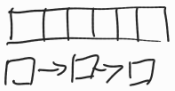
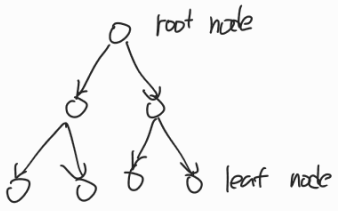
결정 트리 (decision tree)
- 장 : 예측, 해석, 사용 편리, 다용도, 성능
- 단 : 훈련 세트에 매우 민감

트리 (자료구조)

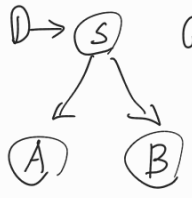




- 분류, 회귀, 다중 출력 작업 가능 => 다재다능
- 복잡한 데이터 셋도 학습 가능
- 랜덤 포레스트의 기본 구성 요소



지니계수(불순도) 측정



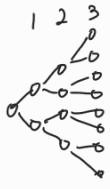
$G_T = 1 - \sum_{k=1}^{n} P_{i,k}^2$

(Gini=0)

예측
- root node에서 시작 계속해서 leaf node로 추가적인 검사 진행 (지니불순도 측정) 노드의 모든 샘플이 같은 class 속하면 순수하다.

확률 추정
- 한 sample이 클래스 k에 속할 확률 추정, 리프노드 찾기 위해 해당 node에 있는 class의 훈련 sample 비율 반환

CART 훈련 알고리즘
- 결정 트리 훈련 위해 예측과 확률 추정 쓰이게 사용



- 2개의 subset으로 나눔
- 최대 깊이 또는 불순도 줄이는 분할 발견 불가능할 때 stop
- 계산 복잡도 $O(\log_2 m)$ => 매우 빠름

엔트로피
- 열역학 : 분자의 무질서함
- 새넌의 정보 이론 : 메시지의 평균 정보량 (한 클래스만 포함하면 엔트로피 0)
- 지니불순도는 빈도가 높은 class를 특정 가지로 고립, 엔트로피는 균형적

규제 매개변수
- 제한이 없는 경우 과대적합
- 파라미터 수가 결정 X = 비파라미터 모델
- 트리의 최대 깊이 제이 가능

회귀
- CART 알고리즘에서 불순도 최소화가 아닌 평균 제곱근오차(MSE) 최소화하도록 분할
- 분류처럼 과대적합되기 쉽다.

불안정성
- 훈련 세트 회전에 민감
- 작은 변화에도 민감

랜덤 포레스트

- 배깅(or 페이스팅)을 적용한 결정 트리의 앙상블.
- 다른 코어 ⇒ 학습 속도가 빠르다.
- 트리노드 분할할 때 전체 특성이 아닌 무작위로 선택한 특성 후보 중 최적의 특성 찾음.

엑스트라 트리

- 각 노드는 무작위 특성의 서브셋 만들어 분할
- 후보 특성 사용 무작위 분할 → 최상의 분할 선택.

랜덤	편향↑	분산↓	과소적합
일반	편향↓	분산↑	과대적합

• 엑스트라 랜덤 트리 앙상블 : 무작위한 트리의 랜덤 포레스트, 편향이 늘어남(과소적합), 분산 낮춤.

특성 중요도

. 랜덤 포레스트는 상대적 중요도 측정 쉽다. ⇒ 어떤 특성이 중요한지 빠르게 파악
⇒ 결과 해석에 이점. (설명 쉽다.)