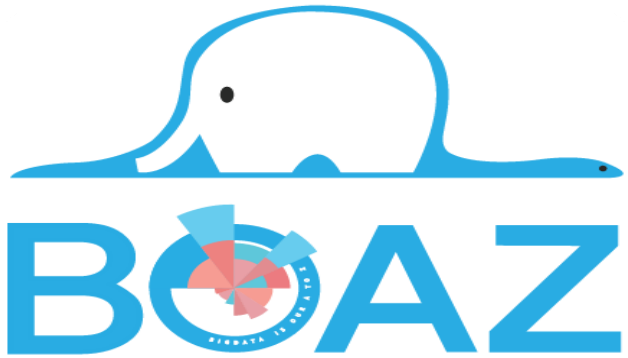
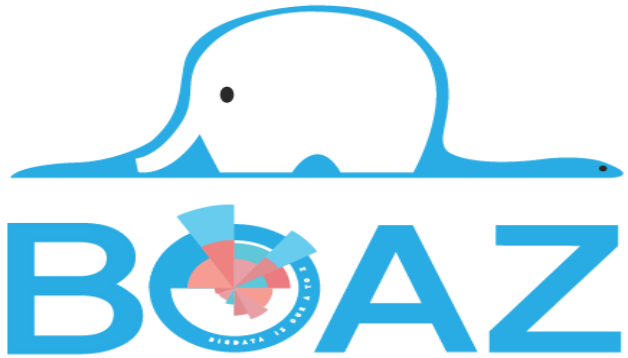


Decision Trees & Ensembles

A조



Decision Trees
(의사결정나무)

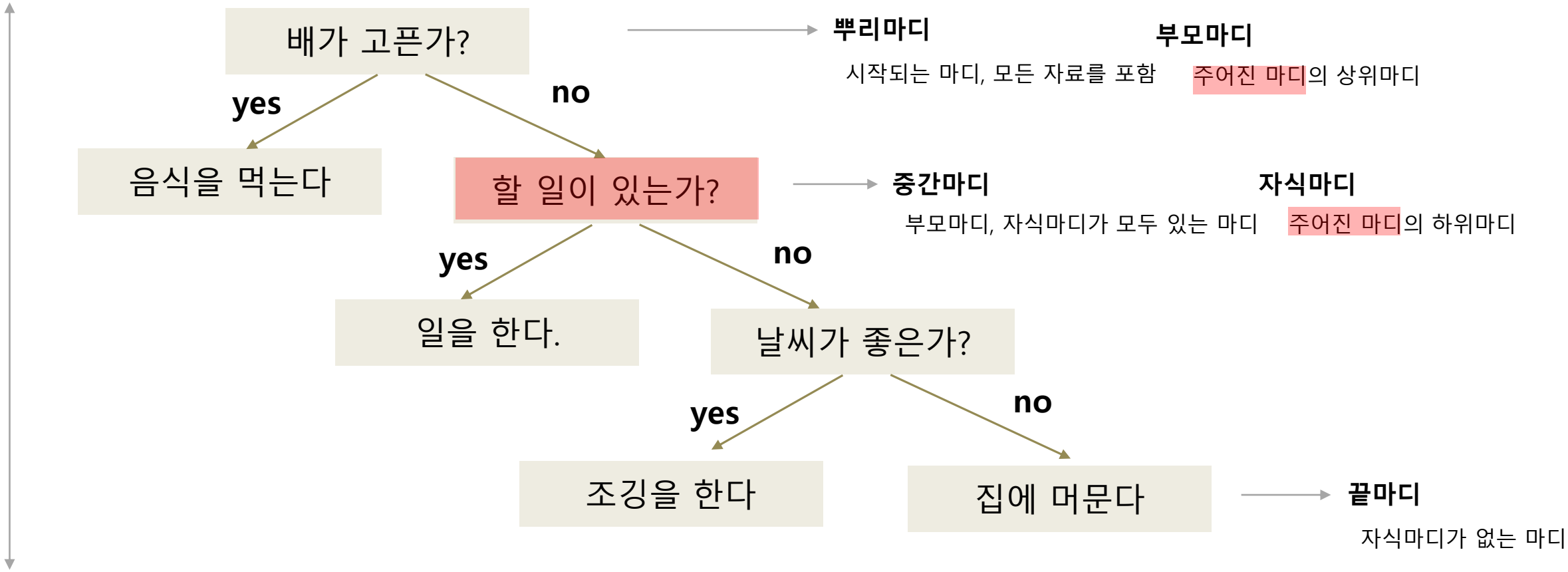
TABLE
OF -CONTENTS

▶ 의사결정나무 모형의 정의

의사결정규칙을 도표화하여 **관심대상이 되는 집단을**
몇 개의 소집단으로 **분류**하거나 **예측을 수행**하는 분석 방법

▶ 의사결정나무 모형의 목적

세분화(segmentation)	데이터를 비슷한 특성을 갖는 몇 개의 그룹으로 분할해 그룹별 특성을 발견하는 경우 ex) 시장세분화, 고객세분화
분류(classification)	관측개체를 여러 예측변수들에 근거해 목표변수의 범주를 몇 개의 등급으로 분류하고자 하는 경우 ex) 고객을 신용도에 따라 우량/불량으로 분류
예측(prediction)	자료에서 규칙을 찾아내고 이를 이용해 미래의 사건을 예측하고자 하는 경우 ex) 고객속성에 따라 대출한도액 예측
차원축소 및 변수선택	매우 많은 수의 예측변수 중에서 목표변수에 큰 영향을 미치는 변수들을 골라 내는 경우
교호작용효과의 파악	여러 개의 예측변수들을 결합해 목표변수에 작용하는 규칙(교호작용효과)을 파악하고자 하는 경우



깊이

뿌리마디부터 끝마디까지의 중간마디들의 수

- 노드를 가장 효율적으로 선정하고 배치하기 위해선 **정보획득량**이라는 개념과 **엔트로피**라는 개념이 필요
- 부모마디보다 **자식마디의 순도가 증가**하도록 분리해야 함

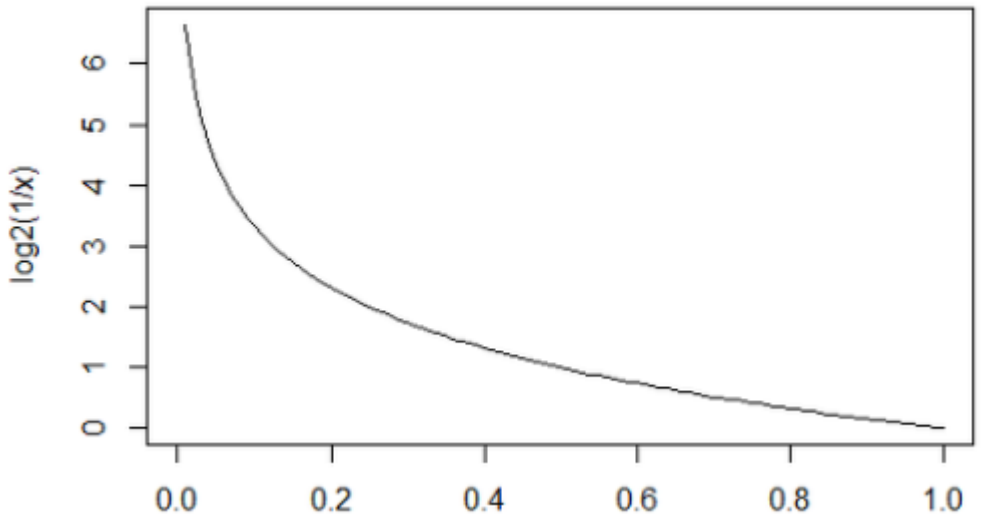
정보획득량

어떤 사건이 얼마만큼의 정보를 줄 수 있는지 수치화한 값

정보함수

- 정보 함수는 정보의 가치를 반환함
- 발생할 확률이 작은 사건일수록 정보의 가치가 큼
- 발생할 확률이 큰 사건일수록 정보의 가치가 작음

[정보함수]



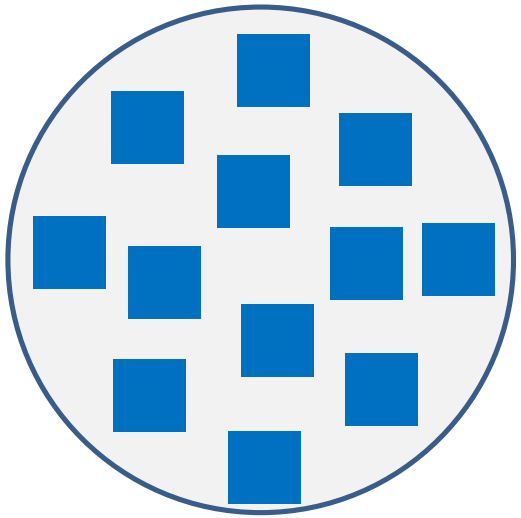
$$I(x) = \log_2 \frac{1}{p(x)}$$

엔트로피 - 모델의 불순도(impurity)를 측정함

- 의사결정 트리에서 가장 비균질한 예측값이 루트 노드에 가장 가깝게 위치함

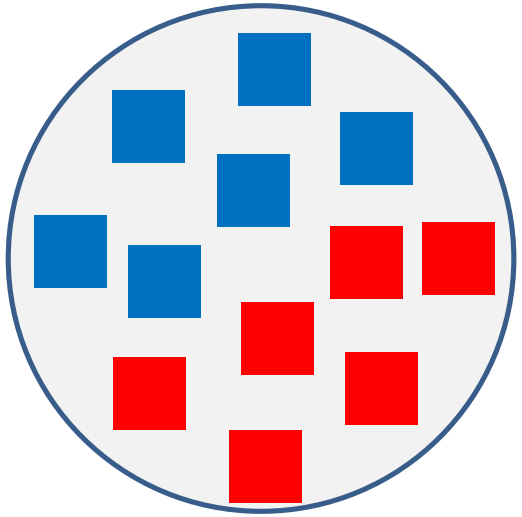
$$H(X) = -\sum_{i=1}^{n} P(X=i)\log_2 P(X=i)$$

Entropy = 0

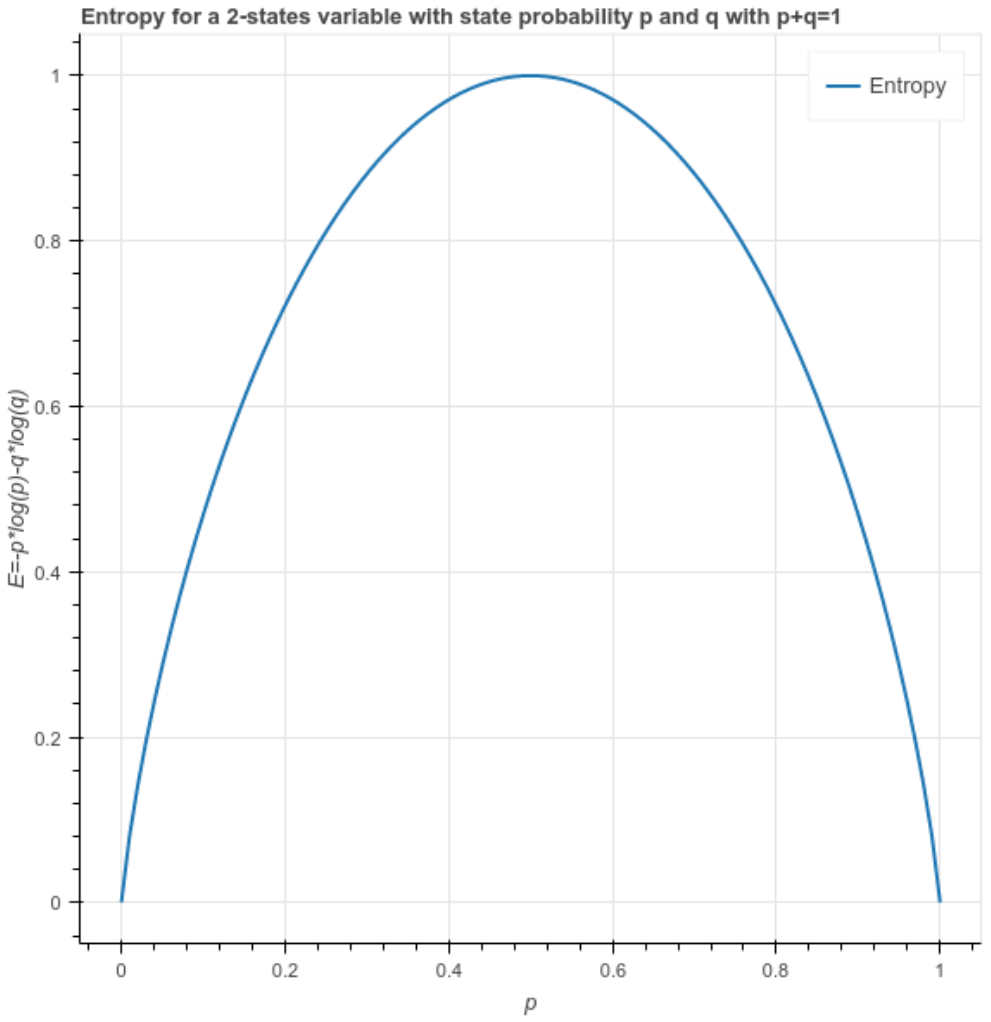


표본 : 완전균질

Entropy = 1



표본 : 동등하게 분리



Ex. 정상적인 동전과 한쪽으로 치우친 동전이 있다. 앞면과 뒷면이 나올 확률은 정상적인 동전은 1/2이지만, 치우친 동전은 앞면이 나올 확률이 1/3이고 뒷면이 나올 확률은 2/3다. 두 동전의 엔트로피를 구하고 모델링 관점에서 어느 것이 나은지 밝혀보라.

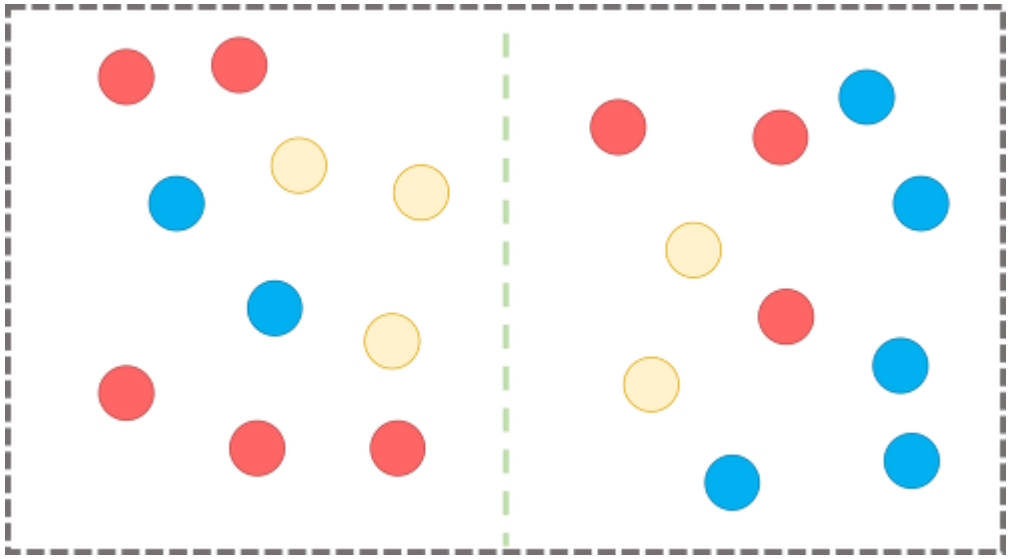
정상적인 동전의 엔트로피

$$-\frac{1}{2}*\log_2\frac{1}{2}-\frac{1}{2}*\log_2\frac{1}{2}=1bits$$

치우친 동전의 엔트로피

$$-\frac{1}{3}*\log_2\frac{1}{3}-\frac{2}{3}*\log_2\frac{2}{3}=0.9183bits$$

정보획득량 전체 엔트로피에서 분류 후 엔트로피 뺀 값



	좌	우	총계
빨간	5	3	8
파랑	2	5	7
노랑	3	2	5
총계	10	10	20

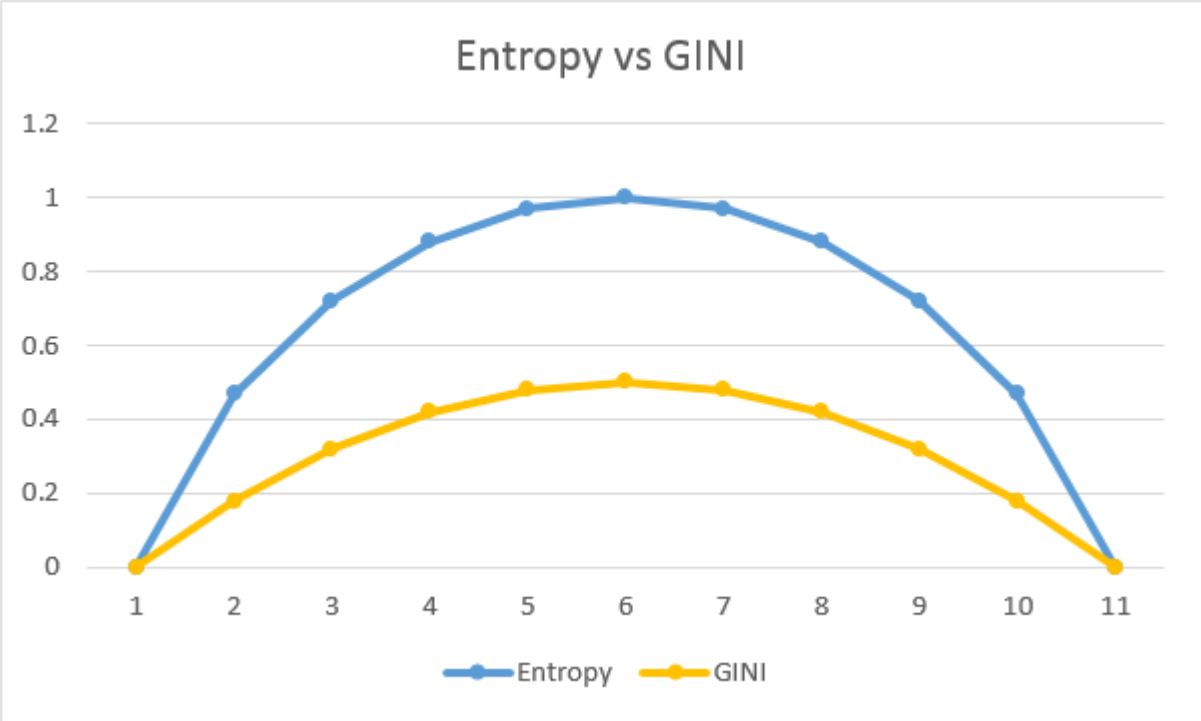
$$G(S) = E(S) - E(S')$$

$$E(S') = \sum \frac{S_a}{S} E(S_a)$$

$$E(S) = -\frac{8}{20}\log_2\frac{8}{20} - \frac{7}{20}\log_2\frac{7}{20} - \frac{5}{20}\log_2\frac{5}{20} \approx 1.558$$

$$E(S') = \frac{10}{20}\left(-\frac{5}{10}\log_2\frac{5}{10} - \frac{2}{10}\log_2\frac{2}{10} - \frac{3}{10}\log_2\frac{3}{10}\right) + \frac{10}{20}\left(-\frac{3}{10}\log_2\frac{3}{10} - \frac{5}{10}\log_2\frac{5}{10} - \frac{2}{10}\log_2\frac{2}{10}\right) \approx 1.485$$

지니



[지니와 엔트로피의 유사성]

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

i = 부류개수를 뜻함

- 지니 불순도는 잘못된 분류를 측정하는 도구
- 다부류 분류기에 적용됨
- 엔트로피와 거의 동일하지만 훨씬 더 빨리 계산 가능

EX

- 다음 예제의 반응 변수는 테니스를 칠 것인지, 말 것인지의 두 부류값만을 가진다.
- 다음의 표는 여러 날에 걸쳐 다양한 조건을 모두 기록하며 작성했다.
- 주어진 과제는 최종 결과값(예/아니요)을 결정하는데 있어 가장 중요한 역할을 하는 변수가 무엇인지 찾아내는 것이다.

일	전망	기온	습도	바람	테니스유무
D1	맑음	높음	높음	약함	아니요
D2	맑음	높음	높음	강함	아니요
D3	흐림	높음	높음	약함	예
D4	비	보통	높음	약함	예
D5	비	낮음	보통	약함	예
D6	비	낮음	보통	강함	아니요
D7	흐림	낮음	보통	강함	예
D8	맑음	보통	높음	약함	아니요
D9	맑음	낮음	보통	약함	예
D10	비	보통	보통	약함	예
D11	맑음	보통	보통	강함	예
D12	흐림	보통	높음	강함	예
D13	흐림	높음	보통	약함	예
D14	비	보통	높음	강함	아니요

EX - (1) 습도 변수를 택한 경우

CHAID - 카이제곱통계량

- 습도는 두 부류밖에 없고 기댓값은 변수를 어떻게 구분하는지 계산하기 위해 고르게 분포돼 있어야 한다.

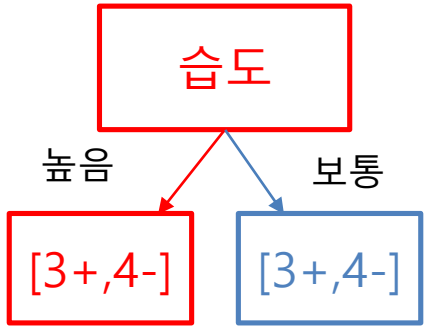
습도분류	테니스 유무		기대		차이	
	아니요	예	아니요	예	아니요	예
높음	4	4	2.5	4.5	1.5	-1.5
보통	1	6	2.5	4.5	-1.5	1.5
	5	9	5	9		

$$\chi^2\text{값계산} : \sum \frac{(O-E)^2}{E} = \frac{(1.5)^2}{2.5} + \frac{(-1.5)^2}{4.5} + \frac{(-1.5)^2}{2.5} + \frac{(1.5)^2}{4.5} = 2.8$$

$$\text{자유도계산} = (r-1)*(c-1) = (2-1)*(2-1) = 1$$

∴ P-value = 0.0942

EX - (1) 습도 변수를 택한 경우



C4.5

$$\text{엔트로피} = -\left(\frac{9}{14}\right)*\log_2\left(\frac{9}{14}\right) - \left(\frac{5}{14}\right)*\log_2\left(\frac{5}{14}\right) = 0.9402$$

$$\text{엔트로피}_{\text{높음}} = -\left(\frac{3}{7}\right)*\log_2\left(\frac{3}{7}\right) - \left(\frac{4}{7}\right)*\log_2\left(\frac{4}{7}\right) = 0.9851$$

$$\text{엔트로피}_{\text{보통}} = -\left(\frac{1}{7}\right)*\log_2\left(\frac{1}{7}\right) - \left(\frac{6}{7}\right)*\log_2\left(\frac{6}{7}\right) = 0.5916$$

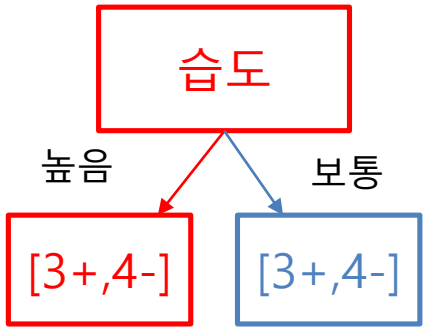
$$\text{정보이득} = \text{전체엔트로피} - \left(\frac{7}{14}\right)*\text{엔트로피}_{\text{높음}} - \left(\frac{7}{14}\right)*\text{엔트로피}_{\text{보통}}$$

$$= 0.9402 - \left(\frac{7}{14}\right)*0.9851 - \left(\frac{7}{14}\right)*0.5916 = 0.1518$$

** 위와 비슷한 방법으로 **모든 변수에 관한 정보 이득값을 계산**하고 **가장 높은 정보 이득값을 가진 최적 변수를 선택**한다.

EX - (1) 습도 변수를 택한 경우

CART



$$\text{지니}_{\text{높음}} = 1 - \left(\frac{3}{7}\right)^2 - \left(\frac{4}{7}\right)^2 = 0.489$$

$$\text{지니}_{\text{보통}} = 1 - \left(\frac{1}{7}\right)^2 - \left(\frac{6}{7}\right)^2 = 0.2448$$

$$\text{예상지니} = \left(\frac{7}{14}\right)*0.489 + \left(\frac{7}{14}\right)*0.2448 = 0.3669$$

** 위와 비슷한 방법으로 **모든 변수에 관해 기대 지니 값을 계산**하고 **최저 기댓값을 가진 변수를 최적 변수로 선택**한다.

EX - (2) 바람 변수를 택한 경우

CHAID - 카이제곱통계량

- 바람변수는 두 부류밖에 없고 기대 변수는 쉽게 구분 가능하도록 고르게 분포되어 있다.

바람분류	테니스 유무		기대		차이	
	아니요	예	아니요	예	아니요	예
약함	2	6	2.5	4.5	-0.5	1.5
강함	3	3	2.5	4.5	0.5	-1.5
	5	9	5	9		

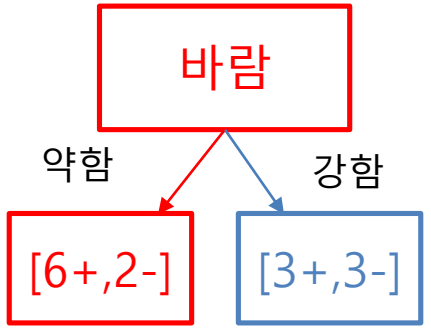
$$\chi^2\text{값계산} : \sum \frac{(O-E)^2}{E} = \frac{(-0.5)^2}{2.5} + \frac{(1.5)^2}{4.5} + \frac{(0.5)^2}{2.5} + \frac{(-1.5)^2}{4.5} = 1.2$$

$$\text{자유도계산} = (r-1)*(c-1) = (2-1)*(2-1) = 1$$

∴ **P-value = 0.2733**

EX - (2) 바람 변수를 택한 경우

C4.5



$$엔트로피 = -\left(\frac{9}{14}\right)*\log_2\left(\frac{9}{14}\right) - \left(\frac{5}{14}\right)*\log_2\left(\frac{5}{14}\right) = 0.9402$$

$$엔트로피_{약함} = -\left(\frac{6}{8}\right)*\log_2\left(\frac{6}{8}\right) - \left(\frac{2}{8}\right)*\log_2\left(\frac{2}{8}\right) = 0.8112$$

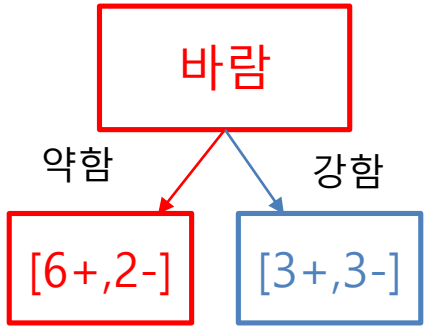
$$엔트로피_{강함} = -\left(\frac{3}{6}\right)*\log_2\left(\frac{3}{6}\right) - \left(\frac{3}{6}\right)*\log_2\left(\frac{3}{6}\right) = 1$$

$$정보이득 = 전체엔트로피 - \left(\frac{8}{14}\right)*엔트로피_{약함} - \left(\frac{6}{14}\right)*엔트로피_{강함} = 0.0482$$

** 위와 비슷한 방법으로 **모든 변수에 관한 정보 이득값을 계산**하고 **가장 높은 정보 이득값을 가진 최적 변수를 선택**한다.

EX - (2) 바람 변수를 택한 경우

CART



$$지니_{약함} = 1 - (\frac{6}{8})^2 - (\frac{2}{8})^2 = 0.375$$

$$지니_{강함} = 1 - (\frac{3}{6})^2 - (\frac{3}{6})^2 = 0.5$$

$$예상지니 = (\frac{8}{14}) * 0.375 + (\frac{6}{14}) * 0.5 = 0.4285$$

** 위와 비슷한 방법으로 **모든 변수에 관해 기대 지니 값을 계산**하고 **최저 기댓값을 가진 변수를 최적 변수로 선택**한다.

EX - 결과 해석

변수	CHAID	C4.5	CART
습도	0.0942	0.1518	0.3669
바람	0.2733	0.0482	0.4285

낮을 수록 높을 수록 낮을 수록

∴ 세 가지 척도를 모두 계산한 결과 습도가 바람에 비해 더 나은 분류기로 증명
세 가지 척도 모두 비슷한 결과 도출!

모델학습 – (1) 재귀적 분귀 (recursive partitioning)

구매가	유지비	문의 수	탑승 가능인원	안전	평가
Vhigh	Vhigh	2	2	Low	Unacc
Vhigh	Vhigh	2	2	Med	Acc
Vhigh	High	2	2	High	Unacc
Vhigh	High	4	2	Low	Acc
High	High	4	2	Med	Unacc
High	Med	4	2	High	Acc
Med	Med	4	2	Low	Unacc
Med	Low	5	2	Med	Acc
Med	Low	5	2	High	Acc
Med	Low	5	4	Low	Acc

$$E(S) = -\frac{4}{10}\log_2\frac{4}{10} - \frac{6}{10}\log_2\frac{6}{10} \approx 0.971$$

$$E(S') = -\frac{1}{10}\log_2 1 - \frac{9}{10}\left(\frac{6}{9}\log_2\frac{6}{9} + \frac{3}{9}\log_2\frac{3}{9}\right) \approx 0.826$$

$$G(S) = E(S) - E(S') \approx 0.145$$

구매가	유지비	문의 수	탑승 가능인원	안전	평가
Vhigh	Vhigh	2	2	Low	Unacc
Vhigh	Vhigh	2	2	Med	Acc
Vhigh	High	2	2	High	Unacc
Vhigh	High	4	2	Low	Acc
High	High	4	2	Med	Unacc
High	Med	4	2	High	Acc
Med	Med	4	2	Low	Unacc
Med	Low	5	2	Med	Acc
Med	Low	5	2	High	Acc
Med	Low	5	4	Low	Acc

구매가	유지비	문의 수	탑승 가능인원	안전	평가
Vhigh	Vhigh	2	2	Low	Unacc
Vhigh	Vhigh	2	2	Med	Acc
Vhigh	High	2	2	High	Unacc
Vhigh	High	4	2	Low	Acc
High	High	4	2	Med	Unacc
High	Med	4	2	High	Acc
Med	Med	4	2	Low	Unacc
Med	Low	5	2	Med	Acc
Med	Low	5	2	High	Acc
Med	Low	5	4	Low	Acc

구매가	유지비	문의 수	탑승 가능인원	안전	평가
Vhigh	Vhigh	2	2	Low	Unacc
Vhigh	Vhigh	3	3	Med	Acc
Vhigh	High	2	2	High	Unacc
Vhigh	High	4	2	Low	Acc
High	High	4	2	Med	Unacc
High	Med	4	3	High	Acc
Med	Med	4	2	Low	Unacc
Med	Low	5	2	Med	Acc
Med	Low	5	2	High	Acc
Med	Low	5	4	Low	Acc

구매가	유지비	문의 수	탑승 가능인원	안전	평가
Vhigh	Vhigh	2	2	Low	Unacc
Vhigh	Vhigh	2	2	Med	Acc
Vhigh	High	2	2	High	Unacc
Vhigh	High	4	2	Low	Acc
High	High	4	2	Med	Unacc
High	Med	4	2	High	Acc
Med	Med	4	2	Low	Unacc
Med	Low	5	2	Med	Acc
Med	Low	5	2	High	Acc
Med	Low	5	4	Low	Acc

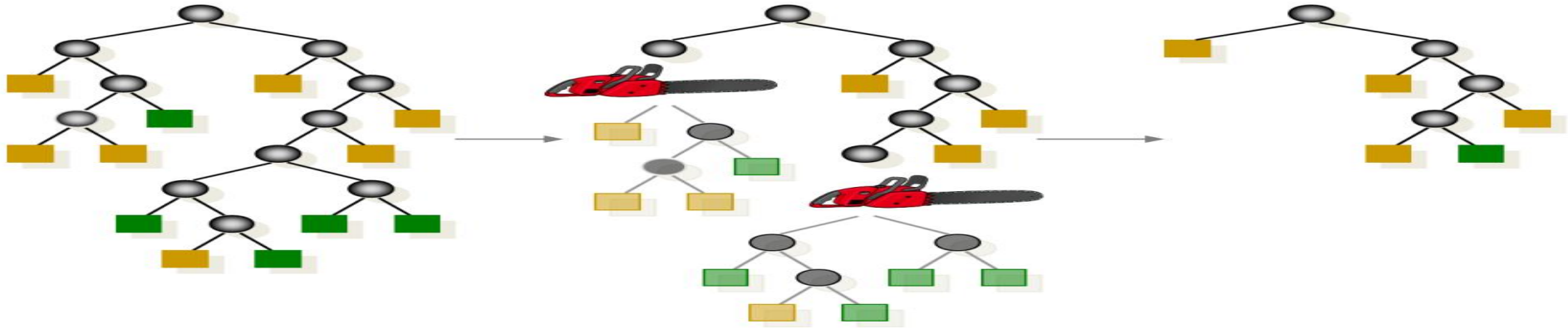
구매가	유지비	문의 수	탑승 가능인원	안전	평가
Vhigh	Vhigh	2	2	Low	Unacc
Vhigh	Vhigh	2	2	Med	Acc
Vhigh	High	2	2	High	Unacc
Vhigh	High	4	2	Low	Acc
High	High	4	2	Med	Unacc
High	Med	4	2	High	Acc
Med	Med	4	2	Low	Unacc
Med	Low	5	2	Med	Acc
Med	Low	5	2	High	Acc
Med	Low	5	4	Low	Acc

가지치기란?

- 나뭇가지를 잘라내는 것과 같다는 의미
- 의사결정나무에서 Full Tree의 분류기가 너무 많을 때 (불필요한 가지들이 있는 경우) 가지치기를 통해 분류기의 갯수를 줄인다.

나무의 불순도가 0인 상태, 가지가 다 만들어진 상태를 뜻함

사용하는 이유는?



분류기가 너무 많아서 학습데이터가 과적합(overfitting)된 경우

(왼쪽이 Full Tree, 오른쪽이 가지를 친 결과물)

Q&A) Q1. 의사결정 나무에서 분류기가 증가하면?

A1. 새로운 데이터에 대한 오분류율이 감소한다.

Q2. 그럼 계속 분류기가 증가해도 되나?

A2. 일정 수준 이상의 분류기를 가질 시, 오분류율이 되레 증가한다.

Why?

→Overfitting 때문에!

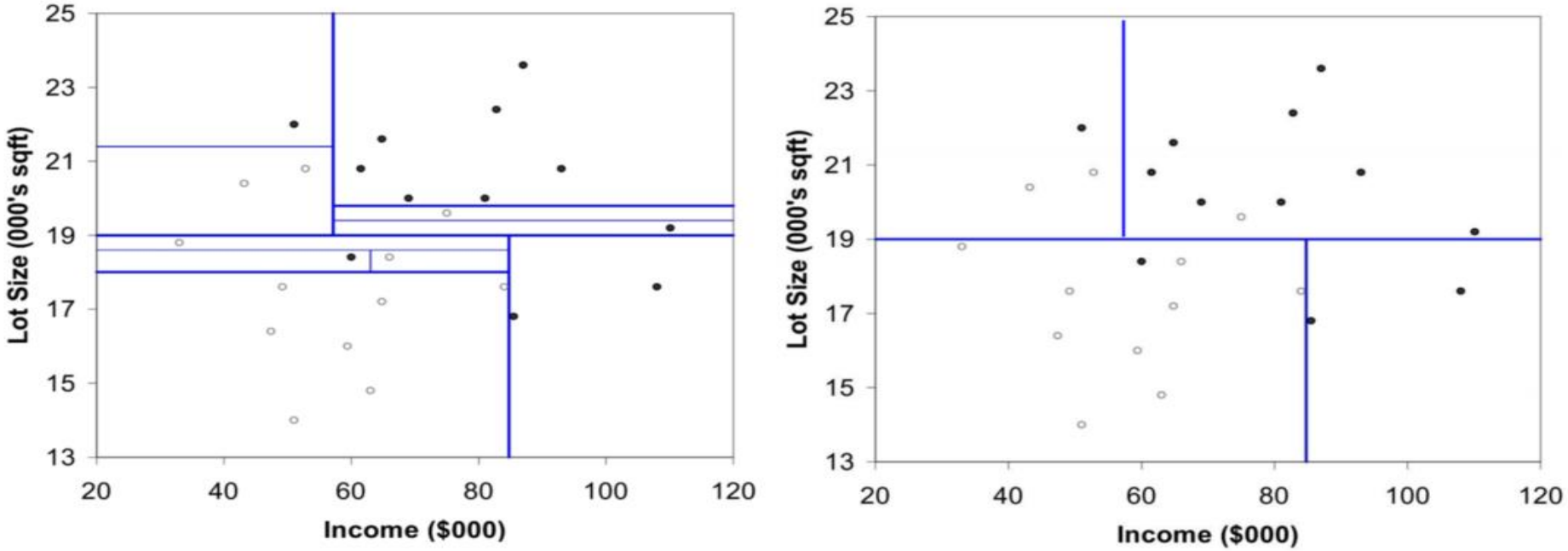
그럼 가지를 치는 시기는?

->검증데이터에 대한 오분류율이 증가하는 시점!

But, 가지치기는 데이터를 버리는 개념이 아니고 분류기를 합치는(merge) 개념으로 이해하자!

Decision Tree

Ex)



- 의사결정나무로 학습한 결과물.
- 왼쪽이 Full tree, 오른쪽이 가지치기한 결과를 시각화한 것.
- 왼쪽 그림에서 모든 terminal node의 불순도는 0.
- 하지만 terminal node가 너무 많으면 새로운 데이터에 대한 예측 성능인 일반화 (generalization) 능력이 매우 떨어짐.
- terminal node를 적절하게 합쳐주면 오른쪽 그림과 같은 결과가 나온다.

가지치기의 비용함수

$$CC(T)=Err(T)+\alpha \times L(T)CC(T)=Err(T)+\alpha \times L(T)$$

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

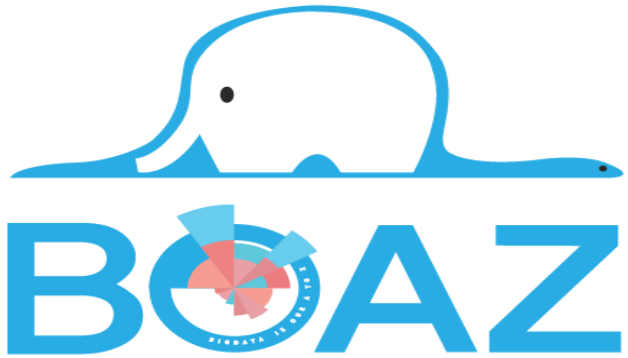
의사결정나무를 마무리하며..

- **장점**

  1. 계산복잡성 대비 높은 예측 성능
  2. 변수 단위로 설명력을 지닌다

- **단점**

  1. 결정경계(decision boundary)가 데이터 축에 수직이어서 비선형(non-linear) 데이터 분류엔 부적합
  ->이를 극복하기 위해 랜덤포레스트가 등장



Ensemble(앙상블)

TABLE OF -CONTENTS

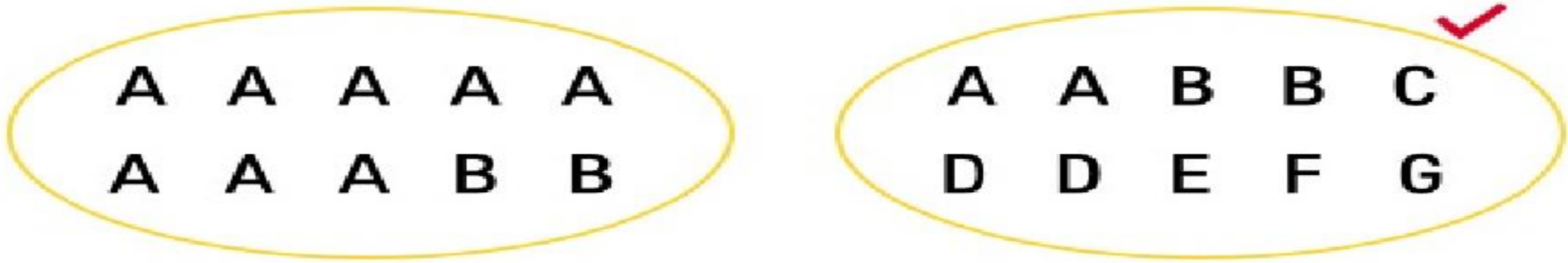
1.앙상블이란?

Ensemble Learning Method

- 여러가지 분류기를 결합함으로써 결과적으로 보다 좋은 성능(일반화 성능)을 내고자 하는 방법

주로 의사결정나무 사용

더 나은 일반화성능?

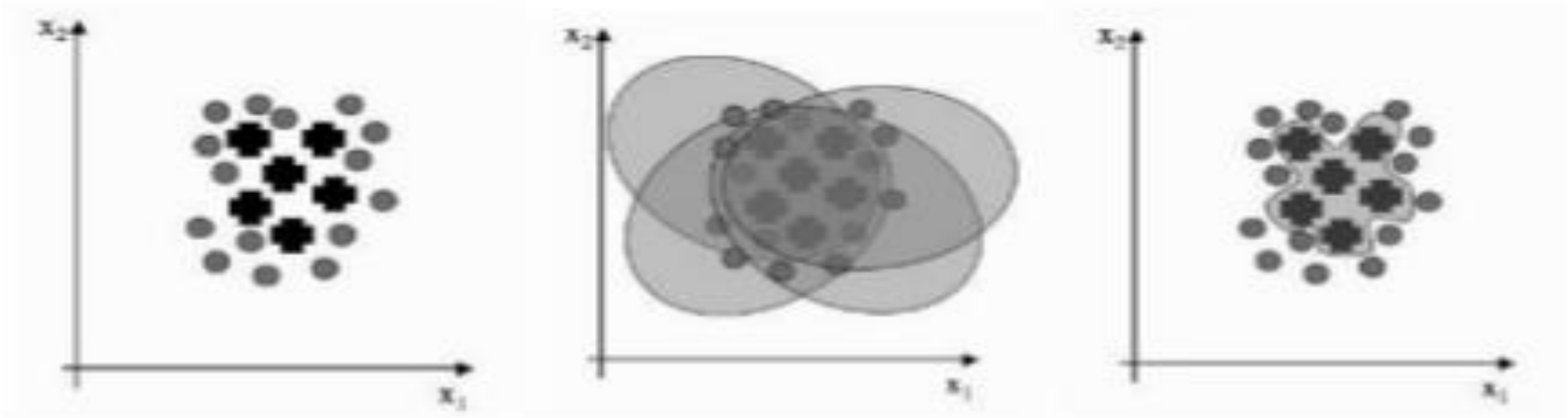


더 좋은 결론도출을 위해선 다양한 배경의 사람들이 필요하다!

->치우친 평가 방지를 위해 다양한 세부 분류기를 가져야 한다!

Ensemble(앙상블)

- 즉, 하나의 만능 알고리즘은 없다.
- 그러므로 서로 다른 여러 개의 모델 또는 학습자를 합성하여 데이터에 대해 다른 결과를 도출
- 여러 학습모델의 결과를 종합해 검정->결과가 비교적 좋다.

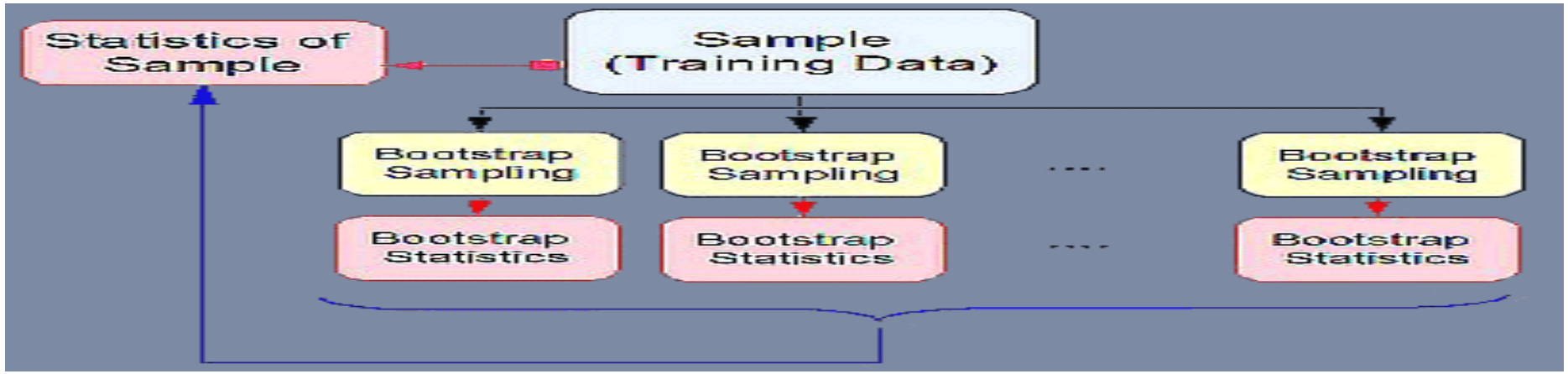


2. Bootstrapping

-앙상블에서 자주 사용하는 샘플 추출 방법

-실제 조사한 결과를 바탕으로 가상의 샘플링을 수행

-수행된 결과를 기반으로 결과의 정확성 평가 및 분포 추정



사용하는 이유는?

1. 데이터를 수집했던 확률변수의 정확한 분포를 모르는 경우
2. 측정된 샘플이 부족한 경우

Q&A) Q1. 딱 하나의 단일 통계치를 얻고자 하면?

A1. 전체의 평균을 구하면 된다.

Q2. 평균의 confidence interval을 구하고 싶다면?

A2. 이 때, bootstrapping을 사용한다.

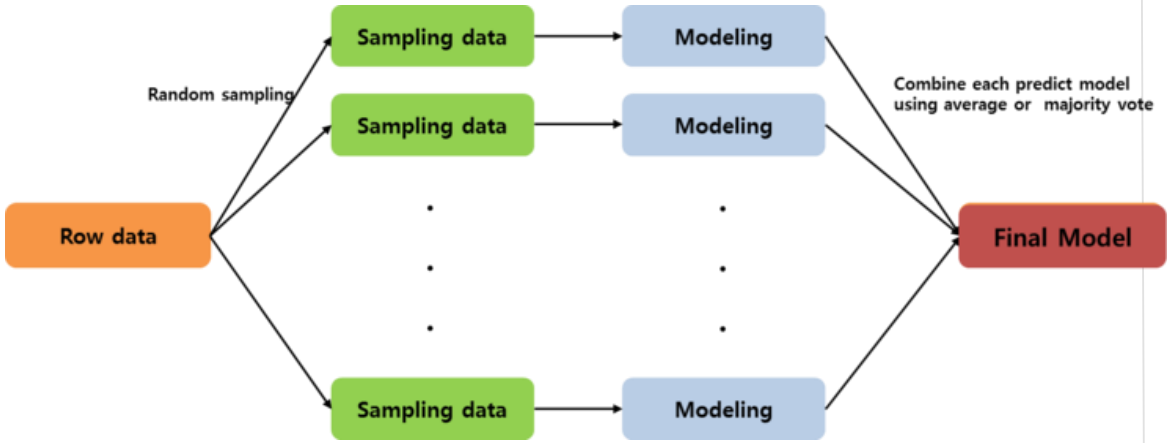
why?

→데이터를 수집했던 확률변수의 정확한 분포를 모르기에 측정된 통계치의 신뢰도를 가늠할 방법이 없다.

그래서 측정된 n개의 데이터 중에서 중복을 허용하여 m개를 뽑고, 그들의 평균을 구하기를 여러 번 반복한다. 그럼 평균의 분포를 구할 수 있게 된다.

3. Bagging(Bootstrap AGGregatING)

- 데이터들로부터 부분 데이터 집합을 형성. 각각에 대해 학습 알고리즘을 적용하여 분류기를 생성.
- 여러 번의 복원 샘플링을 통해 예측모형의 분산을 줄임

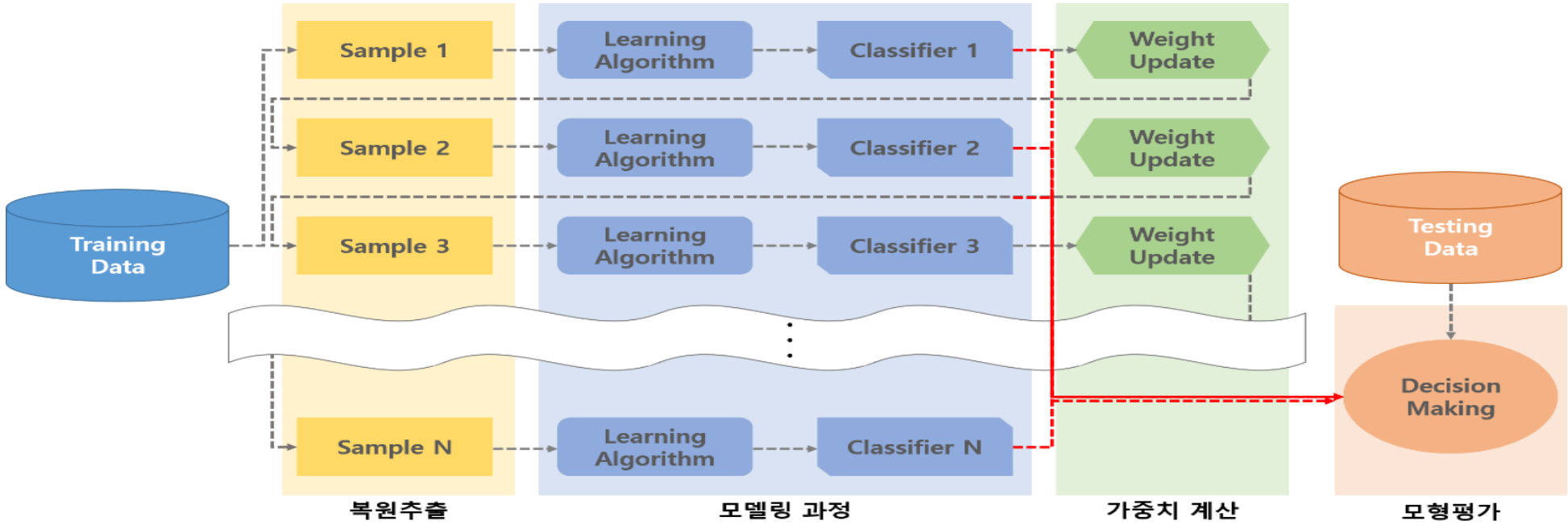


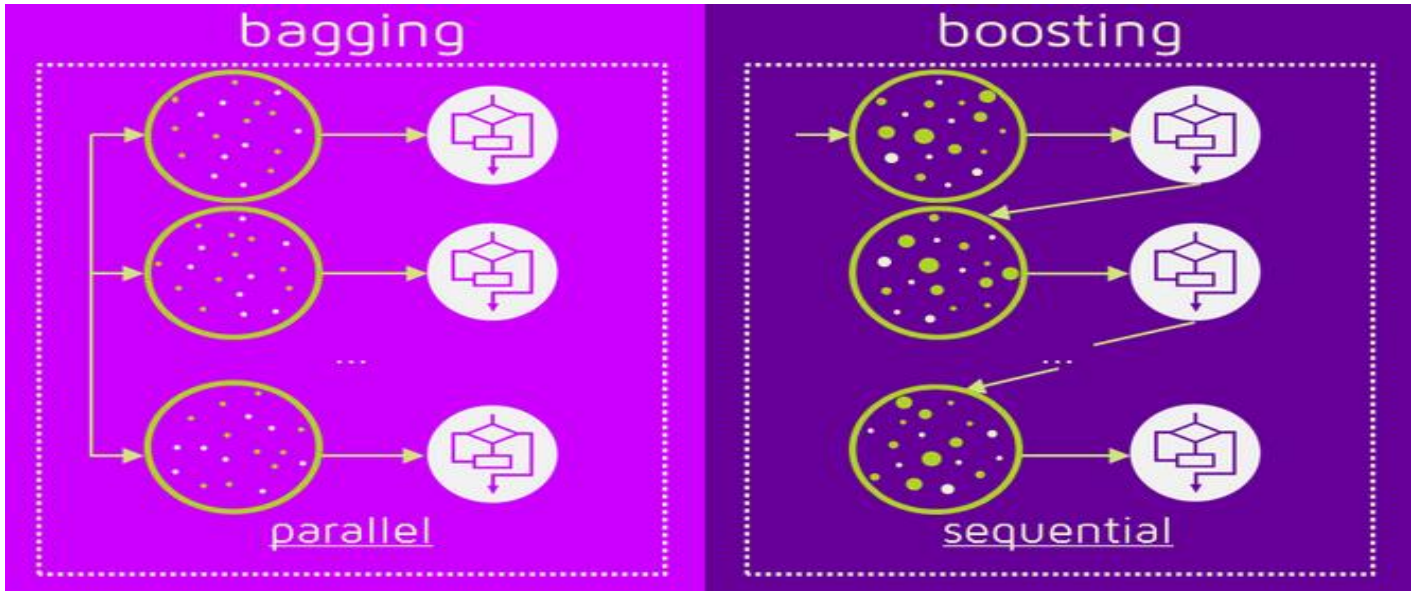
4. Boosting

- 분류기들이 순차적으로 학습하도록 하여, 먼저 학습된 결과가 다름 분류기의 학습에 영향을 줌. 이로써 이전 분류기의 결점을 보완하는 방향.
- 맞추기 어려운 문제를 맞추는 데에 초점.

  (오답에 더 높은 가중치를 부여하여 그에 집중)
- 잘못 분류된 개체들을 더 잘 분류하는 데에 목적을 둠.

4. Boosting


Training Data
Sample 1
Sample 2
Sample 3
Sample N
Learning Algorithm
Learning Algorithm
Learning Algorithm
Learning Algorithm
Classifier 1
Classifier 2
Classifier 3
Classifier N
Weight Update
Weight Update
Weight Update
Testing Data
Decision Making
복원추출
모델링 과정
가중치 계산
모형평가



병렬적 vs 순차적

제일 큰 차이점은 Boosting 의 가중치 분배 때문에 순차적 학습을 시킨 다는 것.

출처 : https://swalloow.github.io

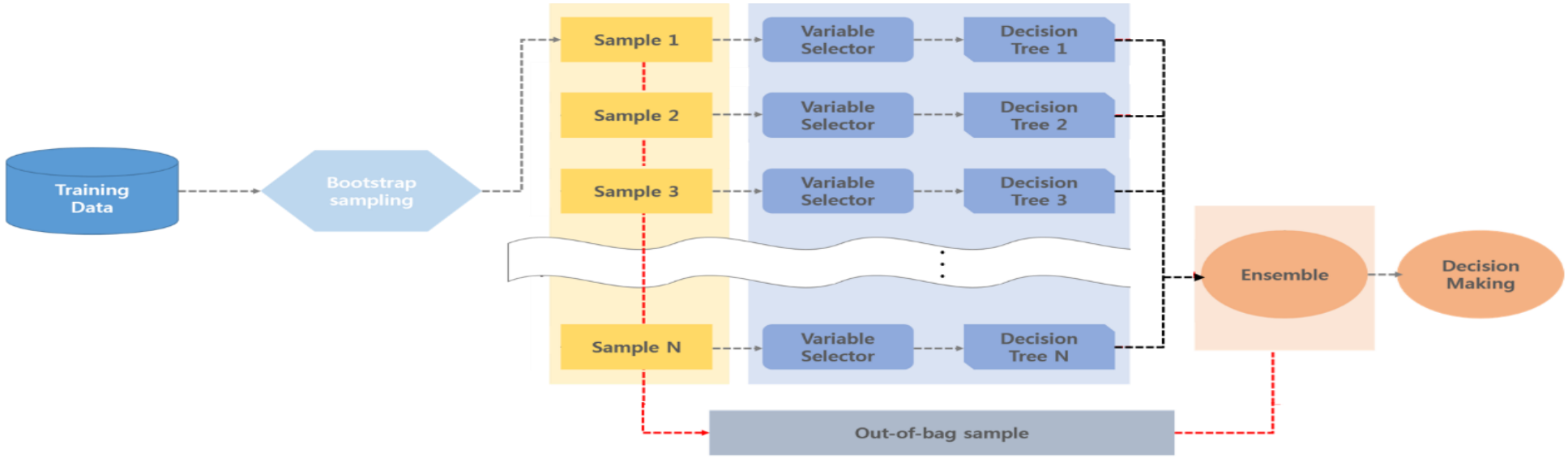
5. Random Forest

- 동일한 데이터로부터 복원 추출을 통해 수많은 데이터 집합을 생성. 이로부터 최종 결과를 도출.
- 변수 선택의 임의성을 더해 예측들이 비상관화.

  —> 일반화 성능 향상시킴

- 여러 번의 복원 샘플링을 통해 예측모형의 분산을 줄임. 비교적 정확.

5. Random Forest



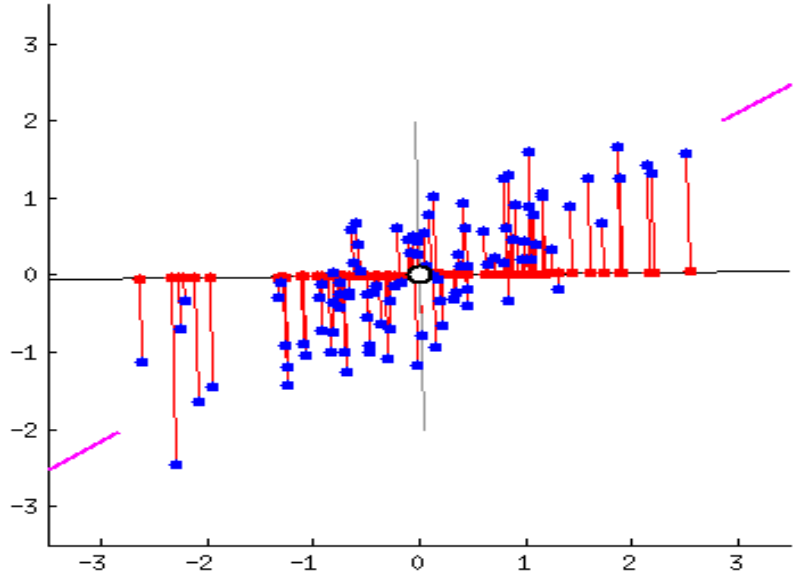
Random Forest

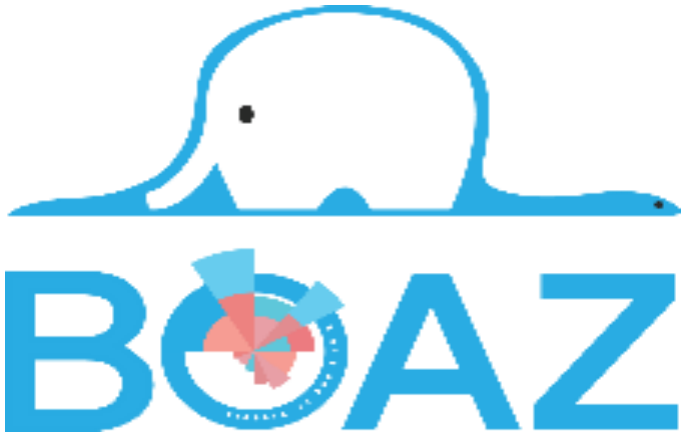
장점	단점
-대부분의 모델에 잘 적용 -노이즈성 데이터를 다루기에 적합 -거대한 데이터들을 다룰 수 있음	-모델이 쉽게 해석되지 않음 -비교적 많은 작업이 요구됨.

+)Rotation Forest

- 주성분분석(PCA)를 적용해 데이터 축을 회전 한 후 학습

출처 : [Making sense of principal component analysis, eigenvectors & eigenvalues]





감사합니다