디지털미디어랩 머신러닝 여름캠프 3주차

(6) Classification과 Decision Tree

YONSEI
UNIVERSITY
DIGITAL
MEDIA LAB.

목차

- Classification (분류) 문제
- Decision Tree
- Iris 데이터 실습

Classification (분류)

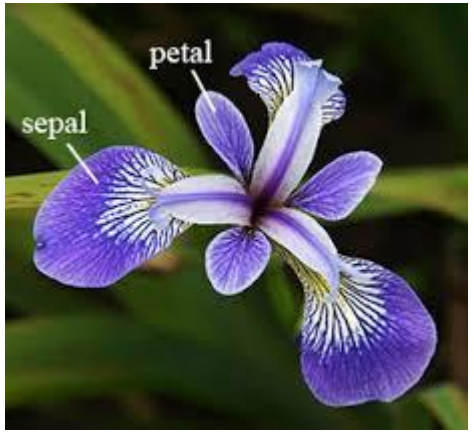
X (hours)	y (P/NP)
1	NP
2	NP
9	P
10	P

X (hours)	y (grades)
1	D
5	C
7	B
10	A

- 목적 변수가 명목형(Nominal) 인 것

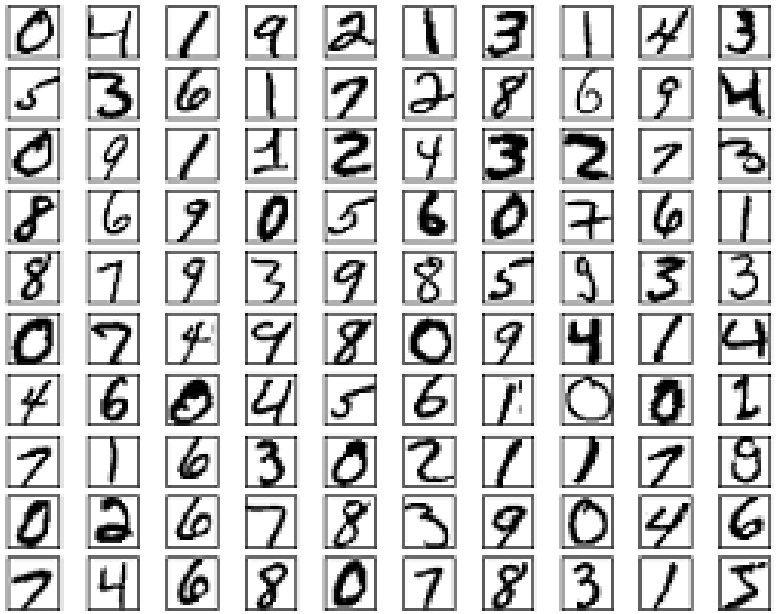
Classification (분류)

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa
12	4.8	3.4	1.6	0.2	setosa
13	4.8	3.0	1.4	0.1	setosa
14	4.3	3.0	1.1	0.1	setosa
15	5.8	4.0	1.2	0.2	setosa



- 각 데이터가 어떤 꽃의 종류인지 알려준다.
- X : sepal length, sepal width, petal length, petal width
- y : 꽃의 종류 (Setosa, Virginica, Versicolor)

Classification (분류)



- 각 이미지 데이터가 어떤 숫자를 의미하는 지 알려준다.
- X : 이미지 데이터
- y : 숫자

Classification (분류)

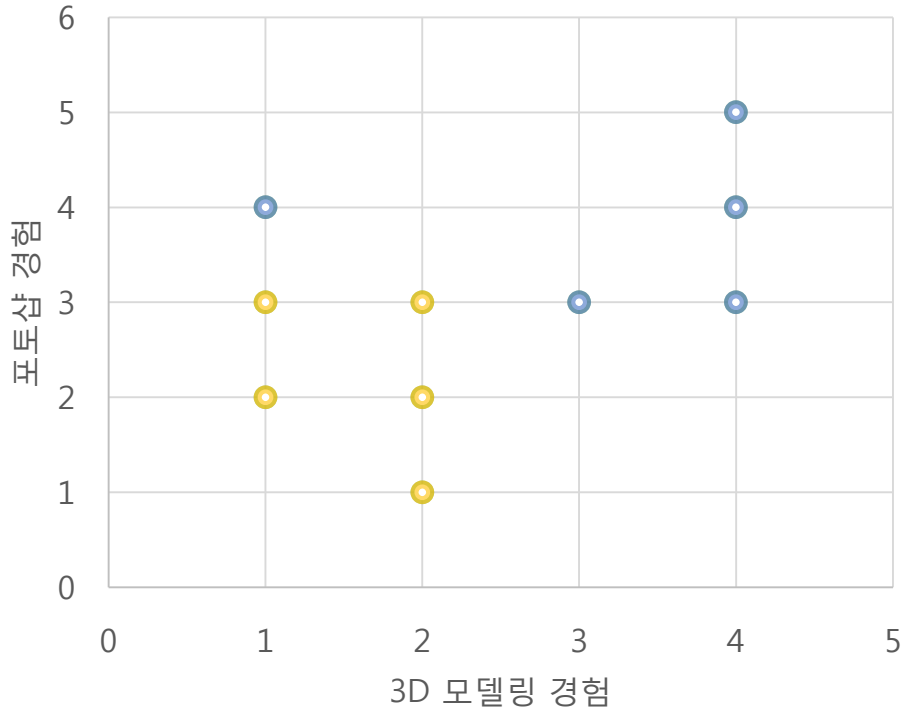
- 스팸메일
- 페이스북 피드 (show, hide)
- 신용카드/계좌이체 사기 탐지 등

Example

ID	3d 모델링 경험	포토샵 경험	디자인전공
디자인1	3	3	True
디자인2	4	4	True
디자인3	4	5	True
디자인4	1	4	True
디자인5	4	3	True
일반1	2	2	False
일반2	2	1	False
일반3	1	3	False
일반4	1	2	False
일반5	2	3	False

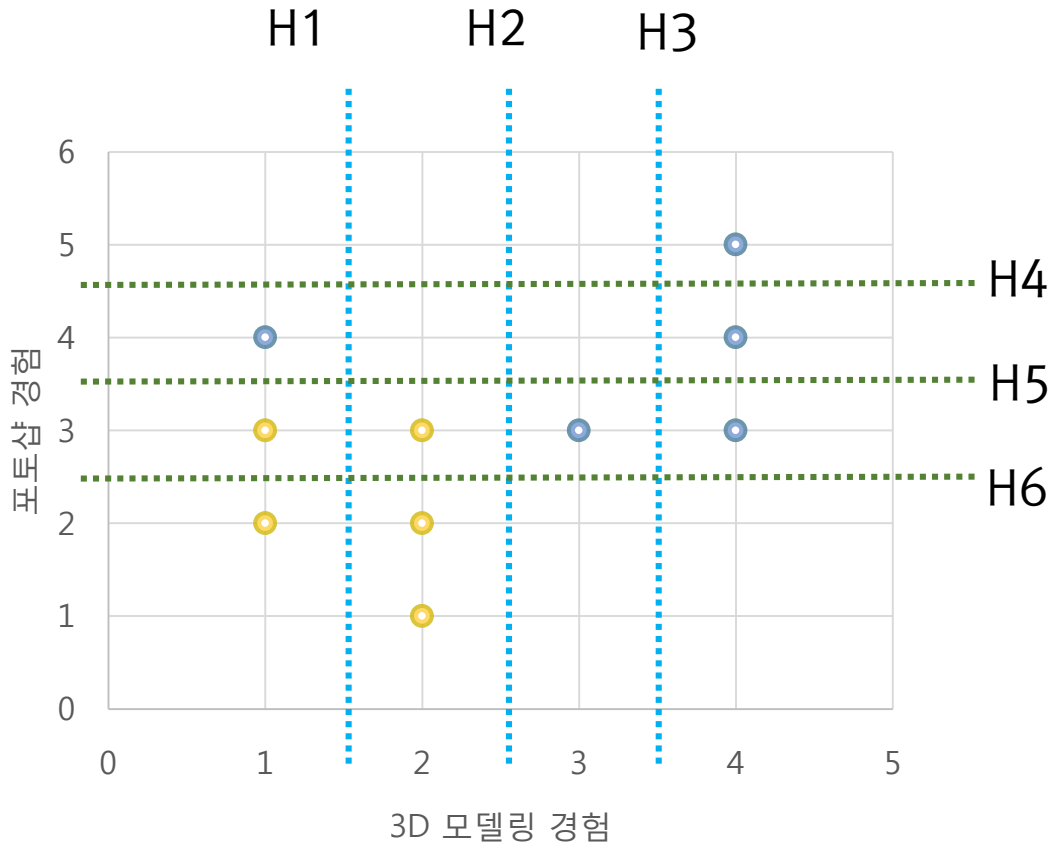
Example

ID	3d 모델링 경험	포토샵 경험	디자인전공
디자인1	3	3	True
디자인2	4	4	True
디자인3	4	5	True
디자인4	1	4	True
디자인5	4	3	True
일반1	2	2	False
일반2	2	1	False
일반3	1	3	False
일반4	1	2	False
일반5	2	3	False



- 디자인 전공자와 디자인 전공자가 아닌 사람을 어떻게 구별할까?

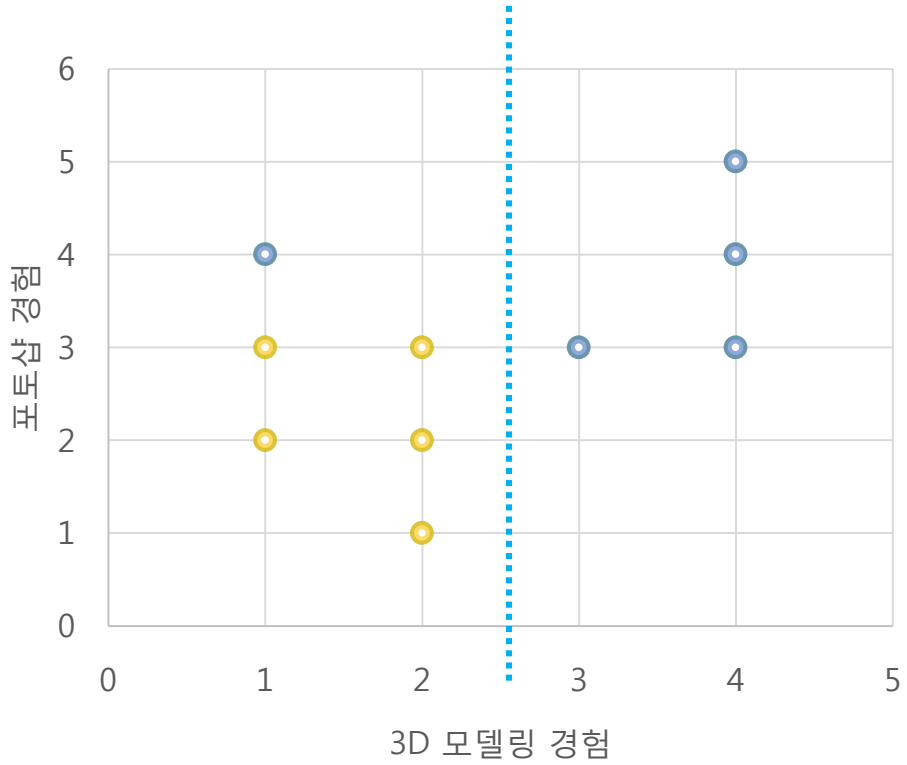
Example

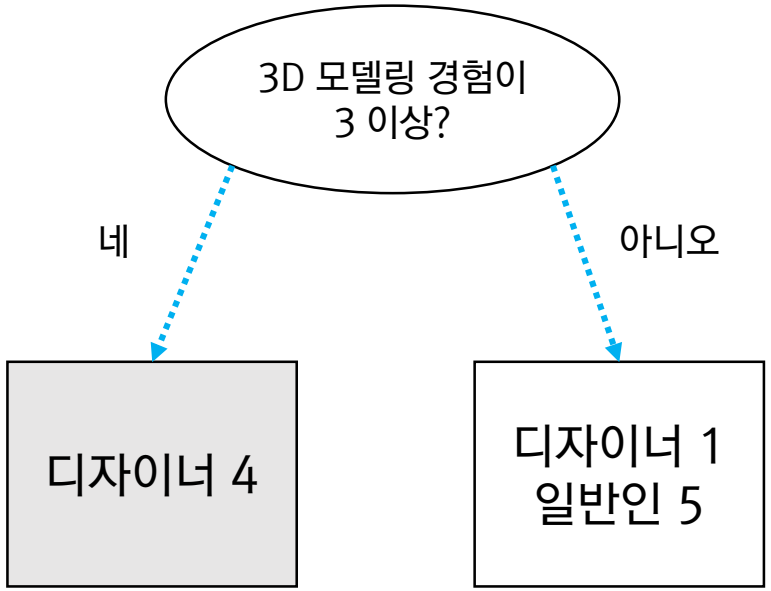


- 전공자와 비전공자를 가장 잘 구분하는 수직, 수평선을 찾는다.

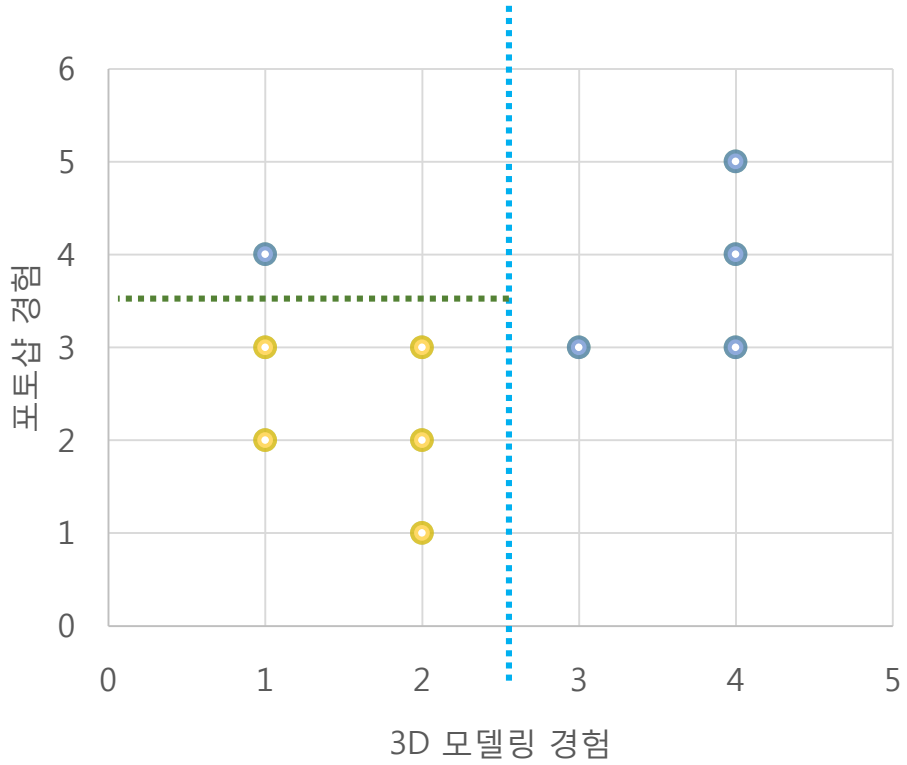
	네	아니오
H1 3D 〉= 2	디4, 일3	디1, 일2
H2 3D 〉= 3	디4	디1, 일5
H3 3D 〉= 4	디3	디2, 일5
H4 포토샵 〉= 4	디1	디4, 일5
H5 포토샵 〉= 3	디3	디2, 일5
H6 포토샵 〉= 2	디5, 일2	일3

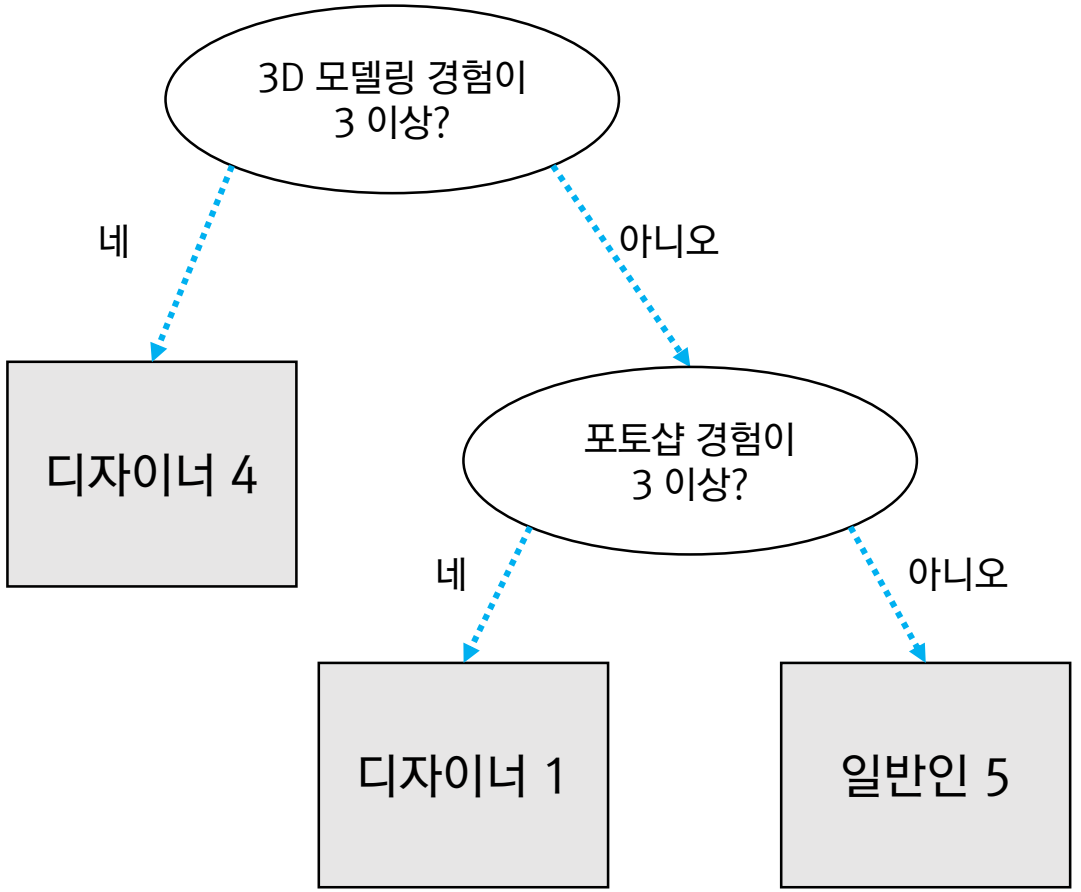
Example


6
5
4
3
2
1
0
포토샵 경험
0
1
2
3
4
5
3D 모델링 경험

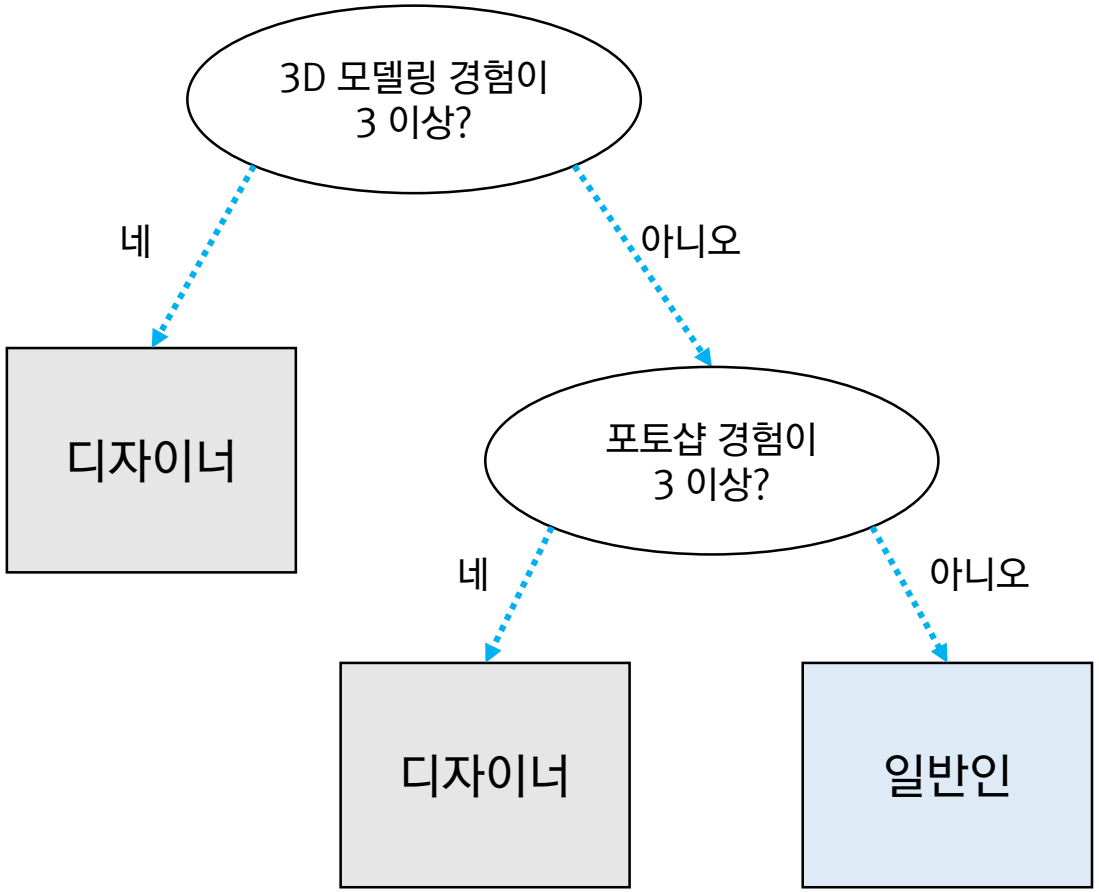

3D 모델링 경험이
3 이상?
네
아니오
디자이너 4
디자이너 1
일반인 5

Example

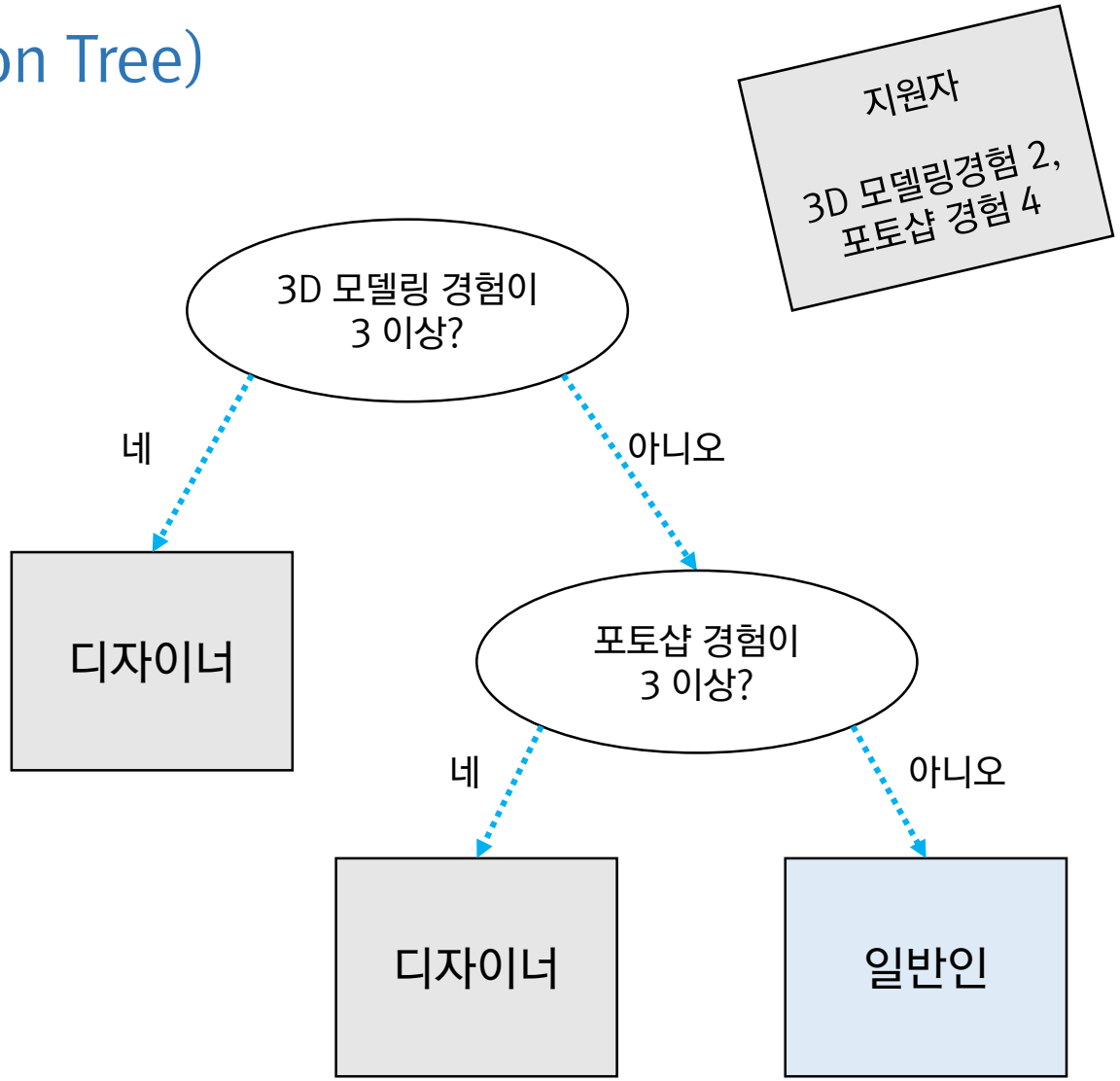

6
5
4
3
2
1
0
포토샵 경험
0
1
2
3
4
5
3D 모델링 경험


3D 모델링 경험이
3 이상?
네
아니오
디자이너 4
포토샵 경험이
3 이상?
네
아니오
디자이너 1
일반인 5

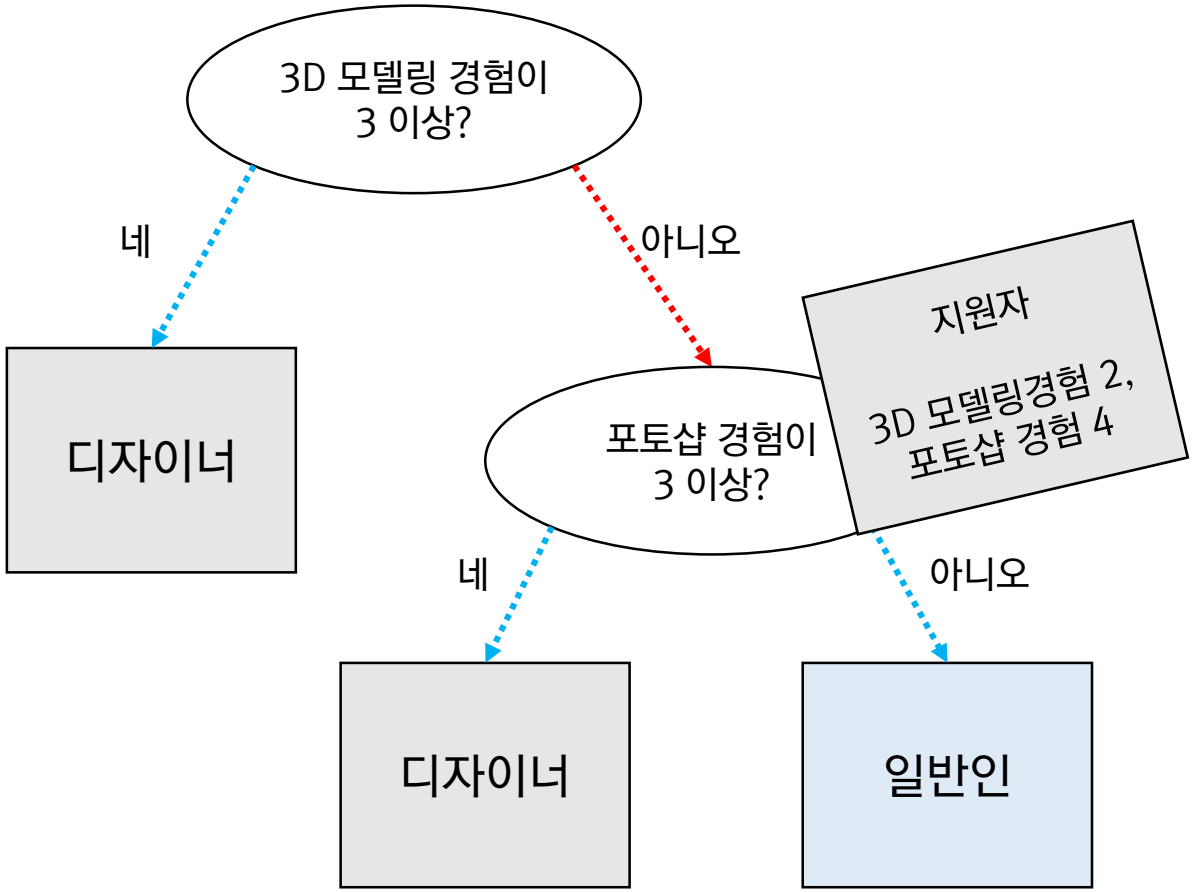
결정 트리 (Decision Tree)



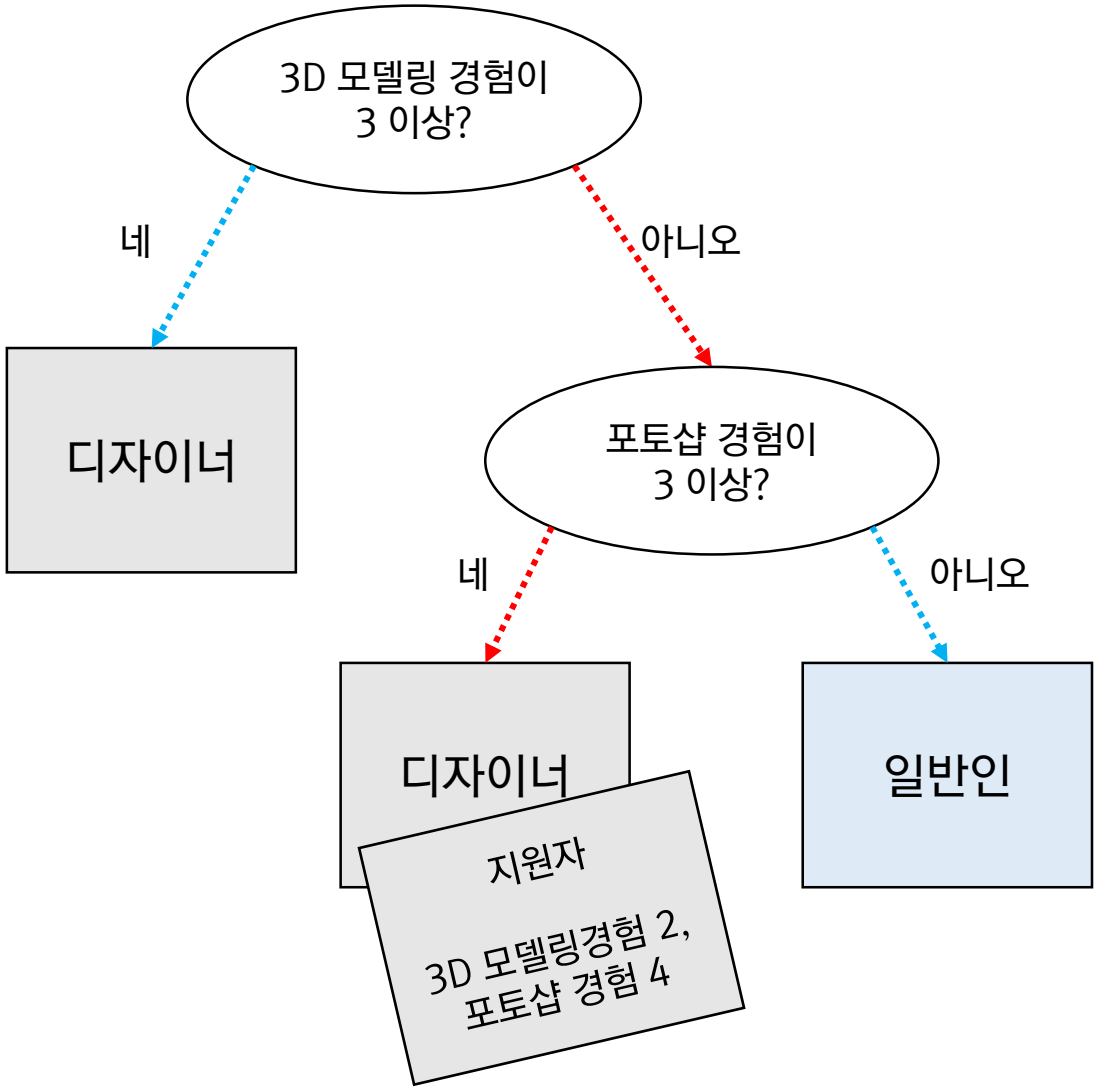
결정 트리 (Decision Tree)



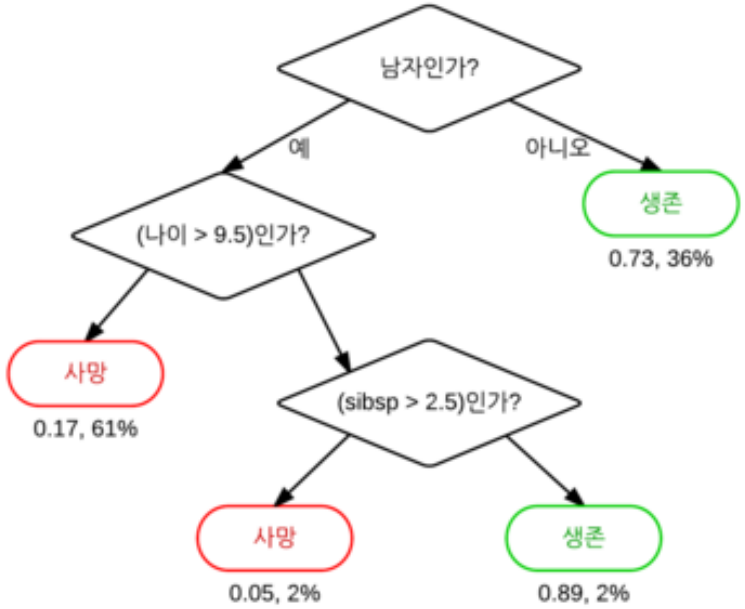
결정 트리 (Decision Tree)



결정 트리 (Decision Tree)



결정 트리 예시 : 타이타닉호 탑승객 생존여부



결정 트리 생성방법

- 엔트로피란? 정보의 무질서 정도를 나타냄.
- 엔트로피를 가장 많이 줄일 수 있는 방법으로

$$Entropy = -p_1 \log(p_1) - p_2 \log(p_2)$$

- 처음 상태의 엔트로피 :
  - P1(디자이너의 비율) = (5/10) = 0.5
  - P2(비디자이너의 비율) = (5/10) = 0.5

  - 엔트로피 = −0.5 * $\log_2$ (0.5) −0.5 * $\log_2$ (0.5) = 1

결정 트리 생성방법

- 초기 엔트로피 : 1
- 정보 획득량 (Information Gain) : 초기 엔트로피 – 현재 엔트로피

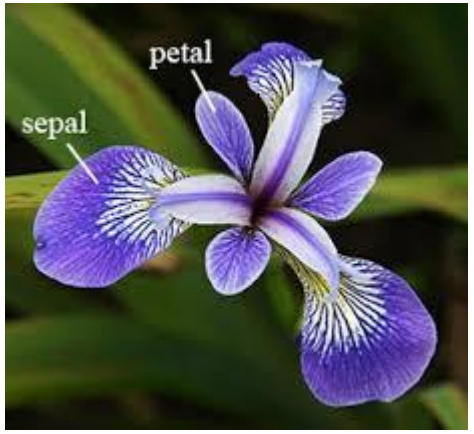
	네	아니오	엔트로피
H1 3D 〉= 2	디4, 일3	디1, 일2	0.965
H2 3D 〉= 3	디4	디1, 일5	0.394
H3 3D 〉= 4	디3	디2, 일5	0.604
H4 포토샵 〉= 4	디1	디4, 일5	0.892
H5 포토샵 〉= 3	디3	디2, 일5	0.604
H6 포토샵 〉= 2	디5, 일2	일3	0.604

결정 트리 생성방법

- 결정 트리는 ?
- 엔트로피를 가장 많이 줄일 수 있는 방법으로
  = 정보 획득량이 가장 많은 방법으로
- 쉽게 이해하면, 가장 분류를 잘 할 수 있는 방법으로 결정 트리를 생성한다.

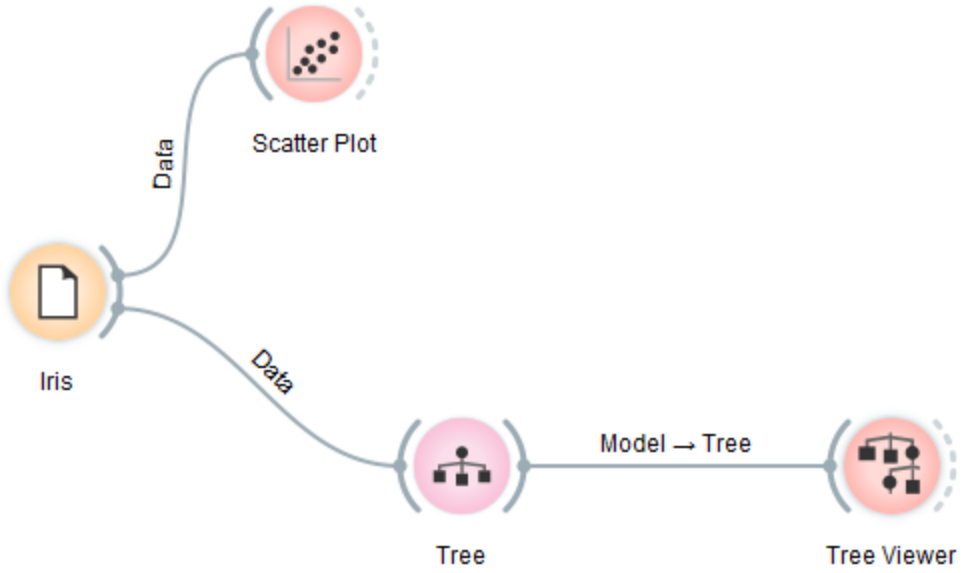
실습 – Iris 데이터

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa
12	4.8	3.4	1.6	0.2	setosa
13	4.8	3.0	1.4	0.1	setosa
14	4.3	3.0	1.1	0.1	setosa
15	5.8	4.0	1.2	0.2	setosa



- 각 데이터가 어떤 꽃의 종류인지 알려준다.
- X : sepal length, sepal width, petal length, petal width
- y : 꽃의 종류 (Setosa, Virginica, Versicolor)

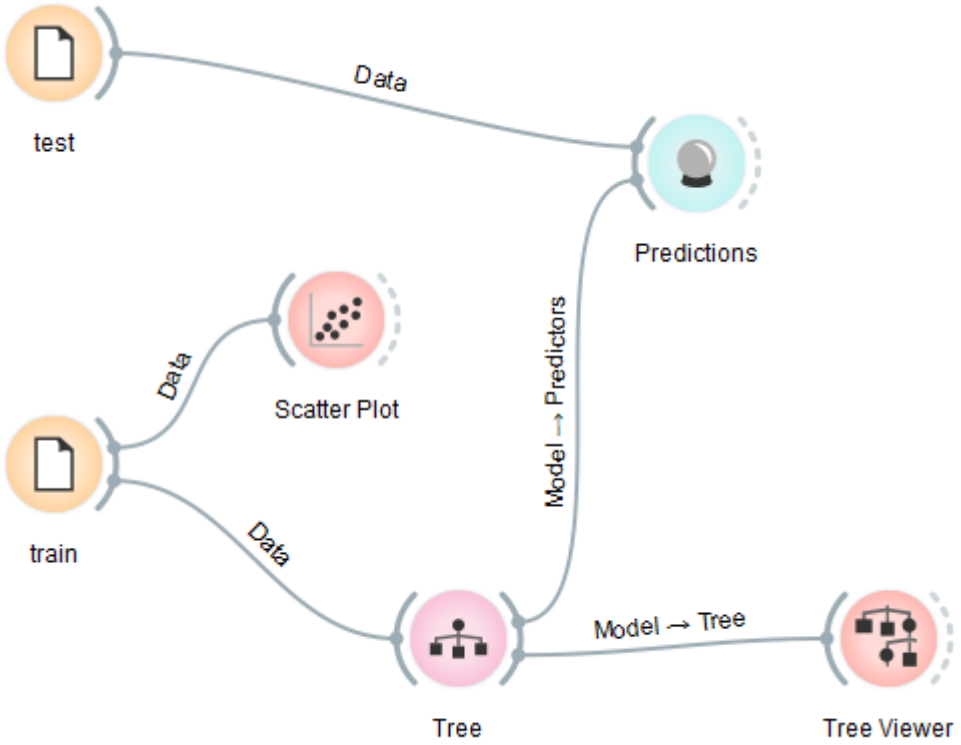
실습 – Iris 데이터



실습 – 디자인 전공여부

	major_design	exp_3d	exp_photoshop
1	1.0	3.000	3.000
2	1.0	4.000	4.000
3	1.0	4.000	5.000
4	1.0	3.000	4.000
5	1.0	5.000	5.000
6	1.0	1.000	4.000
7	1.0	4.000	5.000
8	1.0	3.000	4.000
9	1.0	3.000	4.000
10	1.0	4.000	5.000
11	1.0	4.000	4.000
12	1.0	4.000	4.000
13	1.0	4.000	5.000
14	1.0	3.000	4.000
15	0.0	2.000	2.000
16	0.0	2.000	1.000
17	0.0	1.000	3.000
18	0.0	1.000	2.000
19	0.0	1.000	4.000
20	0.0	1.000	2.000
21	0.0	1.000	1.000
22	0.0	2.000	3.000

실습 – 디자인 전공여부


test
Data
Predictions
Data
Scatter Plot
Model → Predictors
train
Data
Model → Tree
Tree
Tree Viewer

결정 트리의 장점

- 결과를 해석하고 이해하기 쉽다.
- 자료를 가공할 필요가 거의 없다.
- 수치 자료와 범주형 자료 모두 적용 가능
- 안정적
- 대규모 데이터 셋에서도 잘 동작한다.

결정 트리의 단점

- 완전하게 최적의 결정 트리를 만드는 것은 매우 어려운 문제.
- 데이터의 특성이 수직/수평으로 구분되지 못할 때 분류율이 떨어진다.
- 일반화 하지 못할 경우 트리가 복잡해지는 문제

참고 자료

[머신러닝] 의사결정트리 (Decision Tree) 알고리즘 쉽게 이해하기
https://www.youtube.com/watch?v=n0p0120Gxqk