R을 활용한 머신러닝

5장

분할 정복:

의사 결정 트리와 규칙 기반의 분류

Hyunseok Shin

Bio Information Technology Lab.

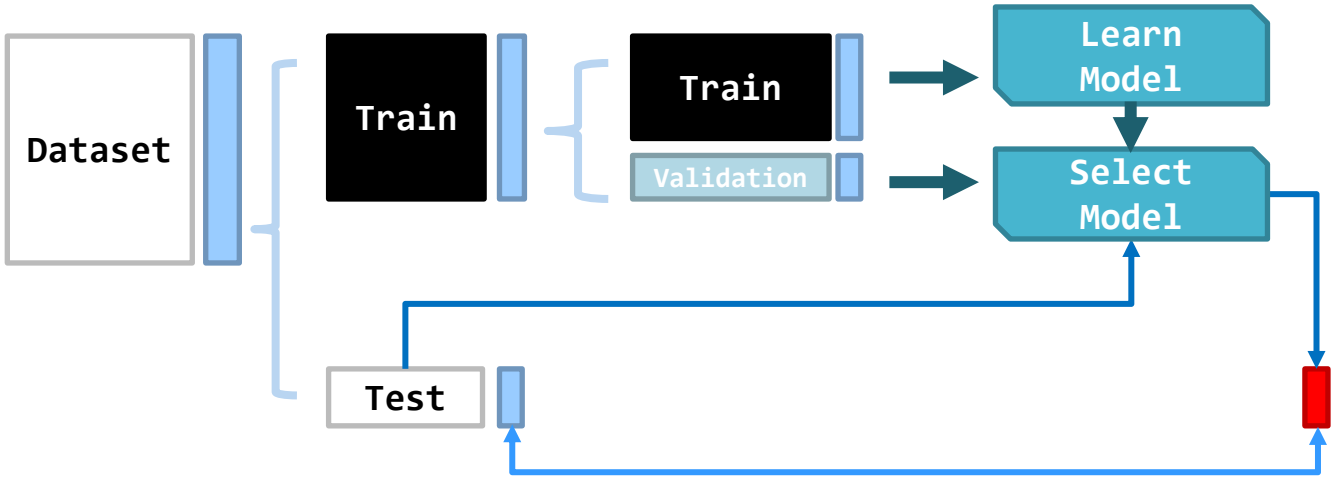
1. 개요

5

Divide and Conquer – Classification Using Decision Trees and Rules

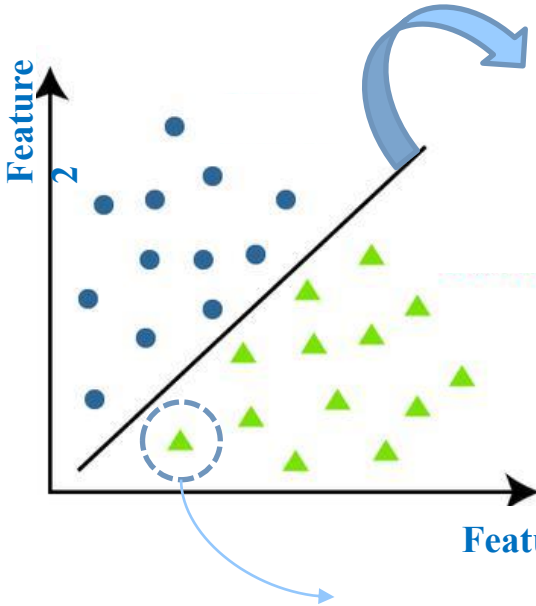
1. 개요

- 모델 개발 과정


Dataset
Train
Train
Validation
Learn Model
Select Model
Test

1. 개요

- 관련 용어
  - 모델(model) = 관계에 대한 가정
  - 학습(Learning) = 가정한 관계를 구체적으로 찾음
  - 분류(Classification) = 어떤 레이블을 가져야 하는지 결정
  - 특징(Feature)과 레이블(Label)



모델 = 하나의 직선이 클래스를 구분 = 관계

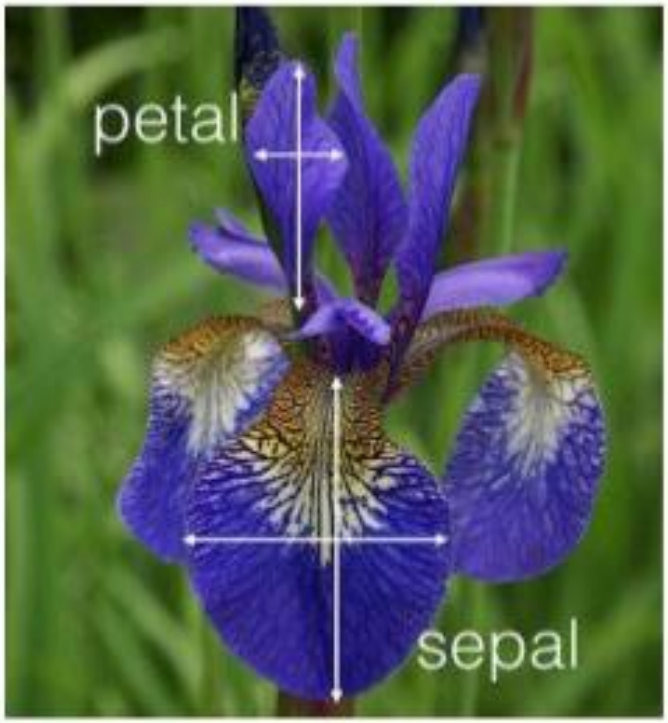
$\Leftrightarrow \quad y = f(x)$

새 데이터에 대해 예측(prediction) 가능

단, 경험 안에서 예측 가능(데이터 = 경험)

1. 개요

- 관련 용어
  - 모델(model) = 관계에 대한 가정
  - 학습(Learning) = 가정한 관계를 구체적으로 찾음
  - 분류(Classification) = 어떤 레이블을 가져야 하는지 결정
  - 특징(Feature)과 레이블(Label)



Training / test data

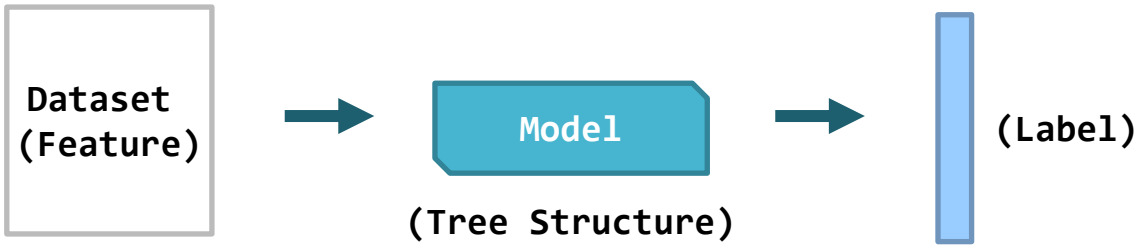
Features				Labels
Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	Iris setosa
4.9	3.0	1.4	0.2	Iris setosa
7.0	3.2	4.7	1.4	Iris versicolor
6.4	3.2	4.5	1.5	Iris versicolor
6.3	3.3	6.0	2.5	Iris virginica
5.8	3.3	6.0	2.5	Iris virginica

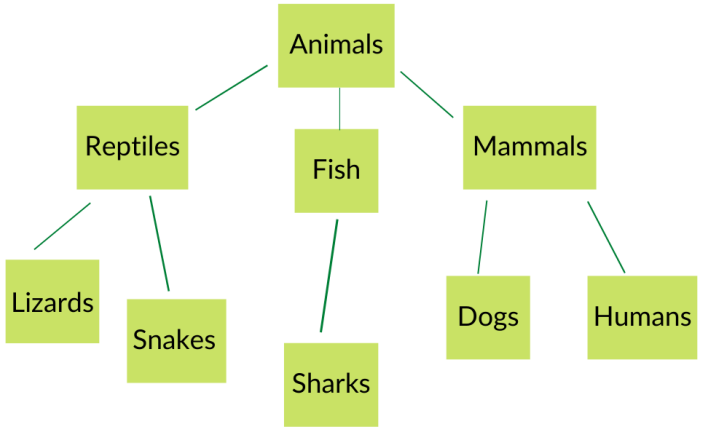
1. 의사 결정 트리

Part 1. 의사 결정 트리의 이해

Tree Structure

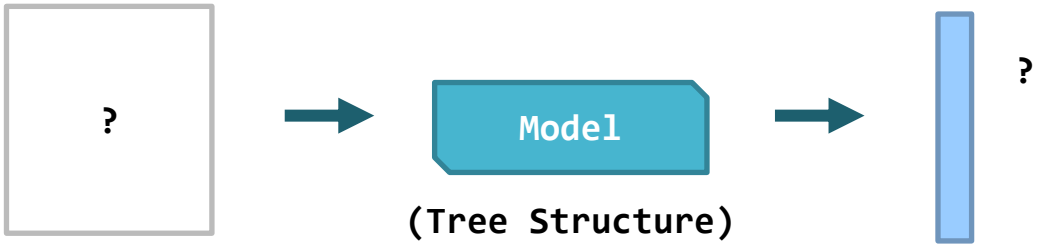
- Decision Tree
  - 단순한 선택 과정을 거쳐 복잡한 결정 → 형태가 Tree와 같음
  - 결정을 하기 위해서는 기준 필요(= 규칙)

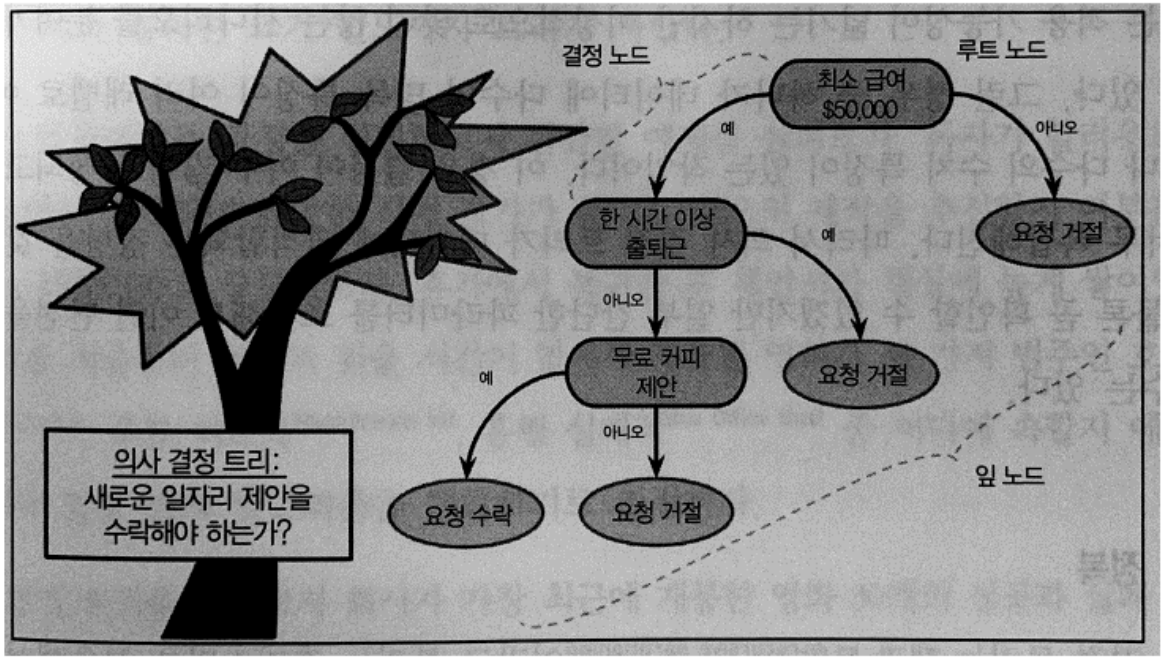




Ex) 일자리 제안 수락 여부

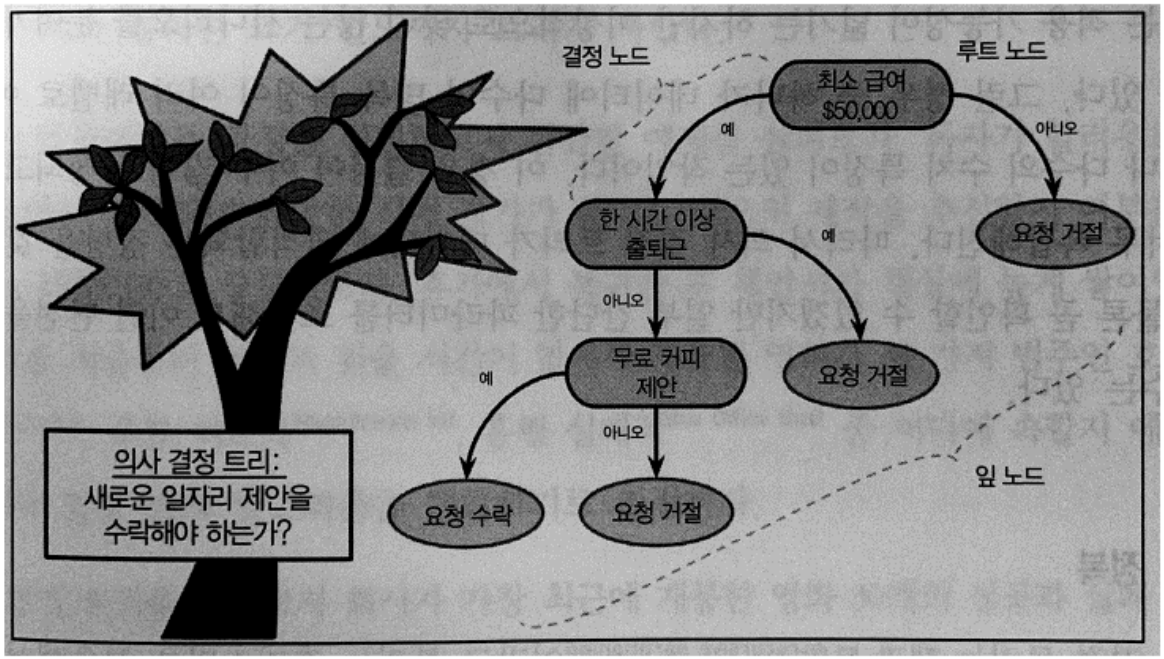
- 새로운 일자리 제안을 수락해야 하는가?
  - Feature? Input data? Output?
  - Root node, Decision node, Leaf node?





의사 결정 트리가 좋은 이유

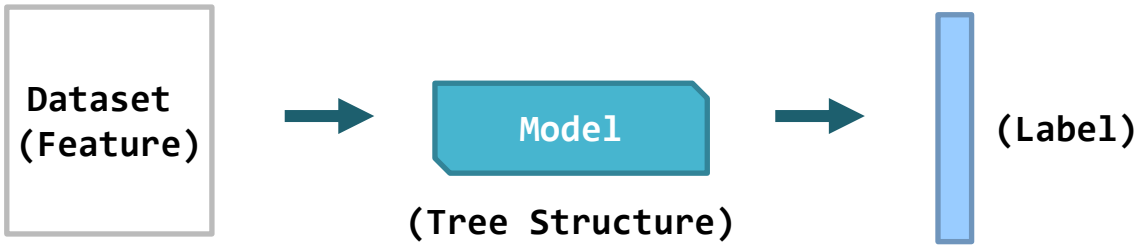
- 결정 과정에 대한 해석 가능성
  - 사람이 읽을 수 있는 구조
  - 분류 방법이 투명
  - 신용평가 모델, 질병 예측모델 등

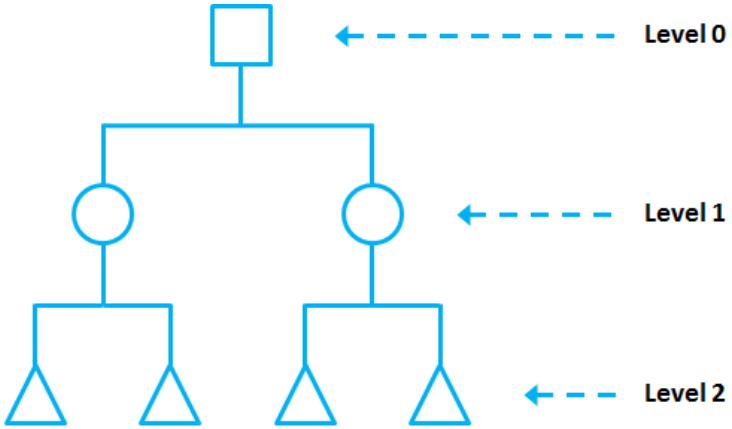


Part 2. 트리는 어떻게 만들까?

Divide and Conquer

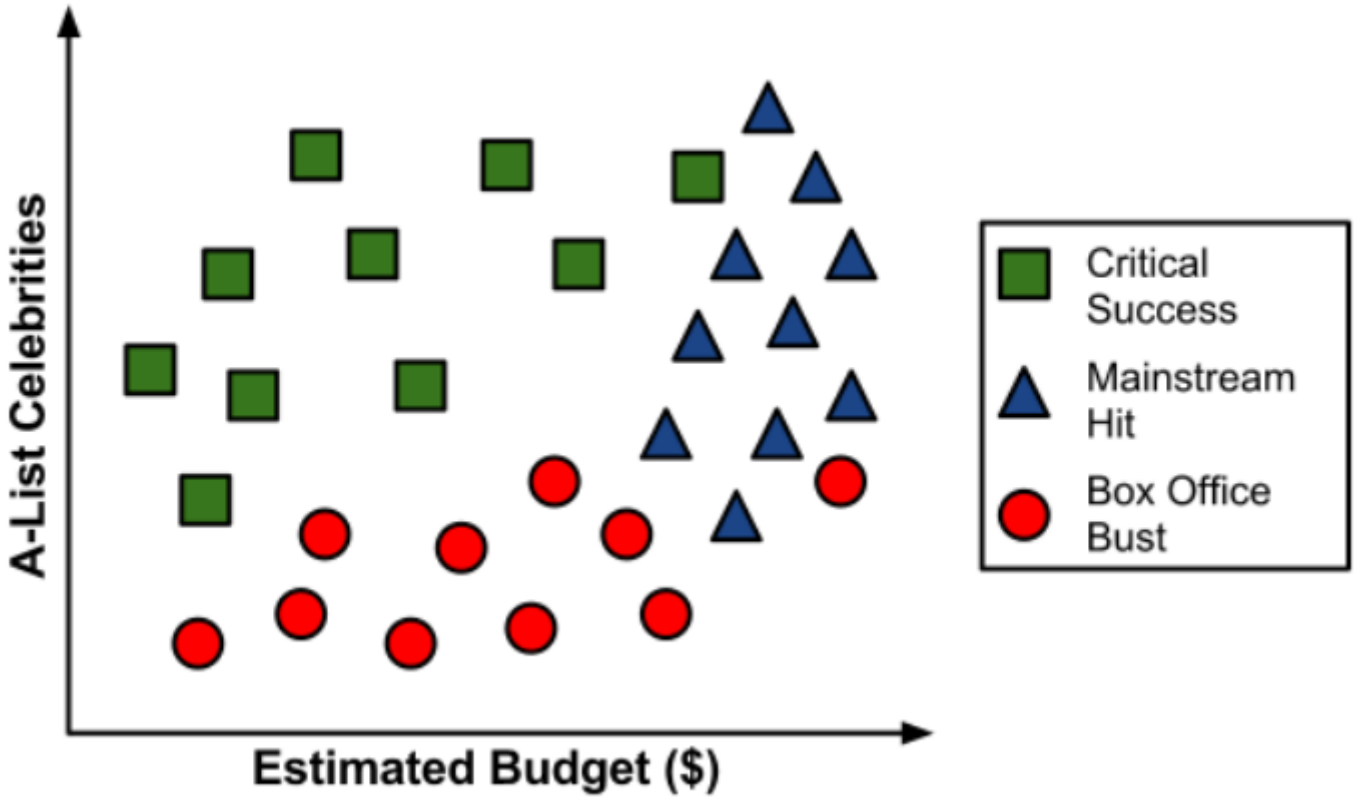
- **동질적**인 집합이 되도록 반복해서 분할
  - 어떤 특징(feature)을 선택할 것인가? (= 특징 선택 기준)
  - 언제까지 반복 수행할 것인가? (= 종료 기준)





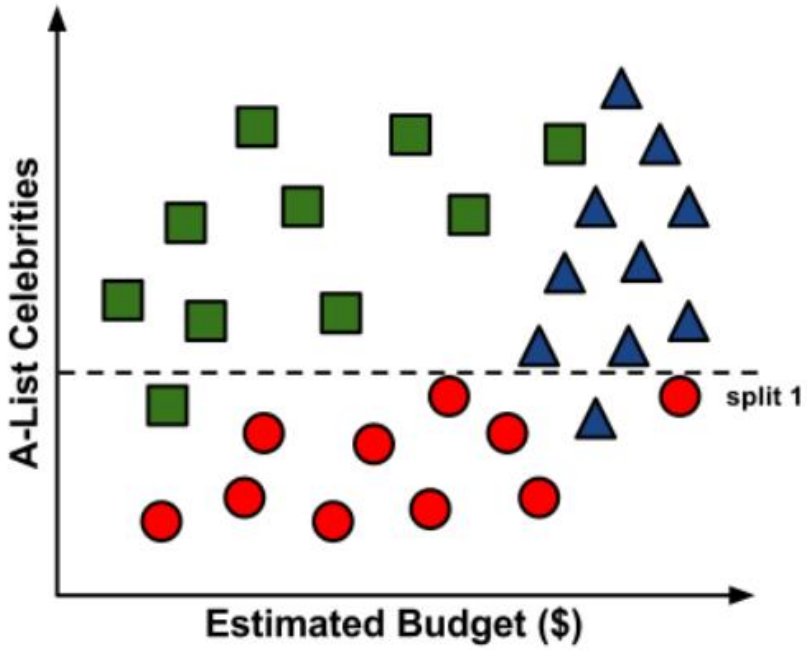
Ex)트리가 만들어지는 과정

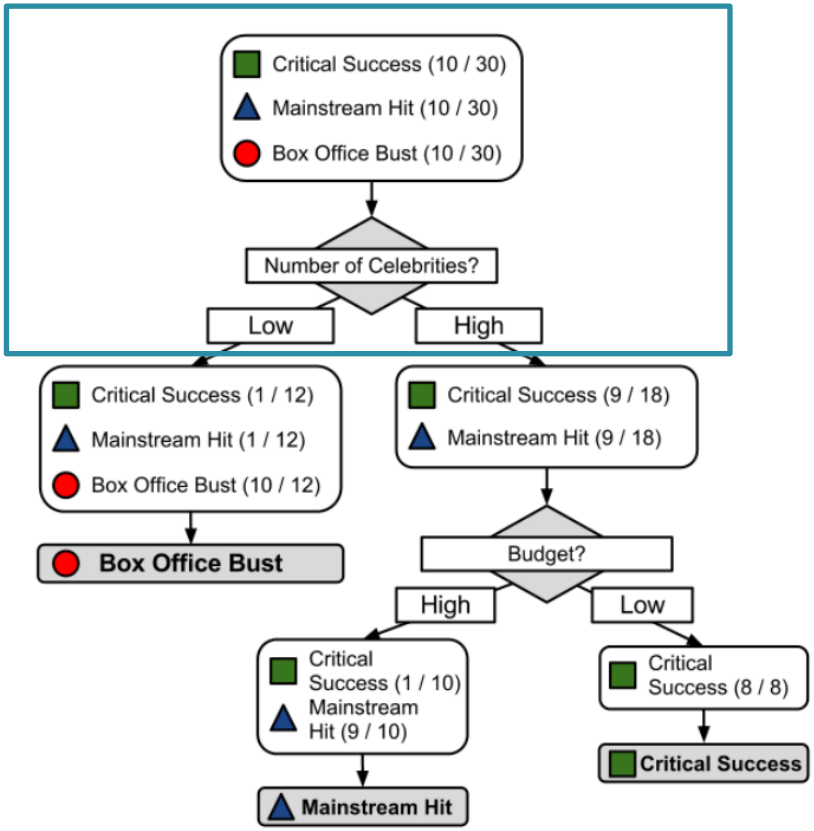
- 의사 결정 트리를 만들고자 최근 개봉 영화 30개의 성공 실패요인 분석
  - Feature? Label?



Ex)트리가 만들어지는 과정

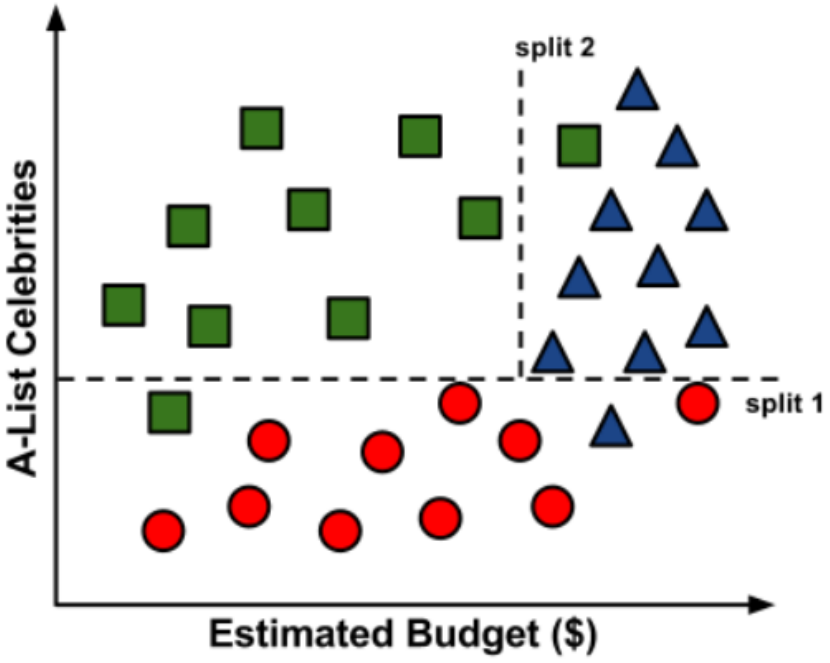
- 유명인 수 많고 적음(feature 1)

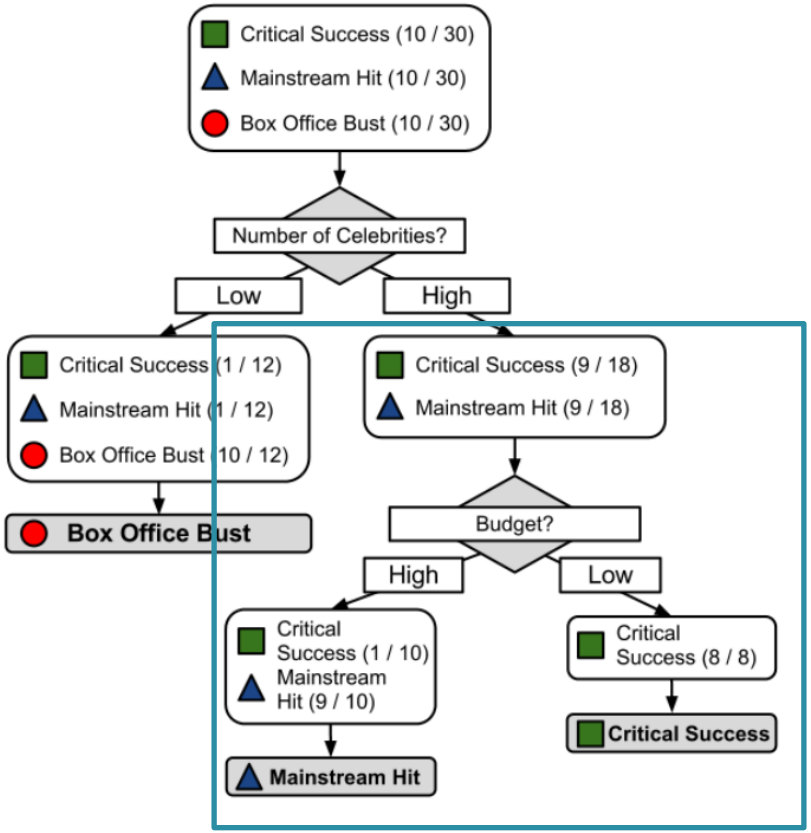




Ex)트리가 만들어지는 과정

- 예산 많고 적음(feature 2)





Part 3. C5.0 알고리즘

C5.0

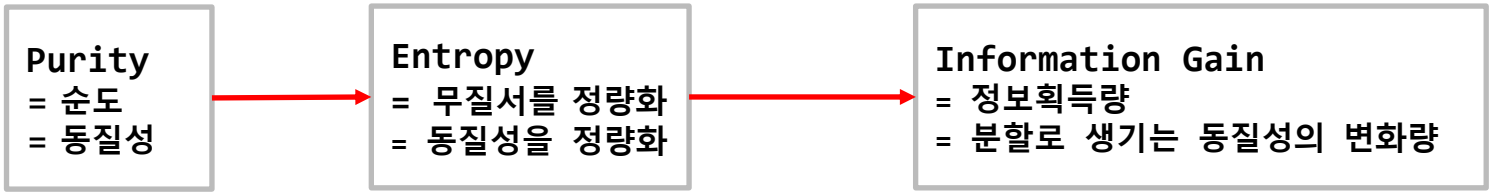
- 가장 잘 알려진 의사결정트리 알고리즘
  - J. Ross Quinlan 구현

[장점]
- 수치, 명목, 누락 데이터를 다룰 수 있음
- 중요하지 않은 특징 제외
- 수학적 배경 없이 해석할 수 있는 모델

[단점]
- 평행 분할에 의존
- 레벨 수가 많은 특징의 분할로 편향될 수 있음

Divide and Conquer

- **동질적**인 집합이 되도록 반복해서 분할
  - 어떤 특징(feature)을 선택할 것인가? (= 특징 선택 기준)



  - 언제까지 반복 수행할 것인가? (= 종료 기준)
    - 조기 종료(early stopping) = 사전 가지치기(pre-pruning)
    - 사후 가지치기(post-pruning)

Divide and Conquer

- **동질적**인 집합이 되도록 반복해서 분할
  - 어떤 특징(feature)을 선택할 것인가? (= 특징 선택 기준)

Purity
= 순도
= 동질성

→

Entropy
= 무질서를 정량화
= 동질성을 정량화

→

Information Gain
= 정보획득량
= 분할로 생기는 동질성의 변화량

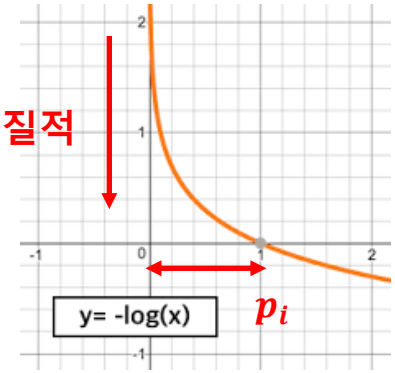
∗ ***Entropy*** 가 작다 = 동질성이 크다

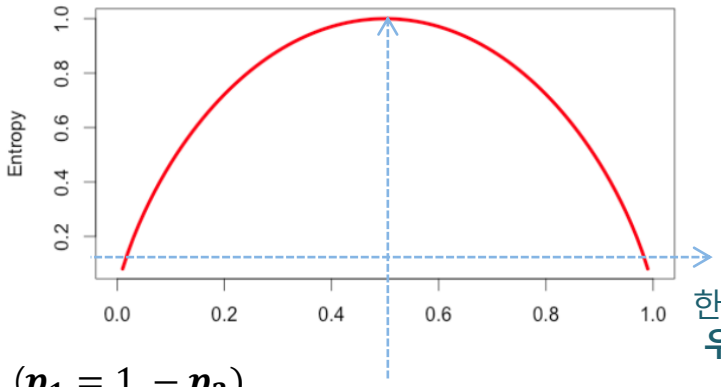
∗ ***InfoGain*** 이 크다 = 동질성이 증가했다

$$\text{Entropy}(S) = \sum_{i=1}^{c} -p_i \log_2(p_i)$$

$$\text{InfoGain}(F) = \text{Entropy}(S_1) - \text{Entropy}(S_2)$$

$$\text{Entropy}(S) = \sum_{i=1}^{n} w_i \, \text{Entropy}(P_i)$$





한쪽 클래스가 우세할 수록

$p_1 = p_2$ 일 때

ex) 빨간공 6개, 흰 공 4개 ($p_1 = 1 - p_2$)

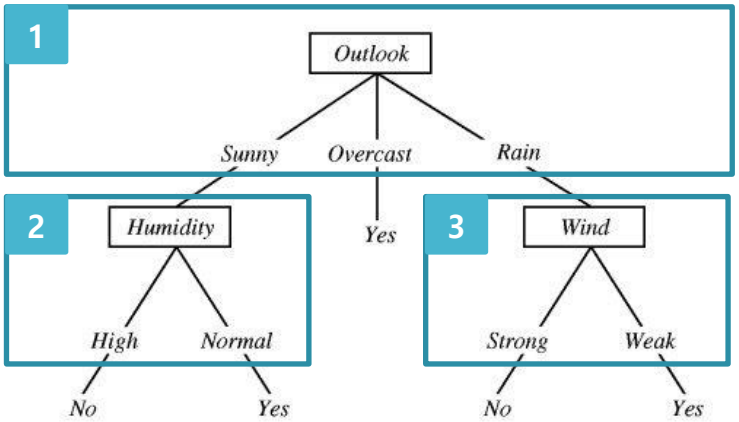
$= -p_1 \log_2(p_1) - p_2 \log_2(p_2)$

$= -0.6 * \log_2(0.6) - 0.4 * \log_2(0.4)$

$= 0.9709506$

#Homework: Play Tennis

- Entropy와 Information Gain 계산



Dataset for Play Tennis

Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

1) 1번에서의 Information Gain을 구하시오

+ outlook의 엔트로피를 구하시오.
+ sunny의 엔트로피를 구하시오.
+ overcast의 엔트로피를 구하시오.
+ rain의 엔트로피를 구하시오.

2) 2번에서의 Information Gain을 구하시오

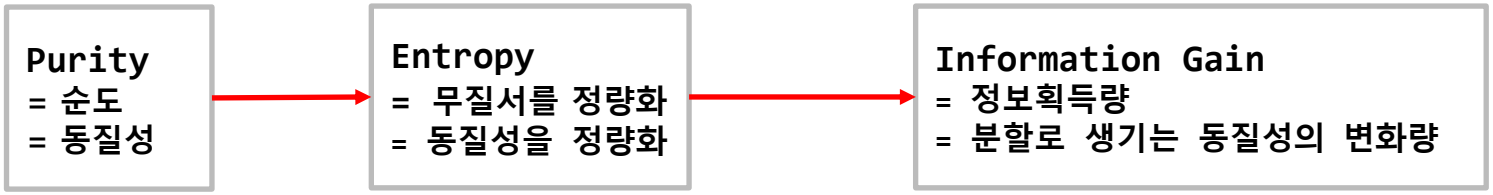
+ humidity의 엔트로피를 구하시오.
+ high의 엔트로피를 구하시오.
+ noraml의 엔트로피를 구하시오.

3) 3번에서의 Information Gain을 구하시오

+ wind의 엔트로피를 구하시오.
+ strong의 엔트로피를 구하시오.
+ weak의 엔트로피를 구하시오.

Divide and Conquer

- **동질적**인 집합이 되도록 반복해서 분할
  - 어떤 특징(feature)을 선택할 것인가? (= 특징 선택 기준)



  - 언제까지 반복 수행할 것인가? (= 종료 기준)
    - 조기 종료(early stopping) = 사전 가지치기(pre-pruning)
    - 사후 가지치기(post-pruning)

Pruning
= 모델 일반화가 목적

Pre-Pruning
= 결정이 특정 개수에 도달 시
= 결정 노드에 포함된 셋이 작을 때

불필요한 작업을 줄여 줌
최적의 모델이 아닐 수 있음
(트리의 깊이를 사전에 알 수 없기 때문)

Post-Pruning
= 결정이 특정 개수에 도달 시
= 결정 노드에 포함된 셋이 작을 때

Part 4. Code

2. 분류 규칙

Part 1. 분류 규칙의 이해

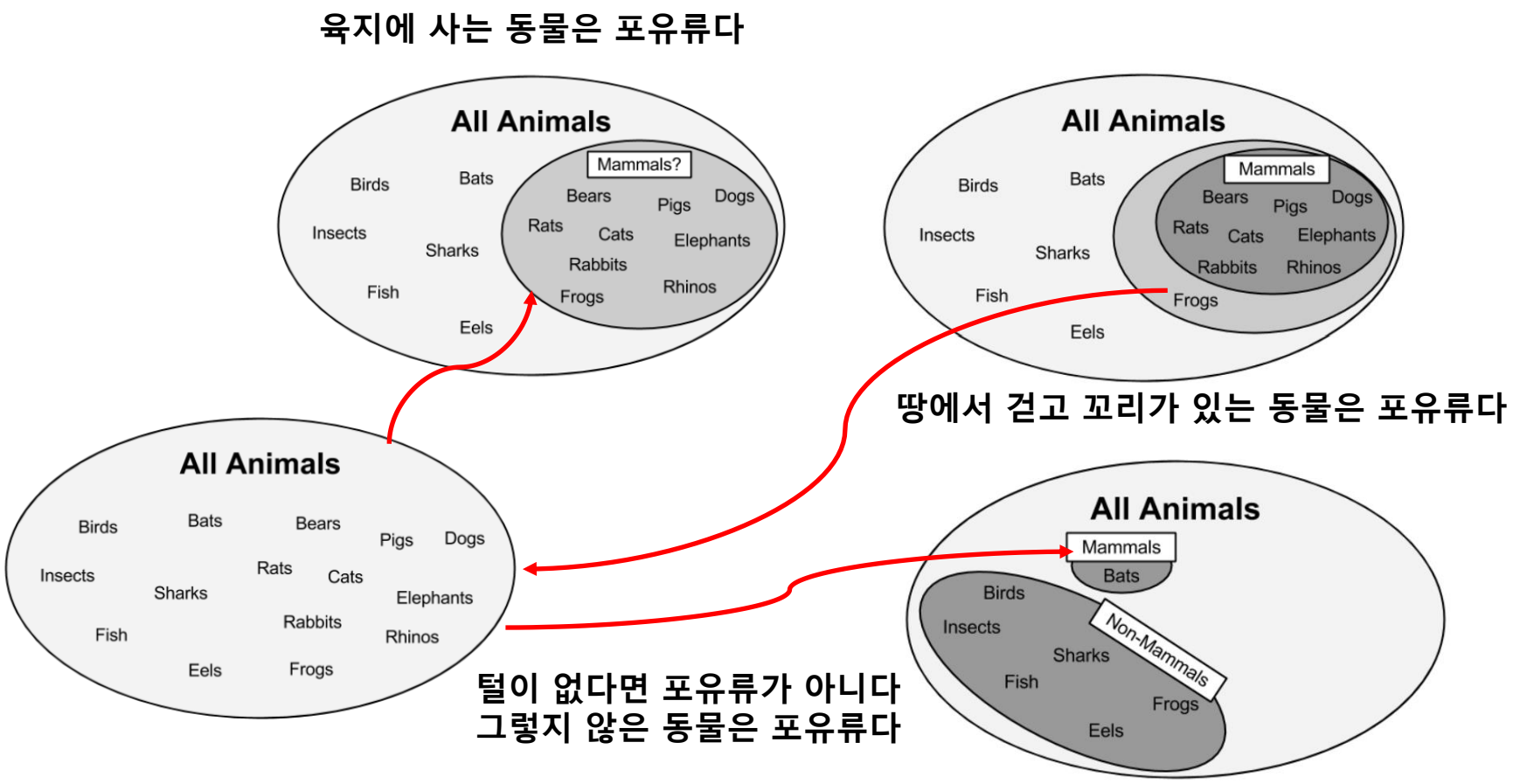
Classification Rules

- 트리와 미묘한 차이 존재
  - Separate and Conquer <-> Divide and Conquer
  - 퍼 온 다음 원하지 않는 것 다시 통에 넣기
  - 퍼 온 것은 다시 통에 넣지 않기

~이면(조건부), ~이다(결론부)

Separate and Conquer

- 트리와 미묘한 차이 존재
  - Separate and Conquer <-> Divide and Conquer



Part 2. 알고리즘

1R (One Rule)

- 가장 작은 오분류율을 가진 단일 규칙을 선택
  - 이동경로 특징이 오류가 적음
    - 동물이 하늘로 이동하면 포유류가 아니다
    - 동물이 땅으로 이동하면 포유류다
    - 동물이 바다로 이동하면 포유류가 아니다

Animal	Travels By	Has Fur	Mammal
Bats	Air	Yes	Yes
Bears	Land	Yes	Yes
Birds	Air	No	No
Cats	Land	Yes	Yes
Dogs	Land	Yes	Yes
Eels	Sea	No	No
Elephants	Land	No	Yes
Fish	Sea	No	No
Frogs	Land	No	No
Insects	Air	No	No
Pigs	Land	No	Yes
Rabbits	Land	Yes	Yes
Rats	Land	Yes	Yes
Rhinos	Land	No	Yes
Sharks	Sea	No	No

Travels By	Predicted	Actual	
Air	No	Yes	×
Air	No	No	
Air	No	No	
Land	Yes	Yes	
Land	Yes	Yes	
Land	Yes	Yes	
Land	Yes	Yes	
Land	Yes	No	×
Land	Yes	Yes	
Land	Yes	Yes	
Land	Yes	Yes	
Land	Yes	Yes	
Sea	No	No	
Sea	No	No	
Sea	No	No	

Rule for Travels By:
Errors = 2 / 15

Has Fur	Predicted	Actual	
No	No	No	
No	No	No	
No	No	Yes	×
No	No	No	
No	No	No	
No	No	No	
No	No	Yes	×
No	No	Yes	×
No	No	No	
Yes	Yes	Yes	
Yes	Yes	Yes	
Yes	Yes	Yes	
Yes	Yes	Yes	
Yes	Yes	Yes	
Yes	Yes	Yes	

Rule for Has Fur:
Errors = 3 / 15

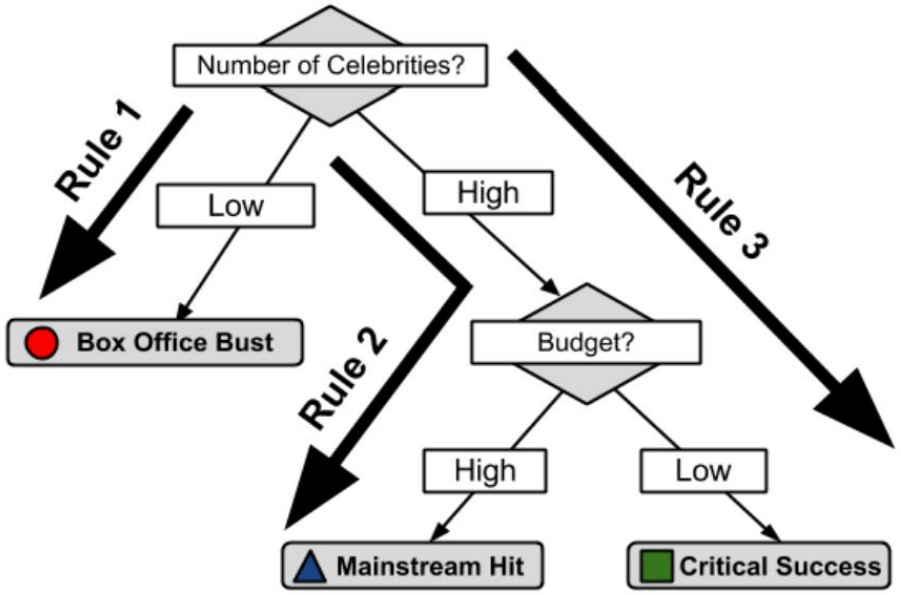
RIPPER

- 속도와 노이즈에 대한 정확도 향상을 위해
  - 1. 복잡한 규칙을 기르고
  - 2. 인스턴스 분리 전 가지치기
  - 3. 최적화

  - 분할 시 정보획득량 기준
  - 엔트로피가 더 이상 줄지 않을 때 가지치기
  - 1과 2단계를 종료 조건에 도달할 때까지 반복

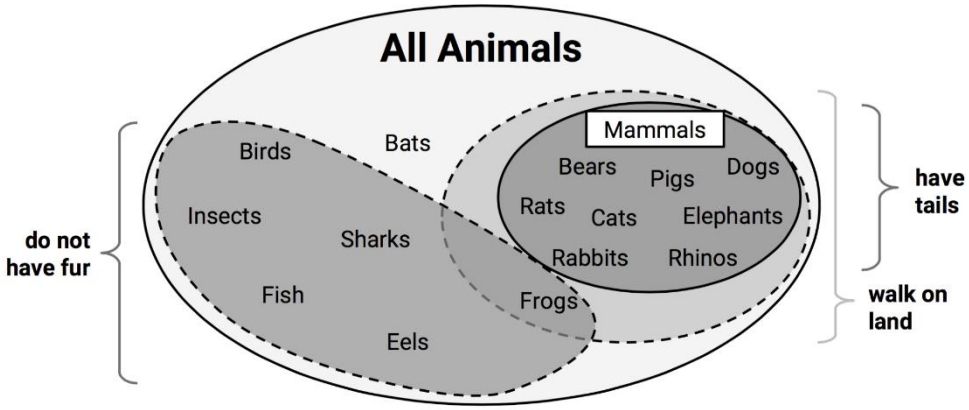
의사결정 트리에서 규칙 구성

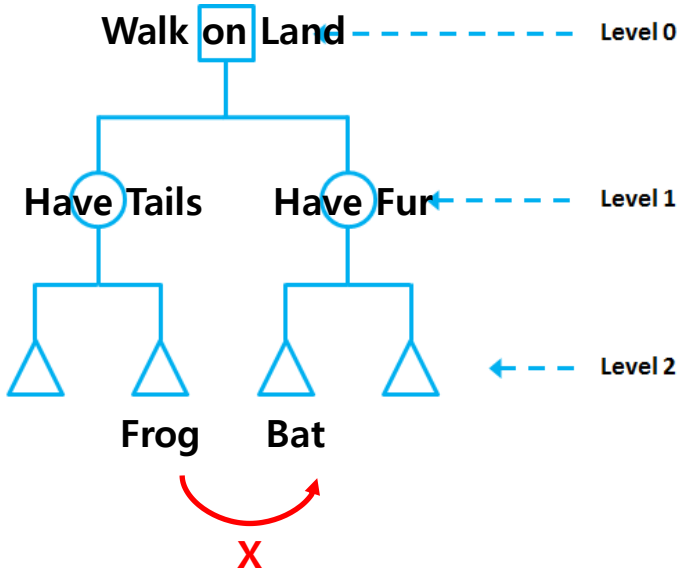
- 생성된 규칙이 규칙 학습 알고리즘으로 학습된 규칙보다 더 복잡



의사결정 트리에서 규칙 구성

- 생성된 규칙이 규칙 학습 알고리즘으로 학습된 규칙보다 더 복잡





오늘도 한걸음

수고했습니다