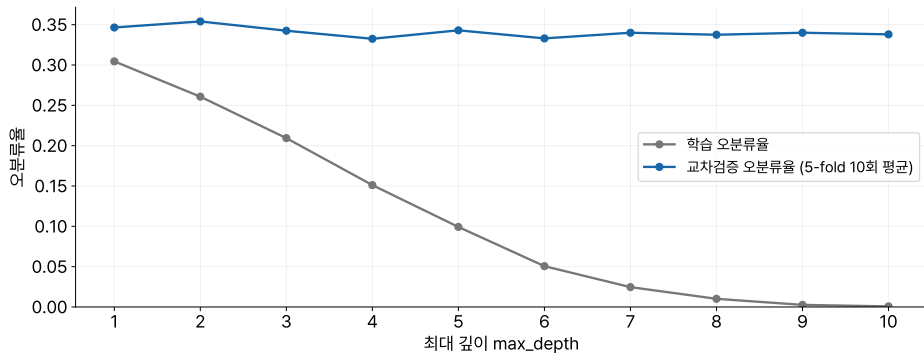
표시	뜻	확인할 점
분할 조건	분기점에서 관측을 나누는 기준	끝점에는 분할 조건이 없음
불순도	그 위치에서 범주가 섞인 정도	학습 관측으로 계산
관측 수	그 위치에 도달한 수	관측이 적으면 추정이 흔들린다
범주별 수 또는 비율	범주 구성과 예측 범주	표시 옵션·가중치 설정 확인

최대 깊이는 표현 가능한 구역 수를 제한한다

최대 깊이	끝점 수의 상한	표현력과 오차
1	2	단순한 분할만 표현
3	8	더 세밀한 구역 표현 가능
d	2^d	학습 오차 감소가 검증 오차 감소를 보장하지 않음

적절한 깊이는 특정 숫자로 고정되지 않으며 검증 성능으로 고른다

깊이에 따른 학습·교차검증 오차



같은 학습 절반 200개에서 5-fold 교차검증을 10번 반복한 평균이다. 학습 오분류율은 깊이 10에서 거의 0이 되지만, 교차검증 오분류율은 0.33에서 0.35 사이에 머물고 fold 사이 표준편차는 약 0.06에서 0.07이다

Adapted from James et al. (2023), Chapter 8 lab, Carseats.

작은 끝점의 예측은 불안정할 수 있다

끝점의 구성	분류 예측	주의
관측 1개	그 관측의 범주	오표기나 특이 관측에 민감
같은 특성값, 다른 범주	다수 범주	특성값이 같으면 더 분할해도 구분할 수 없음

Caution 나무를 깊게 만든다고 모든 관측을 하나씩 분리하거나 학습 오차를 0으로 만들 수 있는 것은 아니다

수치 특성의 단위를 바꿔도 분할은 같다

조건	단위 변환 후 조건	관측의 분할
길이 $x < 2\text{ m}$	길이 $100x < 200\text{ cm}$	같은 관측이 같은 쪽으로 이동

축별 크기 비교를 사용하므로 단순 단위 변환을 위한 표준화는 보통 필요하지 않다

경계의 방향에 따라 필요한 분할이 다르다

자료의 경계

축에 나란한 분할로 표현하면

특성 축에 나란함

적은 분할로 구역을 나눌 수 있다

비스듬한 직선

여러 분할을 조합해 계단 모양으로 근사한다

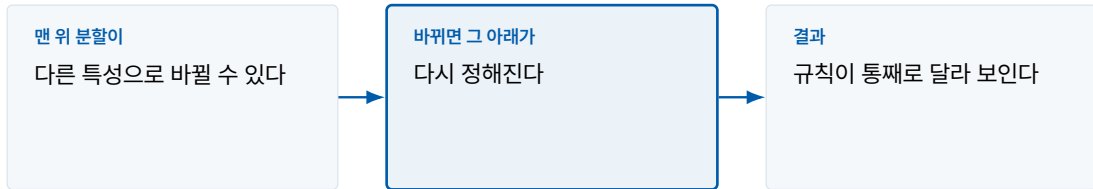
나무의 분할은 한 번에 하나의 특성에 기준값을 적용한다

나무와 선형 분류 모델의 차이

구분	분류 트리	선형 분류 모델
경계	축에 나란한 분할의 조합	직선 또는 초평면
해석	경로의 조건으로 설명	계수와 특성으로 설명
척도	보통 표준화가 필요 없음	정규화(regularization)·최적화 설정에 따라 필요
범주형 입력	구현에 따라 수치 표현 필요	수치 표현으로 변환

경계의 형태뿐 아니라 자료의 양·잡음·검증 성능을 함께 비교한다

자료가 조금 바뀌면



Caution 예측은 비슷해도 규칙이 달라지므로 나무 하나를 설명의 근거로 쓸 때 조심한다

Checkpoint

Reading and Sizing a Tree

- 깊이를 늘릴 때 검증 오차도 반드시 작아지는가
- 척도를 맞추지 않아도 되는 까닭은 무엇인가
- 비스듬한 경계를 나무는 어떻게 흉내 내는가

How Much of the Tree Should You Keep?

Next Lecture: Pruning and Regression Trees

- 앞에서 평균을 예측하는 회귀 트리
- 다 키운 뒤 되돌리는 가지치기
- 규칙과 특성 중요도로 나무 읽기