Foundation of Machine Learning 11주차

정재헌, 우지수 / 2023.03.29

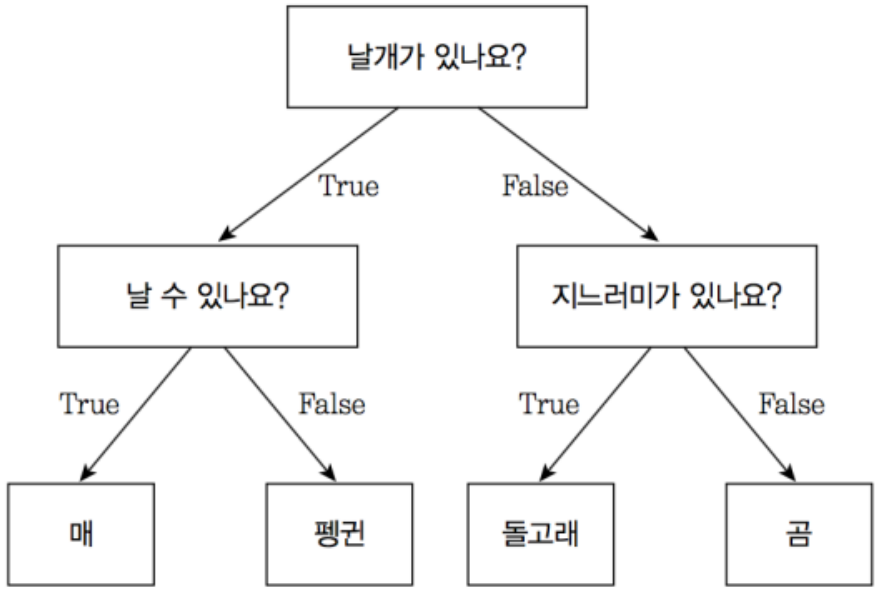
Computational Data Science LAB

CONTENTS

01 Decision Tree

What is Decision Tree

- Decision Tree
  - ✓ 의사결정나무는 입력값에 대한 예측값을 나무 형태로 나타내어주는 모형
  - ✓ Root node, internal node, terminal node
  - ✓ 순도를 최대화 불순도를 최소화하는 방향으로 학습 진행
  - ✓ 장점 : 입력값이 주어졌을 때 설명변수의 영역을 따라가며 출력값이 어떻게 나올 수 있는지 알기 쉬워 모형의 해석이 쉬움
  - ✓ 불순도 측정을 해서 각 class의 데이터들이 얼마나 해당 class에 어울리는지 측정
  - ✓ 연속형 데이터 : 회귀나무, 범주형 데이터 : 분류나무



02 | Regression Tree
What is Regression Tree

- Regression Tree
  - ✓ Data $(x_i, y_i)\,, i = 1, \dots, n$이 주어졌다고 할 때 의사결정 나무는 다음과 같음:

$$f(x) = \sum_{m=1}^{M} c_m I(x \in R_m)$$

  - ✓ 여기서 $R_1, \dots R_M$은 M개로 쪼갠 설명변수들의(겹치지 않는) 영역
  - ✓ 설명변수 벡터x가 R_m에 포함된다면 c_m으로 예측
  - ✓ 이 때 c_m을 추정하기 위해서는 다음의 식을 이용:

$$\widehat{c_m} = ave(y_i \,|x_i \in R_m)$$

02 | Regression Tree
Impurity of Regression Tree

- Impurity
  - ✓ 회귀나무에서의 불순도는 같은 class에 있는 값들이 가급적 비슷한 y값을 갖도록 함
  - ✓ 오차제곱합을 최소화 시킴으로 같은 class에 데이터간 분산을 줄여줌

- c^*
  - ✓ j번째 설명변수 x_j와 분리 기준인 s에 대해 나누어진 두 영역 $R_1(j,s),\ R_2(j,s)$라 한다면:

$$R_1(j,s) = \{x_j : x_j \le s\},\ R_2(j,s) = \{x_j : x_j > s\}$$

- loss function
  - ✓ $\widehat{c_1} = ave(y_i \,|x_i \in R_1(j,s))$
  - ✓ $\widehat{c_2} = ave(y_i \,|x_i \in R_2(j,s))$
  - ✓ $L(j,s) = \sum_{i:x_i \in R_1(j,s)} \left(y_i - \widehat{c_1}(j,s)\right)^2 + \sum_{i:x_i \in R_2(j,s)} \left(y_i - \widehat{c_2}(j,s)\right)^2$

02 | Regression Tree
Finding the Split Point

- Split Point s
  - ✓ 인접한 값 사이의 중간값을 분할 기준으로 선택 따라서 분할 기준 s_j는 다음과 같이 계산:

$$s_j = \frac{1}{2} \cdot \left(x_{\{j(r)\}} + x_{\{j(r+1)\}}\right) for\ r = 1, \dots, n-1$$

  - ✓ 즉, r은 1부터 $n-1$까지 증가하면서 각 분할 기준의 위치를 결정
  - ✓ 이를 통해 $n-1$개의 분할 기준만 검사하면 되므로, 분할 기준을 찾는 데 필요한 계산량이 줄어듬
  - ✓ 분할 기준은 불순도 평가를 진행하며 최적을 분리(loss가 가장 작을 때)를 찾을 때 까지 시도

03 | Classification Tree
What is Classification Tree

- Classification Tree
  - ✓ 회귀나무와 동일하게 불순도의 측도로 성장
  - ✓ 불순도의 측도는 엔트로피 지수, 지니 계수 등
  - ✓ 이는 입력값 $x \in R_m$이 주어졌을 때 R_m에 속하는 y_i의 범주가 가장 많은 범주를 예측값으로 지정
  - ✓ 즉, 예측값을 R_m에서 다수결의 원칙으로 정함

03 | Classification Tree

불순도 측정 – 지니 지수

- 지니지수
  - ✓ m개의 관측치를 포함하는 A에 대한 지니지수:

$$I(A) = 1 - \sum_{k=1}^{m} p_k^2 \quad (p_k : A\text{내에서 클래스 } k\text{에 속하는 관측치의 비율})$$

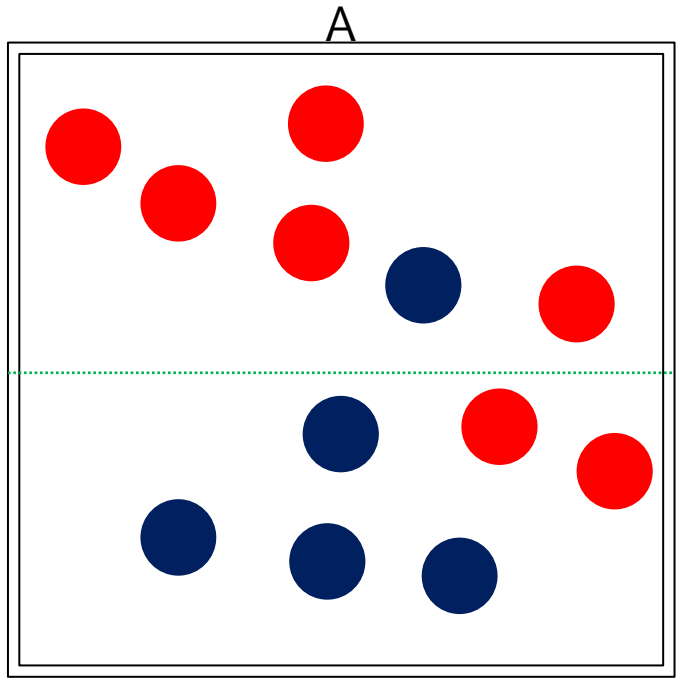
$$I(A) = 1 - \sum_{k=1}^{m} p_k^2 = 1 - \left(\frac{7}{12}\right)^2 - \left(\frac{5}{12}\right)^2 = 0.49$$

  - ✓ m개의 관측치를 포함하는 d개의 class로 나눌 때 지니지수:

$$I(A) = \sum_{1}^{d} \left(r_i \left(1 - \sum_{k=1}^{m} p_k^2 \right) \right) \quad (r_i : \text{전체 관측치 중에서 } i\text{번째 } class\text{에 존재하는 관측치의 비율})$$

$$I(A) = \sum_{1}^{d} \left(r_i \left(1 - \sum_{k=1}^{m} p_k^2 \right) \right)$$

$$= \frac{6}{12}\left(1 - \left(\frac{5}{6}\right)^2 - \left(\frac{1}{6}\right)\right) + \frac{6}{12}\left(1 - \left(\frac{2}{6}\right)^2 - \left(\frac{4}{6}\right)^2\right) = 0.36$$



분할을 통해 얻어지는 지니지수 감소량 (information gain) $\Delta G = 0.49 - 0.36 = 0.13$

03 | Classification Tree

불순도 측정 – 엔트로피 계수

- 엔트로피 계수

✓ m개의 관측치를 포함하는 A에 대한 엔트로피:

$$Entropy(A) = -\sum_{k=1}^{m} p_k \log_2(p_k) \quad (p_k : A\text{내에서 클래스 } k\text{에 속하는 관측치의 비율})$$

$$Entropy(A) = -\sum_{k=1}^{m} p_k \log_2(p_k) = -\left(\frac{7}{12}\right)\log_2\left(\frac{7}{12}\right) - \left(\frac{5}{12}\right)\log_2\left(\frac{5}{12}\right) = 0.98$$

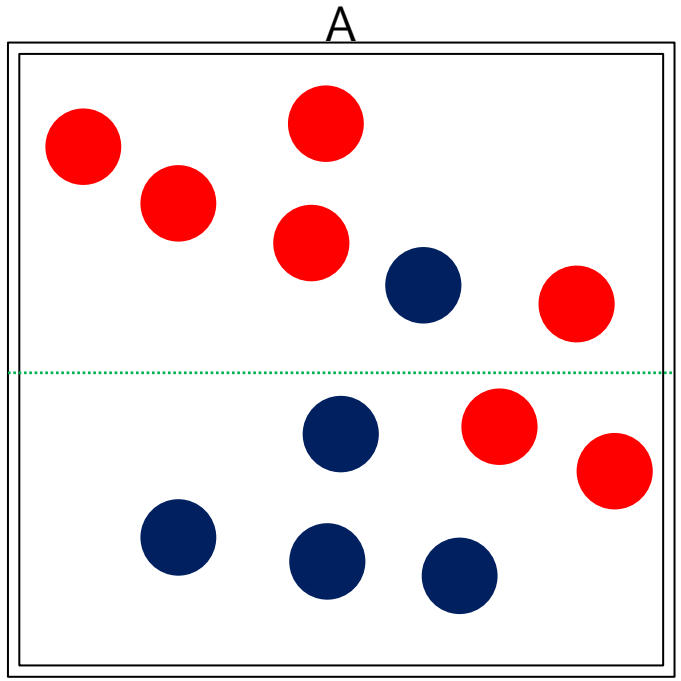
✓ m개의 관측치를 포함하는 d개의 class로 나눌 때 엔트로피:

$$Entropy(A) = \sum_{1}^{d}\left(r_i\left(-\sum_{k=1}^{m} p_k \log_2(p_k)\right)\right) \quad (r_i\text{: 전체 관측치 중에서 } i\text{번째 } class\text{에 존재하는 관측치의 비율})$$

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

$$= \frac{6}{12}\left(-\left(\frac{5}{6}\right)\log_2\left(\frac{5}{6}\right) - \left(\frac{1}{6}\right)\log_2\left(\frac{1}{6}\right)\right)$$

$$+ \frac{6}{12}\left(-\left(\frac{2}{6}\right)\log_2\left(\frac{2}{6}\right) - \left(\frac{4}{6}\right)\log_2\left(\frac{4}{6}\right)\right) = 0.78$$



분할을 통해 얻어지는 엔트로피 감소량(information gain) $\Delta G = 0.98 - 0.78 = 0.20$

04 | Tree Pruning
Pruning

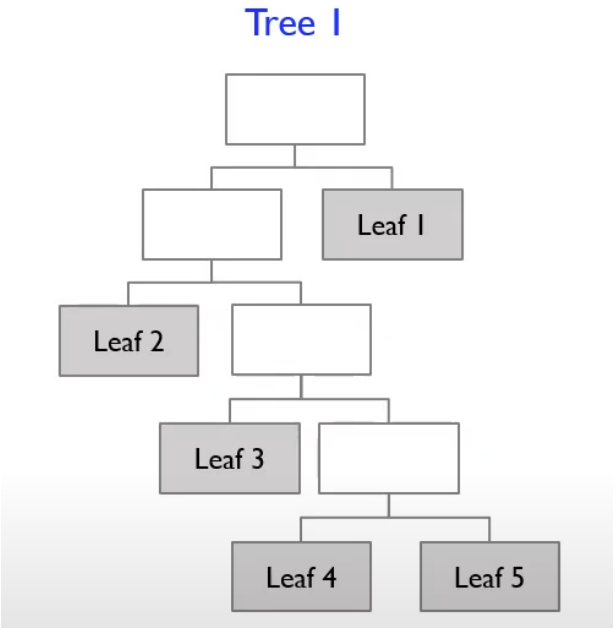
- 사전 Pruning
  - ✓ 트리 생성 과정에서 노드 분할을 결정하기 전에 미리 가지치기 규칙을 설정해줌으로써 트리의 깊이를 제한하는 방법
  - ✓ 예를 들어, 최대 깊이(max_depth)나 최소 노드 크기(min_samples_leaf)를 미리 정해두고, 해당 기준에 도달하면 노드 분할을 중지
- 사후 Pruning
  - ✓ 먼저 완전한 트리를 만든 후, 트리를 가지치기 하면서 성능을 향상시키는 방법
  - ✓ Cost Complexity Pruning(CCP)를 사용해 복잡도와 오분류율(불순도)을 고려, 이를 최소화

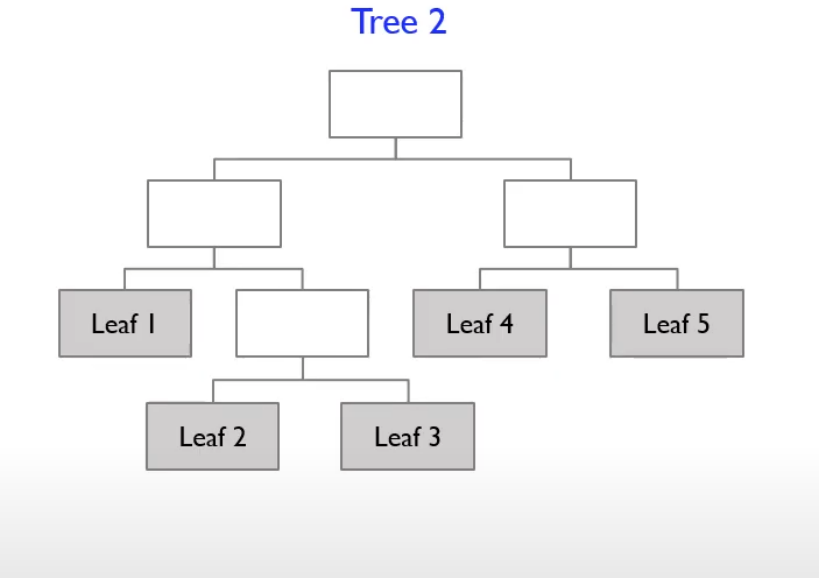
04 | Tree Pruning
Cost Complexity Pruning(CCP)

- Cost Complexity Pruning(CCP)
  - ✓ $C_\alpha(T) = \hat{R}(T) + \alpha \mid T \mid$
  - ✓ $\hat{R}(T)$: T에서의 Empirical risk
  - ✓ α : hyperparameter로 모델의 complexity를 제어하는 역할, 작을수록 모델은 더 복잡해지고 클수록 단순해짐
  - ✓ $\mid T \mid$: terminal node의 개수로서 이 값이 크면 클 수록 모델은 복잡해짐
  - ✓ $C_\alpha(T)$를 최소화해주는 방향으로 가지치기를 진행

04 | Tree Pruning

Pruning



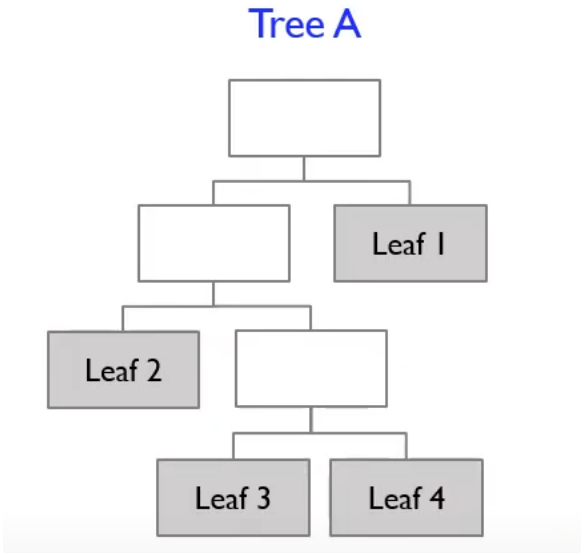


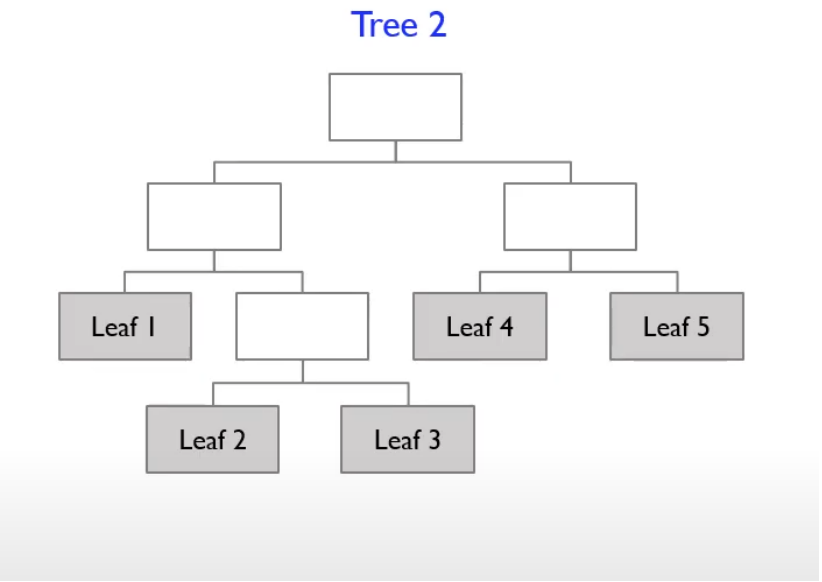
$\hat{R}(T_1) < \hat{R}(T_2)$

- ✓ Leaf node가 5개인 경우 가지치기를 했을 때 Empirical Rick가 적음
- ✓ 가지치기를 했을 때의 장점

04 | Tree Pruning

Pruning



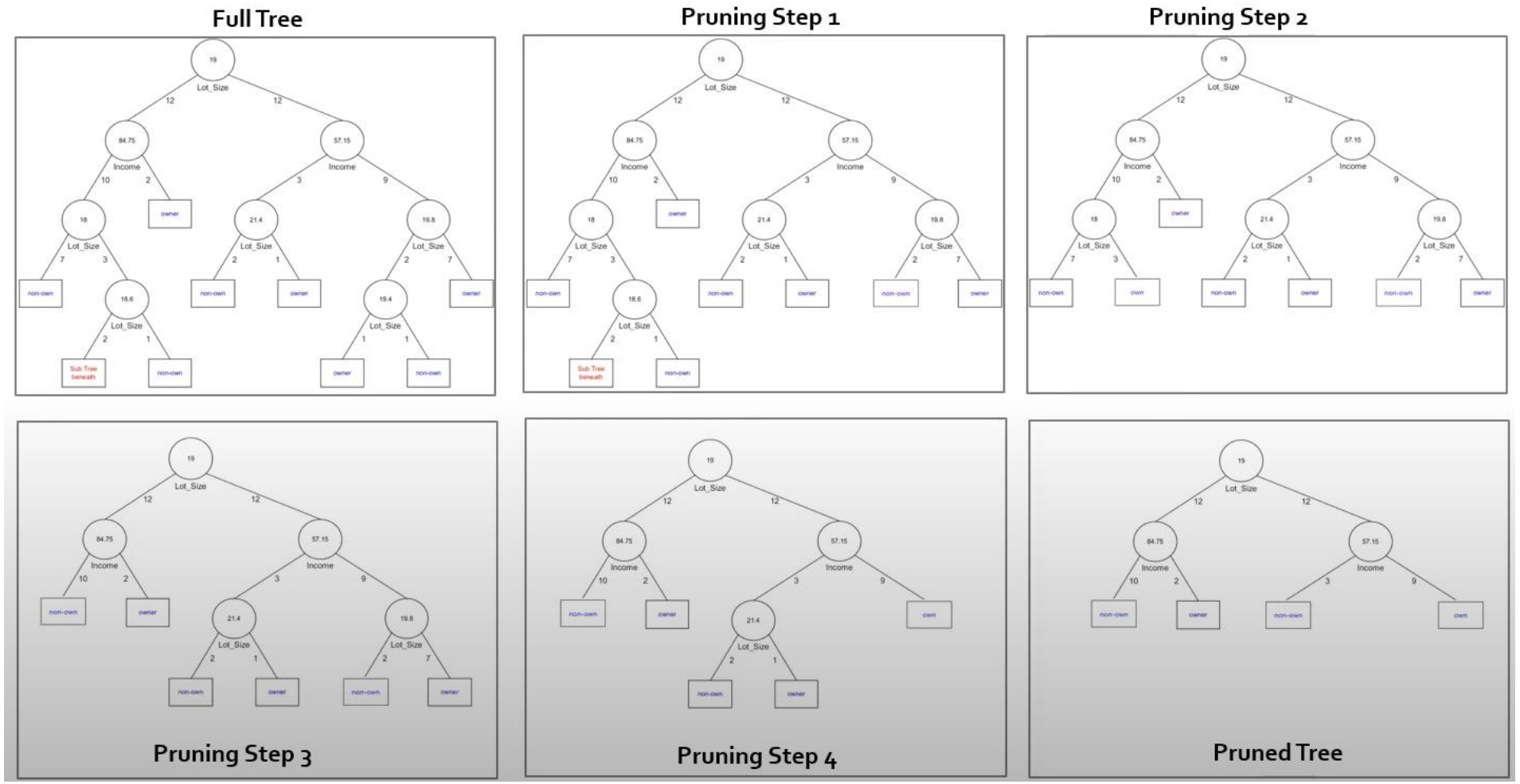


$\hat{R}(T_1)$ $=$ $\hat{R}(T_2)$

- ✓ 가지치기를 했을 때 Empirical Rick가 같고 Leaf node가 다른 경우
- ✓ Complexity가 줄어드는 장점

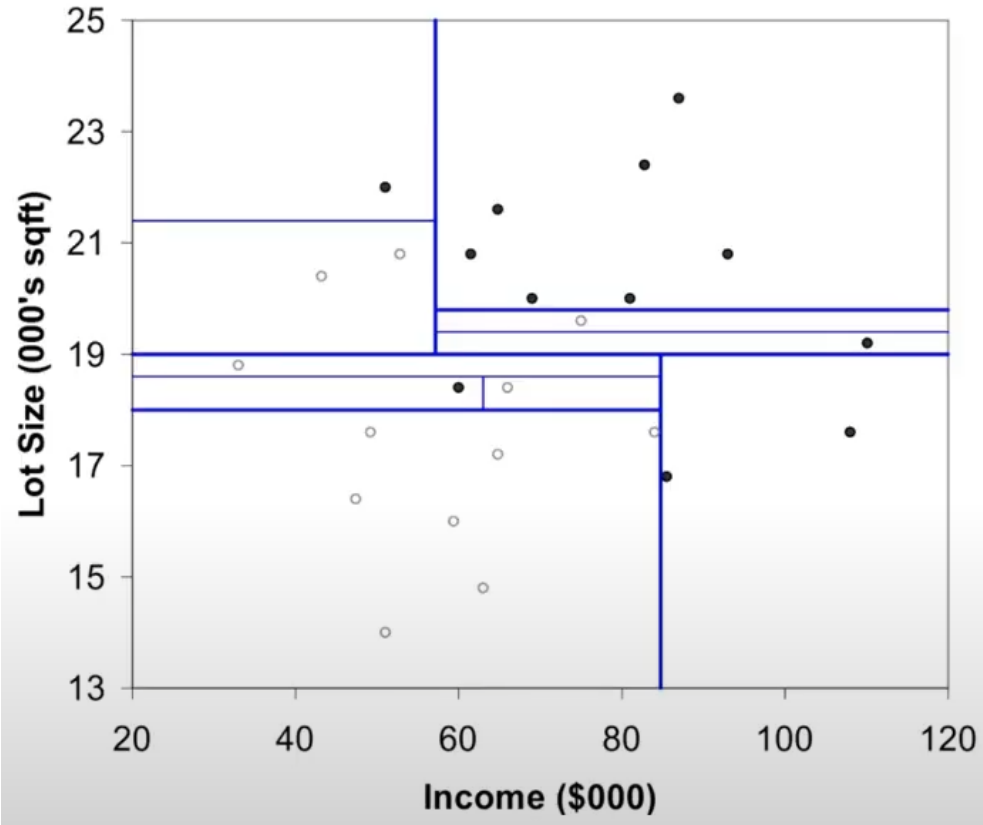
04 | Tree Pruning

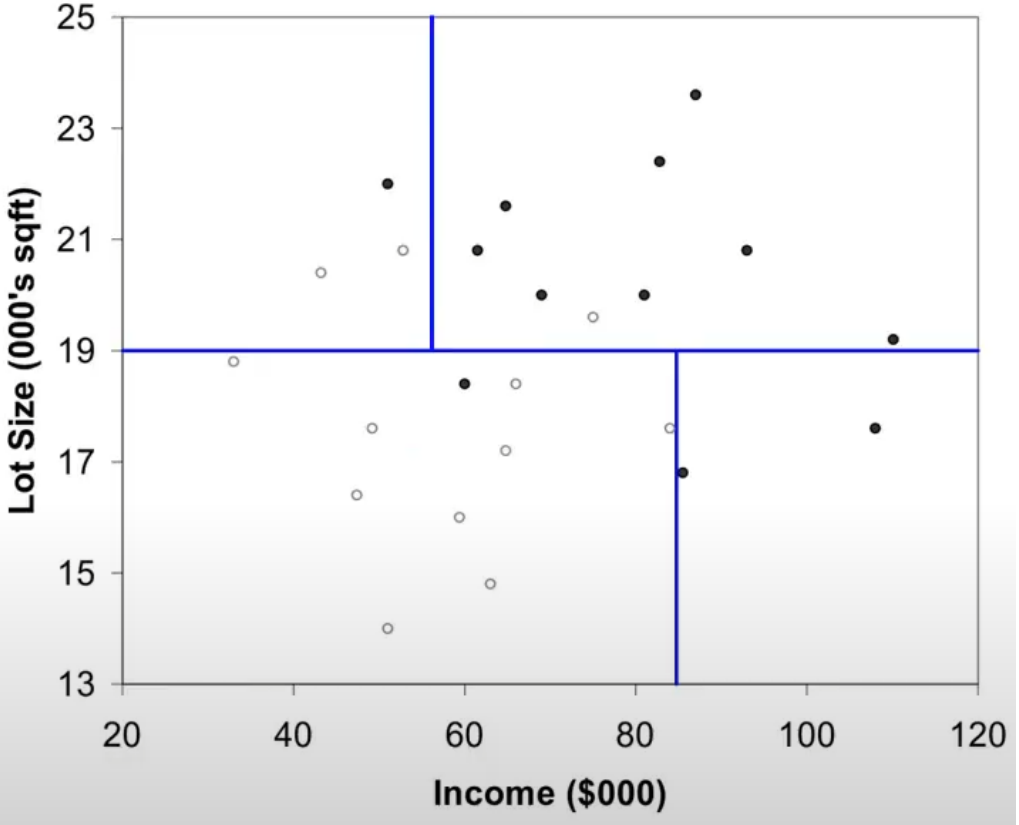
Pruning



04 | Tree Pruning

Pruning





05 | Basic Statistics and a Bit of Bootstrap

Bias and Variance

- Parameters
  - ✓ 확률 분포 P가 있다고 가정했을 때 우리는 P의 모수를 추정하길 원함 ex) 평균, 분산, 등
  - ✓ 이때의 모수를 P의 파라미터라고 함, 확률 분포 P의 함수인 μ는 모든 파라미터를 나타낼 수 있음
  - ✓ Q. 해당 파라미터 μ는 랜덤인가?
  - ✓ A. 실수 R 상에서 밀도 함수 f(x)를 가진 확률 변수 P가 있다면 평균은 아래와 같이 정의:

$$\mu = \int_{-\infty}^{\infty} x f(x) dx$$

  - ✓ 따라서 이는 그저 적분 값이며 아무런 랜덤성이 없음
  - ✓ 평균은 하나의 값으로 고정되어있어 확률 변수의 분포와 밀도 함수를 통해 계산
  - ✓ 평균은 확률 변수와 랜덤한 값이 아니며 그 자체로 확률 변수의 특성을 설명하는 지표

05 | Basic Statistics and a Bit of Bootstrap

Bias and Variance

- Statistics and Estimators
  - ✓ 실수 D가 P의 샘플(표본)이라고 가정, s는 P의 샘플인 D의 함수
    - 통계량 : 표본의 함수
    - 점 추정량 : 어떤 모집단에서 표본을 추출하여 표본의 통계량을 이용하여 모수를 추정
  - ✓ 주어진 데이터에 의해 결정되므로 데이터에 따라 변경
  - ✓ 통계량은 랜덤이기 때문에 확률 분포를 가짐 $=$ 표본 분포
    - 표본분포 : 주어진 모집단으로부터 여러 개의 표본을 추출하고 추출된 표본들을 이용하여 통계량의 분포를 구하는 것
      표본이 커질수록 통계량은 모집단으로 집중
    - 표본분포의 특성을 파악하여 모집단의 특성을 추론 가능
  - ✓ $standard\ error$ (표준오차):
    - 표본분포의 표준편차를 의미
    - 표본의 크기에 따라 변하지만, 표본 데이터에 의존하지 않음
    - 파라미터로서 고정된 값을 가지며 랜덤한 값이 아님
  - ✓ $real\ valued\ parameter$: 실수 값을 가지는 파라미터, 모집단의 특성을 나타내는 값 (μ)
  - ✓ 추정하기 위해선 표본을 이용하여 통계량을 계산

05 | Basic Statistics and a Bit of Bootstrap

Bias and Variance

- Bias and Variance
  - ✓ 편향$(bias)$ → 추정치가 참값으로부터 멀리 떨어져 있는 것

$$Bias(\hat{\mu}) = \mathbb{E}\hat{\mu} - \mu$$

  - ✓ 분산$(Variance)$ → 추정치가 얼마나 가변적인지를 나타내는 척도

$$Var(\hat{\mu}) = \mathbb{E}\hat{\mu}^2 - (\mathbb{E}\hat{\mu})^2$$

    - 분산이 크다면 해당 추정치는 흩어져있는 것이고 분산이 작으면 더 정확한 추정치로 평가
  - ✓ unbiased는 편향이 없기 때문에 참값에 가까운 추정치를 얻을 수 있지만, 분산이 크다면 추정치가 흩어져서 정확한 추정치를 얻기 어려움
  - ✓ 따라서 추정치를 선택할 때에는 편향과 분산을 모두 고려하여 최적의 추정치를 선택해야함
  - ✓ → 따라서 추정을 할 때 'error bars'를 넣음
    - error bars : 표준오차 혹은 신뢰구간으로 나타냄
    - 이걸 넣음으로써 추정치의 불확실성을 나타낼 수 있음
    - 이를 통해 해당 추정치의 신뢰도를 높일 수 있음

$$\hat{\mu} \pm \sqrt{(\widehat{Var\hat{\mu}})}$$

Ex)

- 예를 들어, 추정치의 표준오차가 0.5이고, 해당 추정치가 10이라면
- 95% 신뢰구간은 $10 \pm 1.96 \times 0.5 = (9.01, 10.99)$
- 이를 통해 추정치가 참값에서 얼마나 떨어져 있을 가능성이 있는지를 나타낼 수 있습니다.

05 | Basic Statistics and a Bit of Bootstrap

Bootstrap

- The Bootstrap Sample
  - ✓ 원래 데이터 셋 D로부터 복원 추출을 통해 크기가 n인 샘플을 만드는 것
  - ✓ 이를 통해 샘플링 분포에 대한 정보를 얻고 통계적 추론 가능

    ex) $D = x_1, x_2, x_3, x_4$

    Bootstrap sampling을 통해 크기가 4인 샘플 만들기

    첫 번째 : x_1 / 두 번째 : x_3 / 세 번째 : x_1 / 네 번째 : x_2

    크기가 4인 샘플 : (x_1, x_3, x_1, x_2)

- Bootstrap Sampling을 통해 각 관측치들이 선택될 확률
  - ✓ 복원추출을 통해 D에서 크기가 n인 표본을 생성
  - ✓ 이 때, D에서 관측치 X_i가 선택될 확률은 $\left(1-\frac{1}{n}\right)^n$ 과 동일 $\left(1-\frac{1}{n}\right)^n \approx \frac{1}{e} \approx .368.$
  - ✓ n이 충분히 큰 경우 $\left(1-\frac{1}{n}\right)^n$ 은 $\frac{1}{e}$와 근사적으로 같아지면서 각 관측치가 선택되지않을 확률은 약 0.368
  - ✓ 따라서 각 관측치가 적어도 한 번 선택될 확률은 63.2%

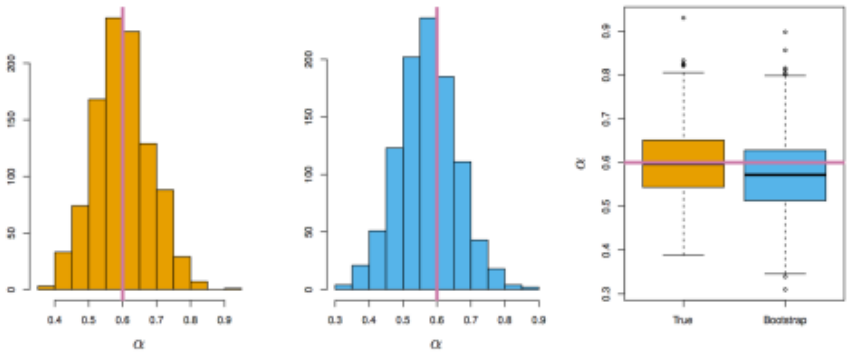
05 | Basic Statistics and a Bit of Bootstrap

Bootstrap methods

- Bootstrap methods
  - ✓ 표본 D에서 추가적으로 표본을 복원 추출하여 B개의 $Bootstrap\ sampling$ 생성 후, 각각의 표본에 대한 통계량을 다시 계산하는 방법
  - ✓ 이렇게 계산된 값은 P로부터 독립적인 표본을 얻었을 때의 결과와 비슷하게 다루어 분석
    - 중요한 점 : $Bootstrap\ Sampling$을 통해 생성된 표본들이 P로부터 독립적인 표본들과 비슷한 성질을 가지고 있음
  - ✓ 따라서 $Bootstrap\ methods$는 P로부터 독립적인 샘플을 얻는 것이 어려운 경우에 유용 이를 통해 분포를 추정하거나 통계적 가설검정 수행 가능
- Bootstrap methods를 활용하여 표준 오차를 계산하는 방법
  - ✓ B개의 $Bootstrap$ 표본 생성
  - ✓ 각각의 표본에 대해 추정량을 계산하고 표본의 분산 혹은 표준편차 구하기
  - ✓ 이를 통해 $Bootstrap$ 분산을 계산하여 추정량의 표준 오차 추정 가능

$$\hat{\mu}(D_n) \pm \sqrt{Boostrap\ Variance}$$

μ = 추정량



Q&A

감사합니다.