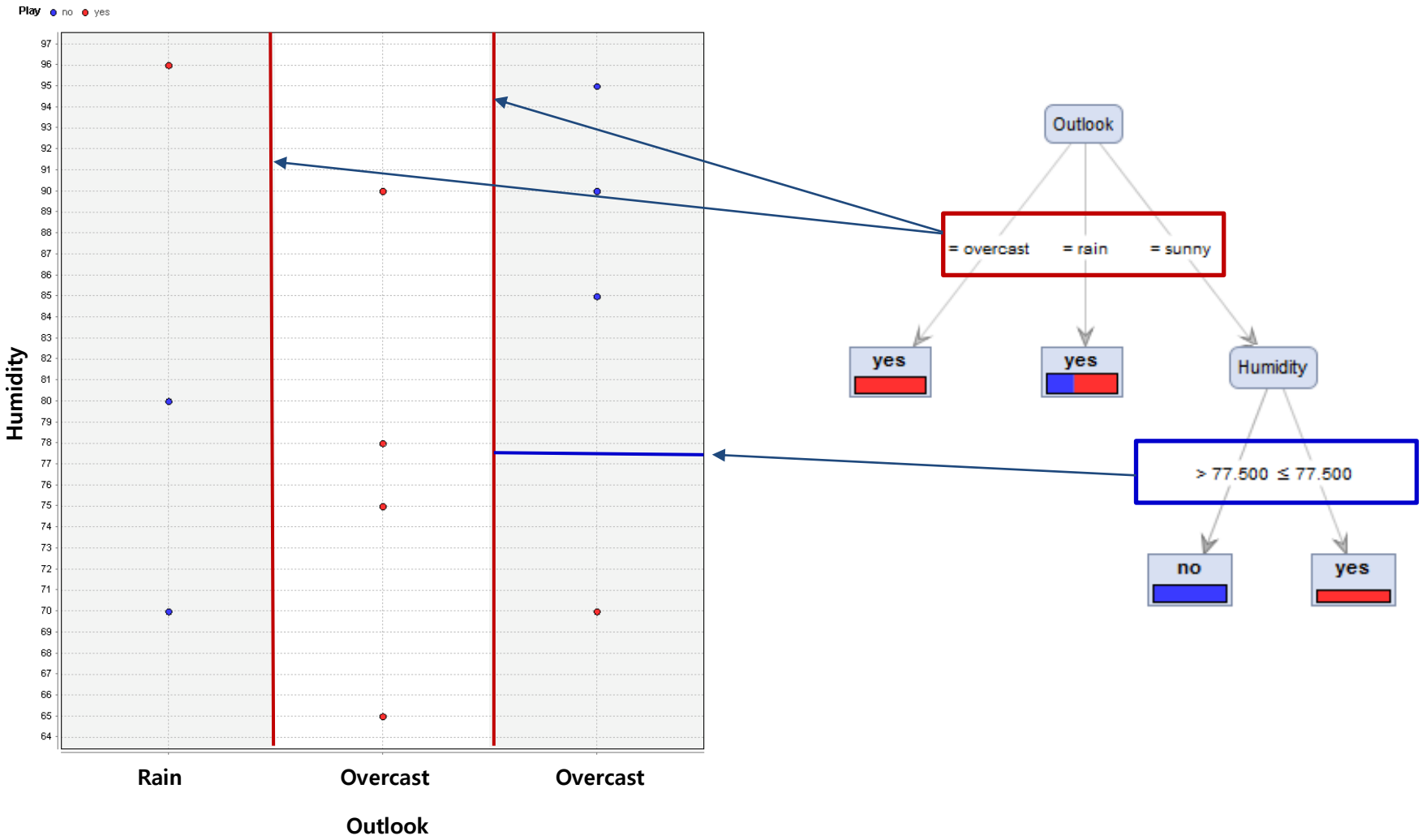
Data Science

04.01-03 의사결정나무 가지치기(pruning)

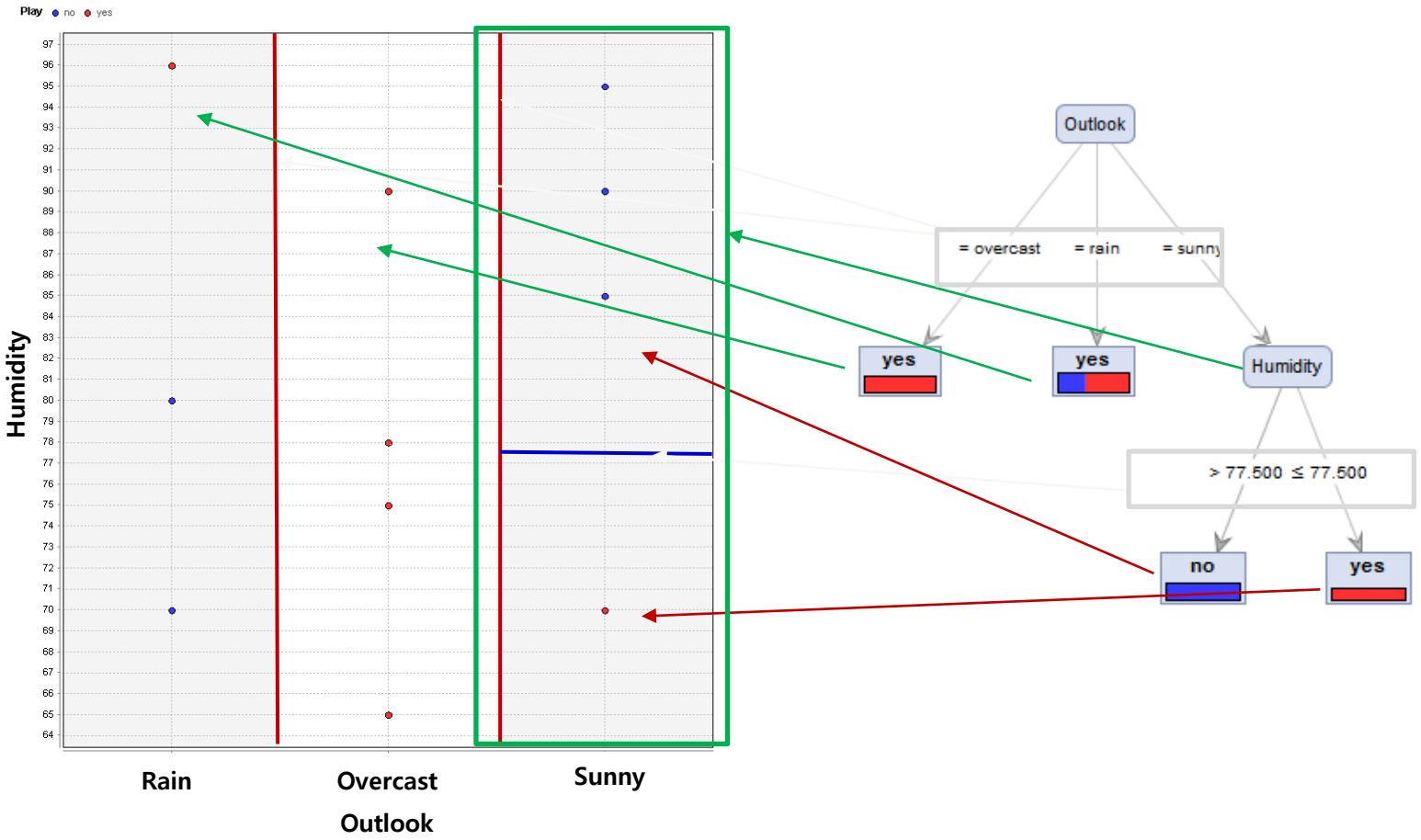
Split 영역

- Split 은 Scatter plot 에서 수평, 수직선의 경계로 나타남 → 결정경계



Split 영역

- Split 은 Scatter plot에서 수평, 수직선의 경계로 나타남
- Node 가 차지하는 영역은 직사각형으로 나타남 → 클래스 영역
  - 아래 레벨의 노드일 수록 차지하는 영역은 더 작아짐.



언제 나무의 성장을 중지하여야 하는가?

나무가 너무 커지는 경우

- 각 잎노드의 사각형이 너무 작아짐
  - 필요 없는 detai로 간주되며 과적합의 가능성.
  - 과적합: 모델이 일반적인(general) 패턴보다는 학습용 데이터에 나타나는 국소적(local) 패턴을 학습
  - 과적합을 막기 위하여 나무의 크기를 조절
    → 가지치기

가지치기

- 미리 가지치기(pre-pruning)
- 나중 가지치기(post-pruning)

미리 가지치기(pre-pruning)

- 나무 모델 생성 도중 Split 를 중지하는 조건
  - 최소불순도 이득(gain)을 사전에 설정하고 이것을 충족하는 노드가 없으면 중지
  - 나무의 최대 깊이를 설정하고 이 깊이까지만 나무를 성장시킴
  - 현재의 서브트리(subtree)에 포함된 사례수가 특정수 이하인 경우 중지함.

나중 가지치기(post-pruning)

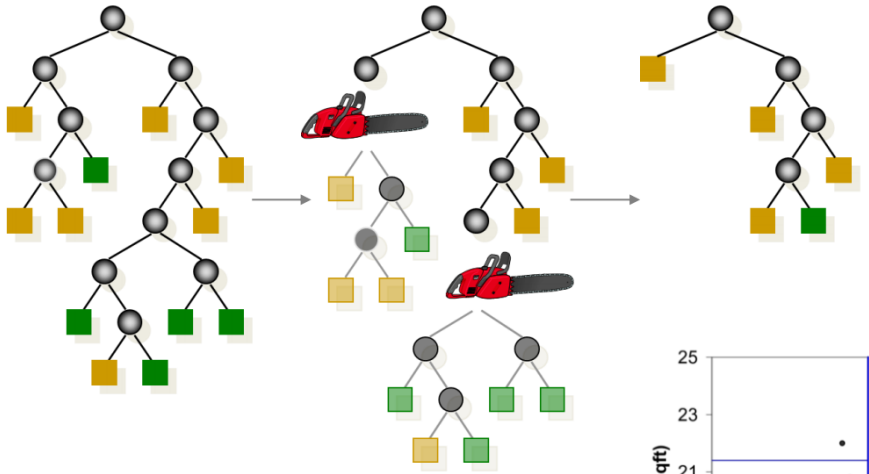
- 나무를 완전히 성장 시킨 후
  - Hunt's 의 알고리즘을 적용 더 이상 성장 시킬 수 없을 때까지 나무를 성장
  - Full tree, 과적합(overfit)되어 있음.
- 나무모델을 검증용 데이터세트에 적용하여 가지치기
  - 하위 레벨의 노드부터 시작하여 subtree에 대하여
    제거 전후의 예측 정확도를 비교하고 제거 여부를 결정
  - 나무모델의 복잡도(=잎 노드수)를 계산하여 예측정확도에 페널티로 부여함.
  - 잎노드의 라벨은 최대비율(majority) 클래스로 결정함.

Note: 검증용 데이터 세트(validation data set):

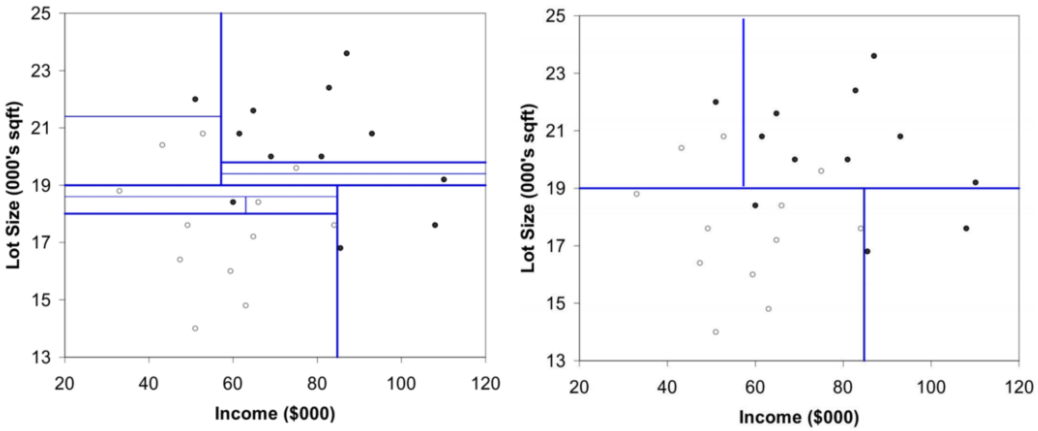
- 모델의 개선을 위하여 모델의 테스트 용으로 사용하는 데이터 셋

나중 가지치기

- 모델의 일반화(generalization)
  - 모델을 단순화 하여 과적합을 감소시켜
  - 새로운 데이터에 대한 예측 성능 향상



서브트리를 하나의 단말 노드로 축소

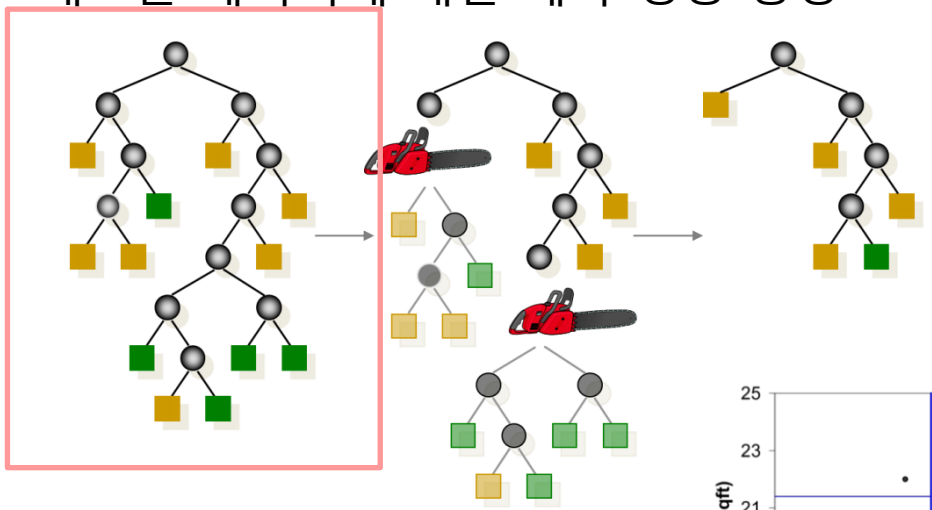


데이터 스페이스에서 작은 사각형들을 병합하는 것과 동일한 효과

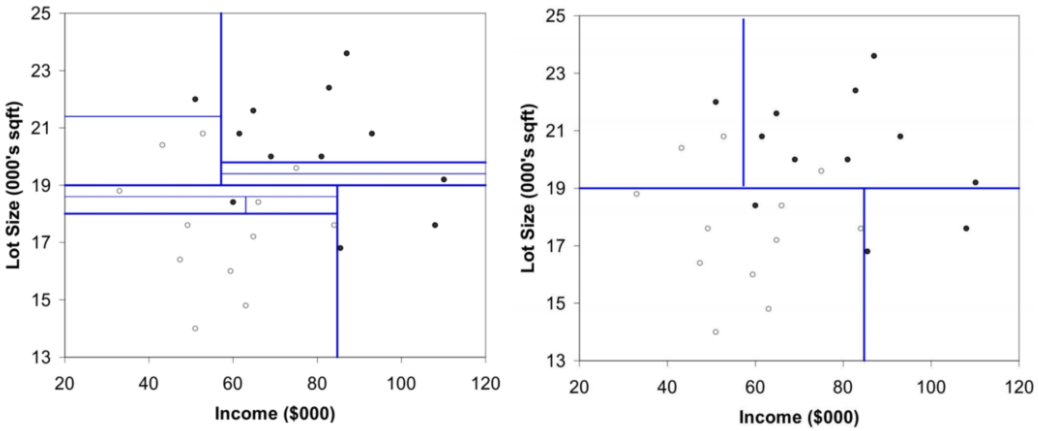
Source: https://ratsgo.github.io/machine%20learning/2017/03/26/tree/

나중 가지치기

- 모델의 일반화(generalization)
  - 모델을 단순화 하여 과적합을 감소시켜
  - 새로운 데이터에 대한 예측 성능 향상



서브트리를 하나의 단말 노드로 축소

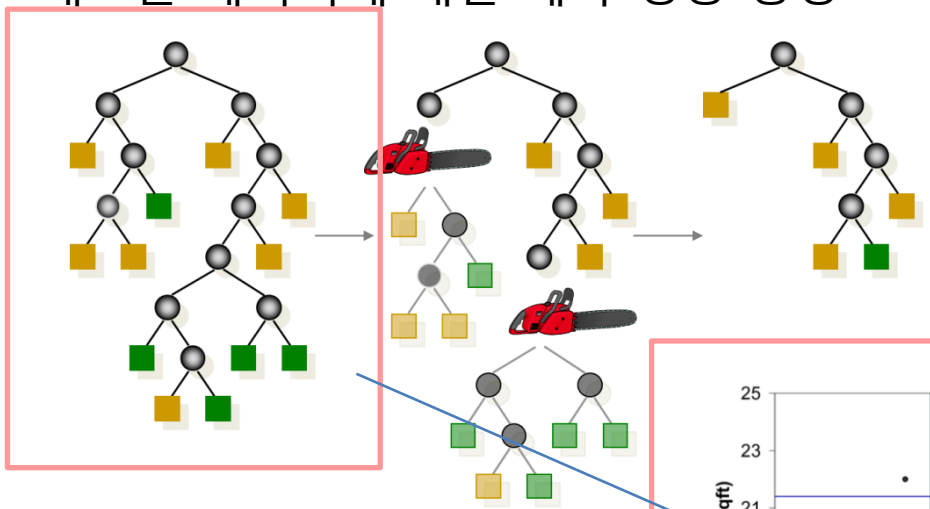


데이터 스페이스에서 작은 사각형들을 병합하는 것과 동일한 효과

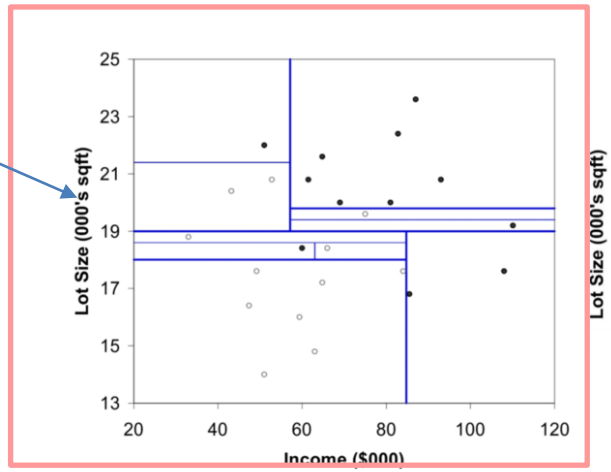
Source: https://ratsgo.github.io/machine%20learning/2017/03/26/tree/

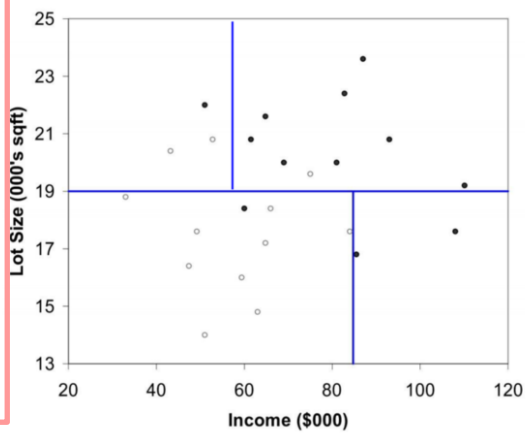
나중 가지치기

- 모델의 일반화(generalization)
  - 모델을 단순화 하여 과적합을 감소시켜
  - 새로운 데이터에 대한 예측 성능 향상



서브트리를 하나의 단말 노드로 축소



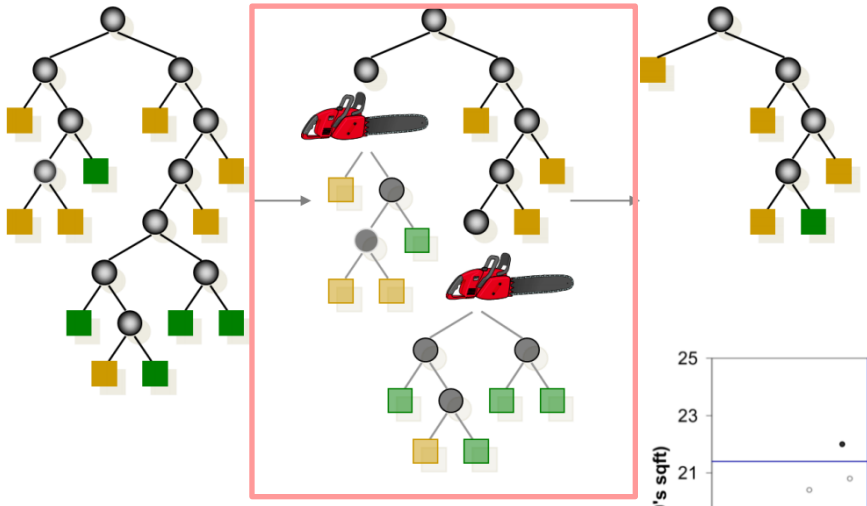


데이터 스페이스에서 작은 사각형들을 병합하는 것과 동일한 효과

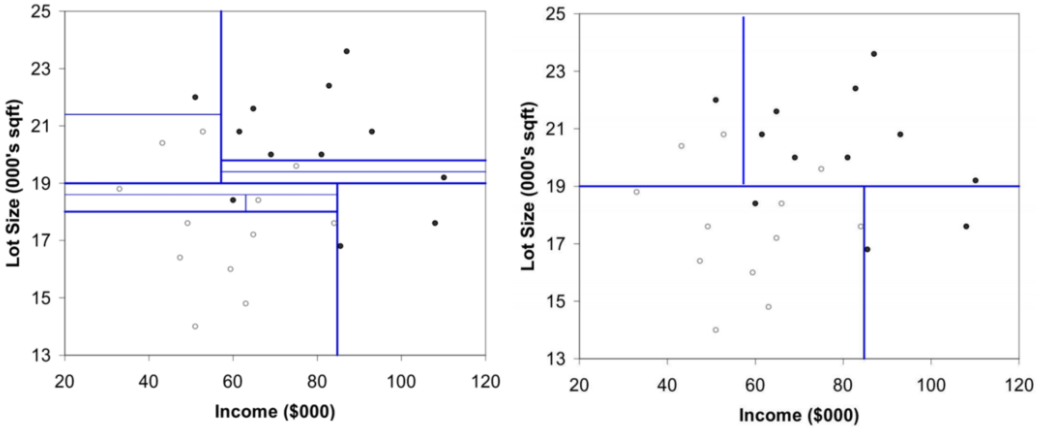
Source: https://ratsgo.github.io/machine%20learning/2017/03/26/tree/

나중 가지치기

- 모델의 일반화(generalization)
  - 모델을 단순화 하여 과적합을 감소시켜
  - 새로운 데이터에 대한 예측 성능 향상



서브트리를 하나의 단말 노드로 축소

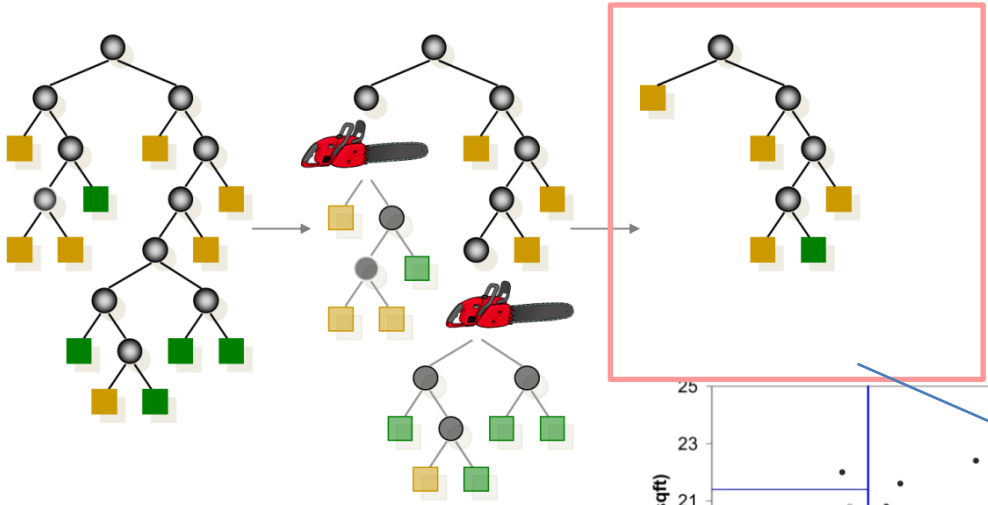


데이터 스페이스에서 작은 사각형들을 병합하는 것과 동일한 효과

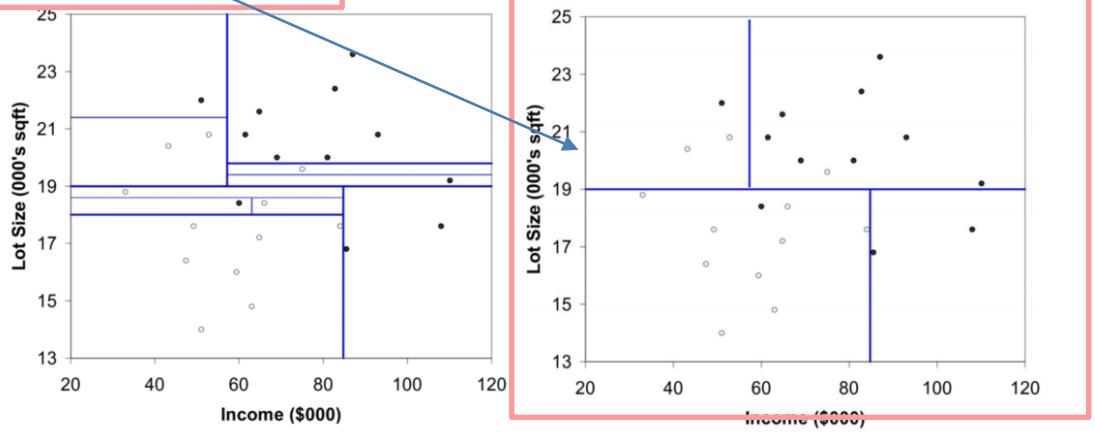
Source: https://ratsgo.github.io/machine%20learning/2017/03/26/tree/

나중 가지치기

- 모델의 일반화(generalization)
  - 모델을 단순화 하여 과적합을 감소시켜
  - 새로운 데이터에 대한 예측 성능 향상



서브트리를 하나의 단말 노드로 축소



데이터 스페이스에서 작은 사각형들을 병합하는 것과 동일한 효과

Source: https://ratsgo.github.io/machine%20learning/2017/03/26/tree/

나중 가지치기

가지치기의 비용함수

- 비용함수(cost function)를 최소로 하도록 가지치기

$$CC(T)=Err(T)+\alpha \times L(T) \quad \text{- (수식 1)}$$

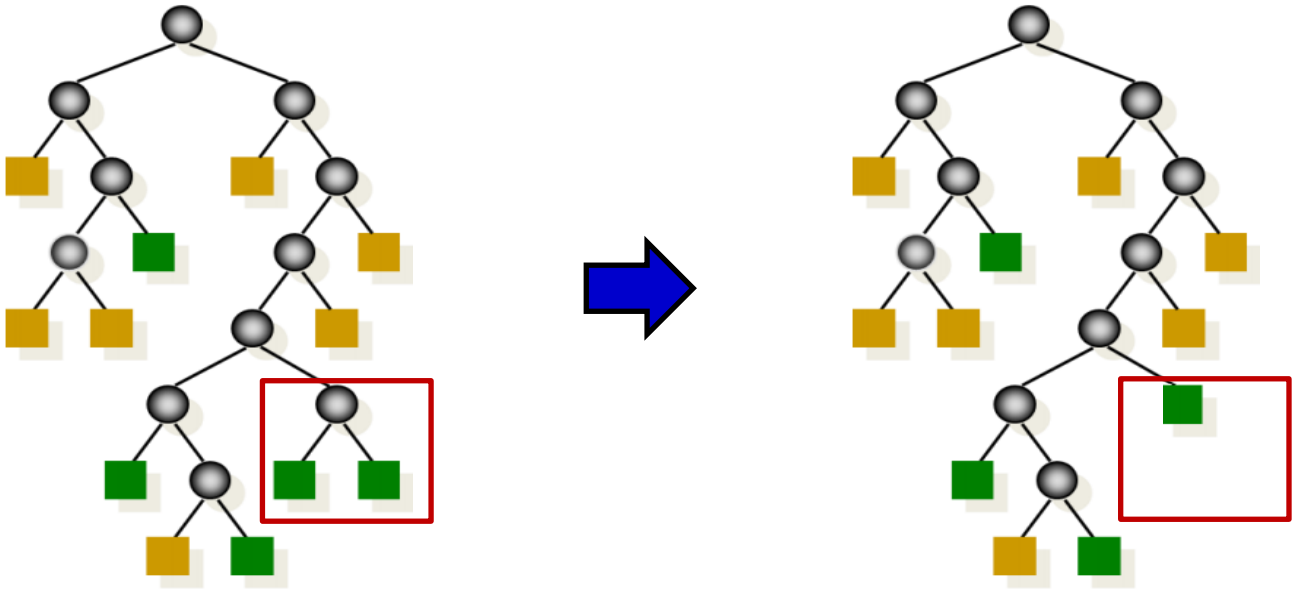
- CC(T)= 의사결정나무의 비용 복잡도
  오류가 적으면서 terminal node 수가 적은 단순한 모델일 수록 작은 값
- Err(T)= 검증데이터에 대한 오분류율
- L(T)= 단말노드-terminal node의 수(구조의 복잡도)
- α (Alpha) =ERR(T)와 L(T)를 결합하는 가중치
  사용자에 의해 부여됨, 보통 0.01~0.1의 값을 씀

Source: https://ratsgo.github.io/machine%20learning/2017/03/26/tree/

나중 가지치기

재귀적 가지치기 알고리즘

- 결정나무를 Full tree (불순도 0) 로 성장,
- 최하위의 서브트리 부터 가지치기 시도 → bottom up
- 가지치기 전, 후의 비용함수의 값을 비교 $CC(T)=Err(T)+\alpha \times L(T)$
- 전 CC > 후 CC 라면 가지 치기 실행
- 더 이상 제거 되는 서브트리가 없다면 STOP



가지치기 전

가지치기 후

Source: https://ratsgo.github.io/machine%20learning/2017/03/26/tree/

나무모델의 장단점

장점

- 계산 복잡도 대비 높은 예측 성능
- 모델을 이해하기 쉽고
- 변수 단위로 설명력을 지닌다는 강점

단점

- 결정경계(decision boundary)가 데이터 축에 수직이어서 비선형(non-linear) 데이터 분류엔 부적합

랜덤포레스트 모델

- 같은 데이터에 대해 의사결정나무를 여러 개 만들어 그 결과를 종합해 예측 성능을 높이는 기법
- 참조: https://ratsgo.github.io/machine%20learning/2017/03/17/treeensemble/

Source: https://ratsgo.github.io/machine%20learning/2017/03/26/tree/