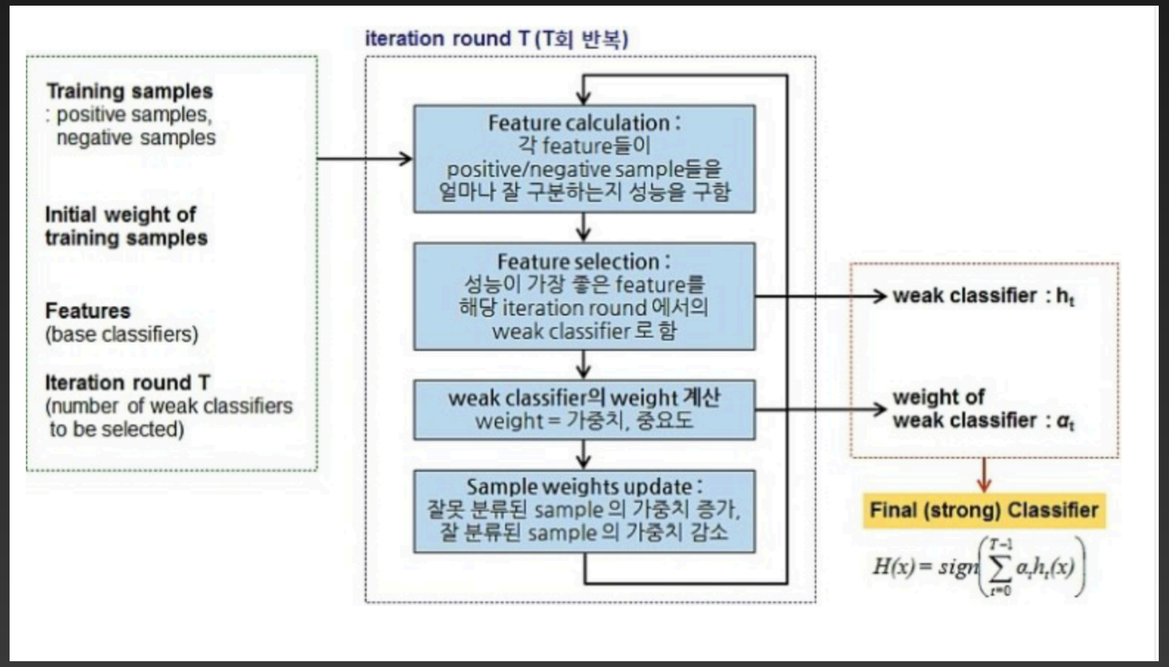
Choose Your Own Algorithm

AdaBoost

진주희

AdaBoost

- Adaptive Boosting의 약자
- 약한 분류기(Weak Classifier)들을 순차적 학습(sequential)시켜 마지막의 강한 분류기의 성능을 증폭시킨다



$$H(x) = \alpha_1 h_1(x) + \alpha_2 h_2(x) + ... + \alpha_t h_t(x) = \sum_{t=1}^{T} \alpha_t h_t(x)$$

H(x) : 최종 강한 분류기

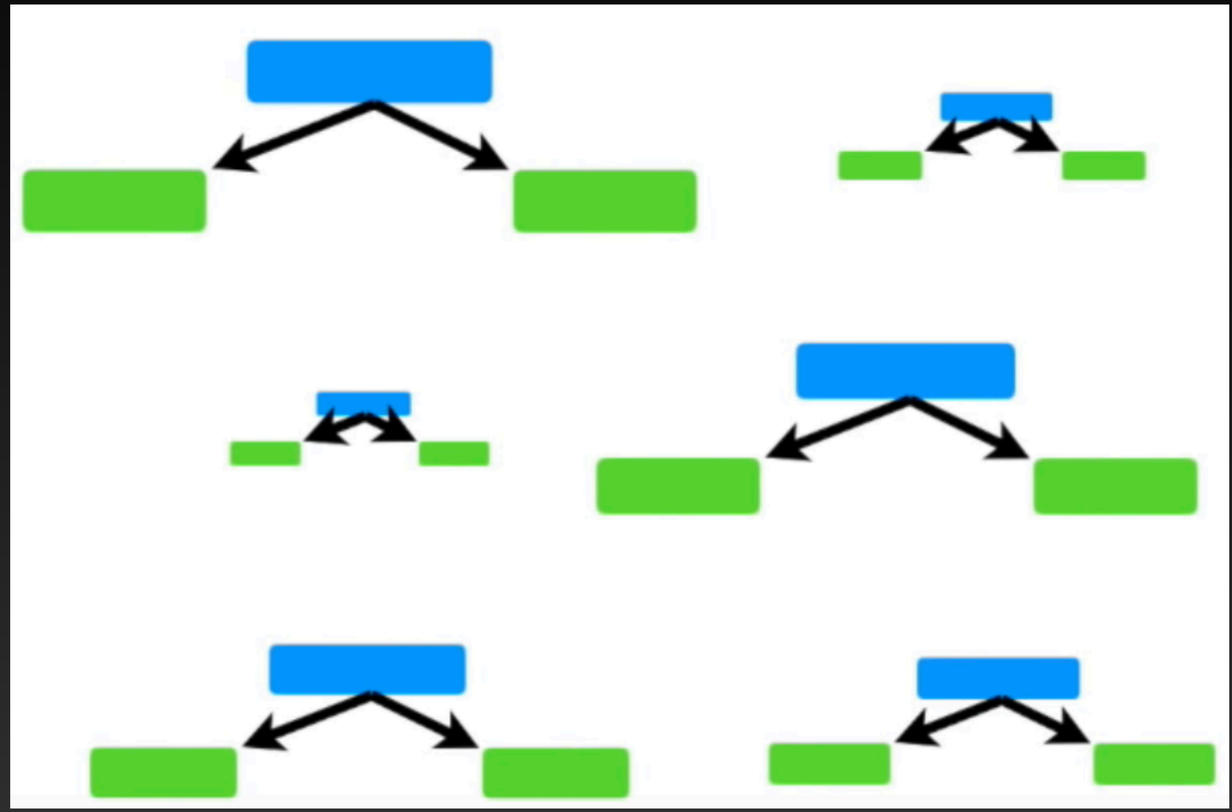
h : 약한 분류기

α : 약한 분류기의 가중치

t : 반복 횟수

AdaBoost 기본특징

1. 약한 학습기(weak learner)로 구성되어 있으며, 이는 stump의 형태이다
2. 각 stump들의 가중치가 다르다
3. 각 stump들의 에러는 다음 stump들의 결과에 영향을 준다



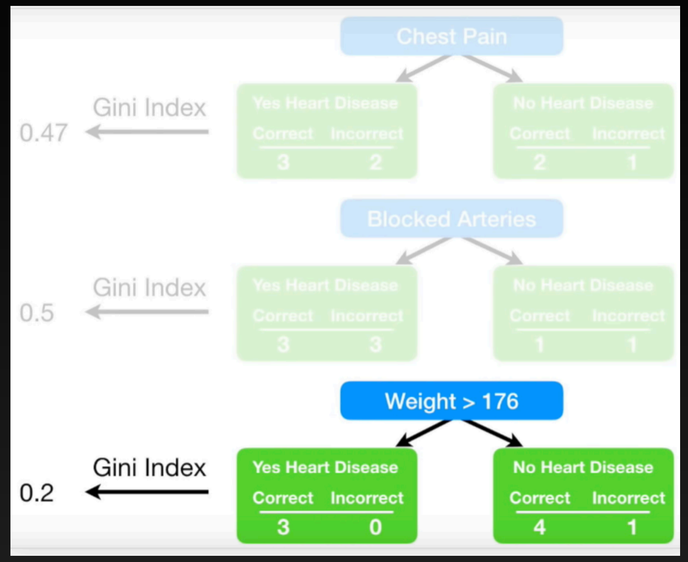
작동 원리

첫 stump 정하기

- 각 stump별로 분류 결과를 바탕으로 지니 인덱스값을 구하여 가장 작은 stump를 최초로 설정

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

- 처음 테스트별 가중치는 같다고 설정
- total error = incorrect / total data



Chest Pain	Blocked Arteries	Patient Weight	Heart Disease	Sample Weight
Yes	Yes	205	Yes	1/8
No	Yes	180	Yes	1/8
Yes	No	210	Yes	1/8
Yes	Yes	167	Yes	1/8
No	Yes	156	No	1/8
No	Yes	125	No	1/8
Yes	No	168	No	1/8
Yes	Yes	172	No	1/8

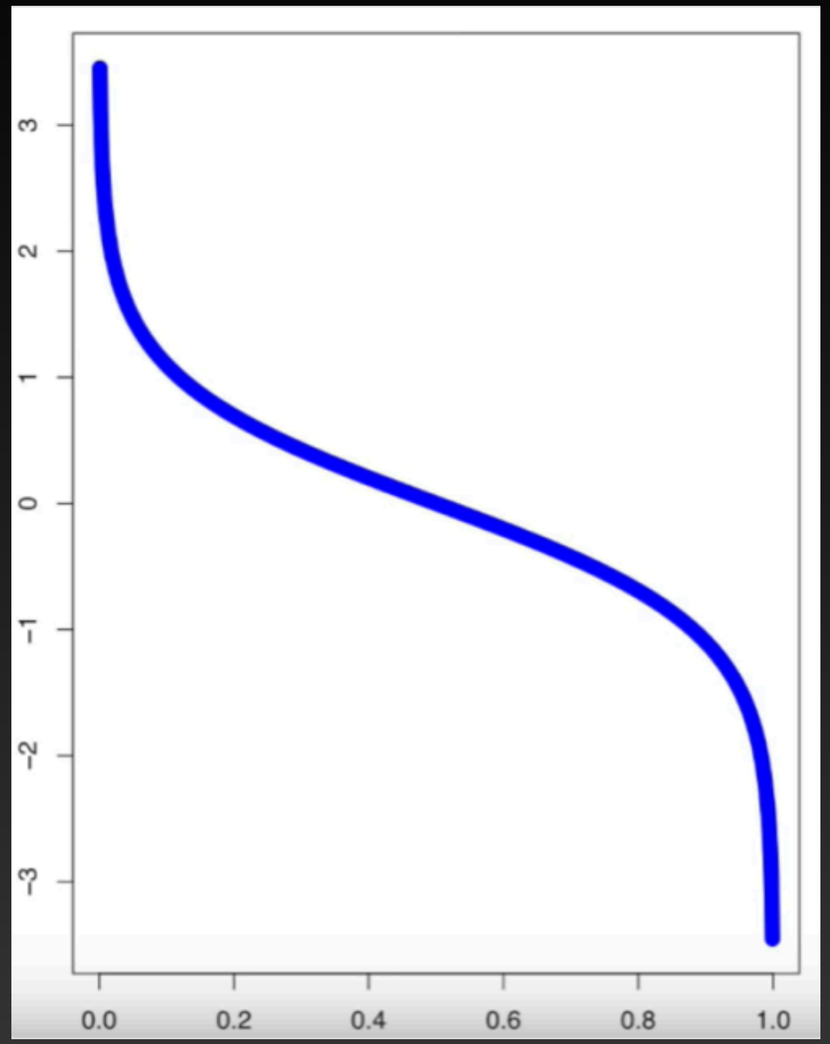
작동원리

Amount of Say

- 최종 분류에 있어서 해당 stump가 얼마만큼의 영향을 주는가

$$\text{Amount of Say} = \frac{1}{2}\log\left(\frac{1 - \text{Total Error}}{\text{Total Error}}\right)$$

- Amount of say 계산으로 완전히 틀린 모델의 결과는 반대로 바꾸고 항상 맞는 모델의 가중치는 증가 시킴



작동 원리

새로운 가중치 구하기

- 이전에 구한 amount of say를 이용 다음 stump로 갈때 틀린 샘플새로운 가중치를 구해줌

$$\text{New Sample Weight} = \text{sample weight} \times e^{\text{amount of say}}$$

- 이러면 가중치가 더 올라가서 다음 stump가 잘 못 분류된 데이터에 더 집중해서 분류함
- 계산 후에는 합이 1이아니므로 정규화

Chest Pain	Blocked Arteries	Patient Weight	Heart Disease	Sample Weight	New Weight	Norm. Weight
Yes	Yes	205	Yes	1/8	0.05	0.07
No	Yes	180	Yes	1/8	0.05	0.07
Yes	No	210	Yes	1/8	0.05	0.07
Yes	Yes	167	Yes	1/8	0.33	0.49
No	Yes	156	No	1/8	0.05	0.07
No	Yes	125	No	1/8	0.05	0.07
Yes	No	168	No	1/8	0.05	0.07
Yes	Yes	172	No	1/8	0.05	0.07

작동원리

새로운 테이블 만들기

- 새로운 sample weight을 바탕으로 랜덤하게 숫자를 뽑아 해당하는 샘플을 추가
- 가중치가 크면 뽑힐 확률이 높아지고 여러번 뽑힐 수 있음
- 이후 다시 sample weight을 같게 설정하고 계산

작동원리

최종 분류

- 여러번 진행해서 각 stump 마다 amount say를 구하고 더해서 더 큰 값을 선택해 최종 분류

- 각각의 stump는 분류력이 낮지만 순차적으로 계산한 여러결과를 종합하면 강한 학습기가 된다.