Machine Learning

Chapter 3 지도 학습(Supervised Learning)

START

Smart Media
스마트미디어인재개발원

학습목표

- Decision Tree 알고리즘을 이해 할 수 있다.
- Label 인코딩과 One-hot 인코딩을 이해 할 수 있다.
- 교차 검증 기법을 이해 할 수 있다.

Decision Tree

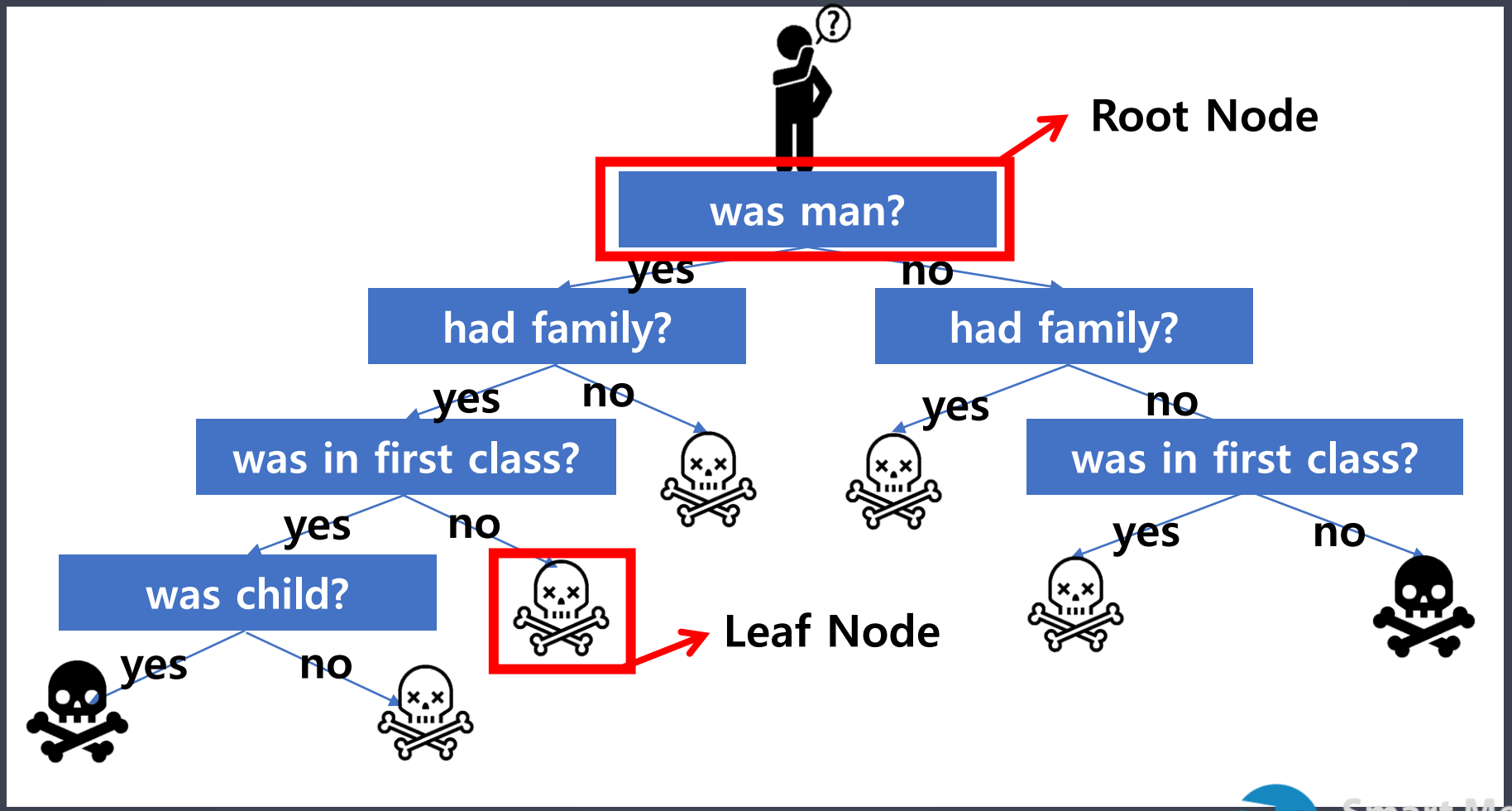
Decision Tree

Decision Tree(결정트리)

- Tree를 만들기 위해 예/아니오 질문을 반복하며 학습한다.
- 다양한 앙상블(ensemble) 모델이 존재한다
  (RandomForest, GradientBoosting, XGBoost, LightGBM)
- 분류와 회귀에 모두 사용 가능

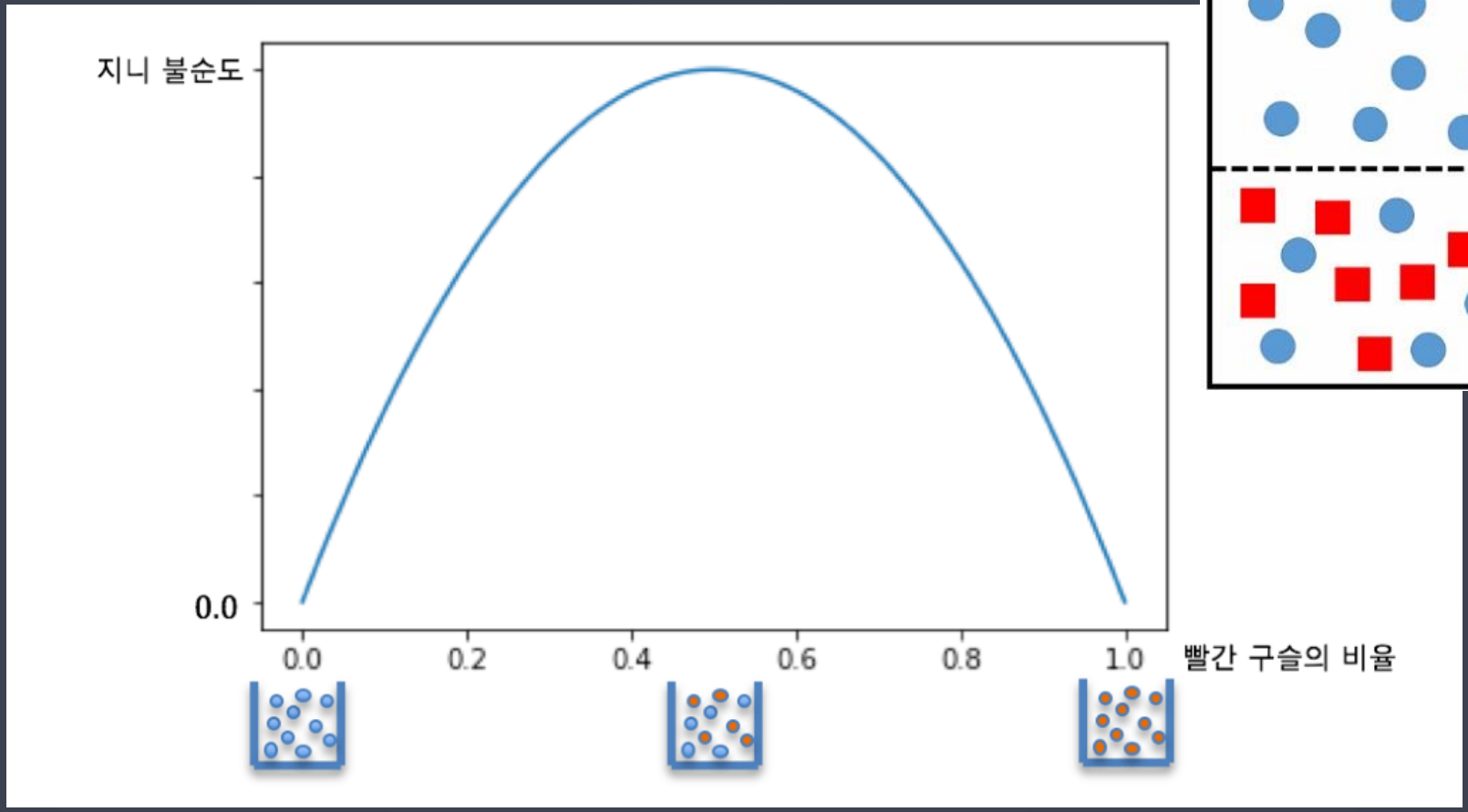
Decision Tree

Decision Tree(결정트리)



Decision Tree

Gini Impurity(지니 불순도)



Decision Tree(결정트리)

- 타깃 값이 한 개인 리프 노드를 순수 노드라고 한다.
- 모든 노드가 순수 노드가 될 때 까지 학습하면 복잡해지고 과대 적합이 된다.
- 새로운 데이터 포인트가 들어오면 해당하는 노드를 찾아 분류라면 더 많은 클래스를 선택하고, 회귀라면 평균을 구한다.

Decision Tree

Decision Tree(결정트리) 과대적합 제어

- 노드 생성을 미리 중단하는 사전 가지치기(pre-pruning)와 트리를 만든후에 크기가 작은 노드를 삭제하는 사후 가지치기(pruning)가 있다
  (sklearn은 사전 가지치기만 지원)
- 트리의 최대 깊이나 리프 노드의 최대 개수를 제어
- 노드가 분할 하기 위한 데이터 포인트의 최소 개수를 지정

Decision Tree

장단점 및 주요 매개변수(Hyperparameter)

- 트리의 최대 깊이 : max_depth
  (값이 클수록 모델의 복잡도가 올라간다.)
- 리프 노드의 최대 개수 : max_leaf_nodes
- 리프 노드가 되기 위한 최소 샘플의 개수 : min_samples_leaf

Decision Tree

장단점 및 주요 매개변수(Hyperparameter)

- 만들어진 모델을 쉽게 시각화할 수 있어 이해하기 쉽다.
  (white box model)
- 각 특성이 개별 처리되기 때문에 데이터 스케일에 영향을 받지 않아 특성의 정규화나 표준화가 필요 없다.
- 트리 구성시 각 특성의 중요도를 계산하기때문에 특성 선택 (Feature selection)에 활용될 수 있다.

Decision Tree

장단점 및 주요 매개변수(Hyperparameter)

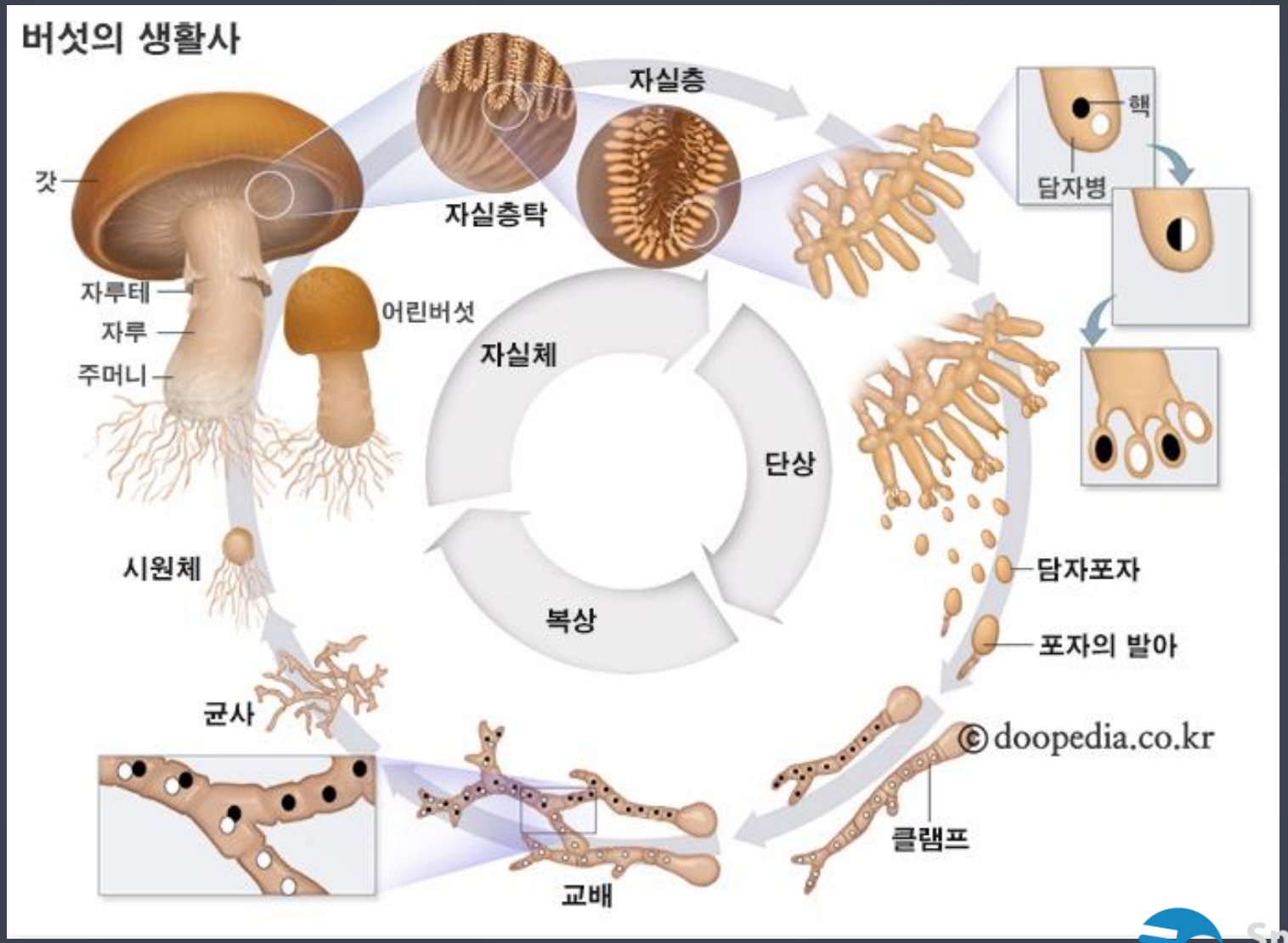
- 훈련데이터 범위 밖의 포인트는 예측 할 수 없다.
  (ex : 시계열 데이터)

- 가지치기를 사용함에도 불구하고 과대적합되는 경향이 있어 일반화 성능이 좋지 않다.

Decision Tree 실습 - ex04

Mushroom 데이터 활용
Decision Tree 분류 실습

Decision Tree 실습 - ex04



Decision Tree 실습 - ex04

poisonous : 독버섯(poisonous), 식용버섯(edible)
cap-shape : 갓 모양(b,c,x,f,k,s) : 원뿔/평면/볼록 등
cap-surface : 갓 표면(f,g,y,s) : 섬유질/비늘모양/부드러움 등
cap-color : 갓 색(n,b,c,g,r,p,u,e,w,y) : 계피/회색/노란색 등
bruises : 타박상(t,f) : 예/아니오
odor : 냄새(a,l,c,y,f,m,n,p,s) : 아몬드,생선,매운 등

gill-attachment(자실층 위치), gill-spacing(자실층 간격), gill-size(자실층 크기), gill-color(자실층 색), stalk-shape(자루 모양), stalk-root(자루 뿌리), stalk-surface-above-ring(자루 표면 위 자루테), stalk-surface-below-ring(자루 표면 아래 자루테), stalk-color-above-ring(자루 색 위 자루테), stalk-color-below-ring(자루 색 아래 자루테), veil-type(베일 유형), veil-color(베일 색), ring-number(링 번호), ring-type(링 타입), spore-pring-color(포자 색), population(인구), habitat(서식지)

Decision Tree 실습 - ex04

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	poisonous	cap-shape	cap-surfac	cap-color	bruises	odor	gill-attach	gill-spacin	gill-size	gill-color	stalk-shap	stalk-root	stalk-surfa	stalk-surfa	stalk-colo
2	p	x	s	n	t	p	f	c	n	k	e	e	s	s	w
3	e	x	s	y	t	a	f	c	b	k	e	c	s	s	w
4	e	b	s	w	t	l	f	c	b	n	e	c	s	s	w
5	p	x	y	w	t	p	f	c	n	n	e	e	s	s	w
6	e	x	s	g	f	n	f	w	b	k	t	e	s	s	w
7	e	x	y	y	t	a	f	c	b	n	e	c	s	s	w
8	e	b	s	w	t	a	f	c	b	g	e	c	s	s	w
9	e	b	y	w	t	l	f	c	b	n	e	c	s	s	w
10	p	x	y	w	t	p	f	c	n	p	e	e	s	s	w
11	e	b	s	y	t	a	f	c	b	g	e	c	s	s	w
12	e	x	y	y	t	l	f	c	b	g	e	c	s	s	w
13	e	x	y	y	t	a	f	c	b	n	e	c	s	s	w
14	e	b	s	y	t	a	f	c	b	w	e	c	s	s	w
15	p	x	y	w	t	p	f	c	n	k	e	e	s	s	w

범주형(이산형) 특성이기 때문에 인코딩 필요

categorical feature

Decision Tree 실습 - ex04

Label 인코딩 or One-hot 인코딩 방식을 이용해 수치화한다.

범주형 변수

$$\hat{y} = w[0]\times x[0] + w[1]\times x[1] + \cdots + w[p]\times x[p] + b > 0$$

범주형 특성

	age	workclass	education	gender	hours-per-week
0	39	State-gov	Bachelors	Male	40
1	50	Self-emp-not-inc	Bachelors	Male	13
2	38	Private	HS-grad	Male	40
3	53	Private	11th	Male	40
4	28	Private	Bachelors	Female	40

연속형 특성

Decision Tree 실습 - ex04

Label Encoding
단순 수치 값으로 mapping하는 작업

I	J	K	L
Ticket	Fare	Cabin	Embarked
A/5 21171	7.25		S
PC 17599	71.2833	C85	C
STON/O2.	7.925		S
113803	53.1	C123	S
373450	8.05		S
330877	8.4583		Q
17463	51.8625	E46	S
349909	21.075		S

L
Embarked
0
1
0
0
0
2
0
0

Decision Tree 실습 - ex04

One-hot Encoding
0 or 1의 값을 가진 여러 개의 새로운 특성으로 변경하는 작업

I	J	K	L
Ticket	Fare	Cabin	Embarked
A/5 21171	7.25		S
PC 17599	71.2833	C85	C
STON/O2.	7.925		S
113803	53.1	C123	S
373450	8.05		S
330877	8.4583		Q
17463	51.8625	E46	S
349909	21.075		S

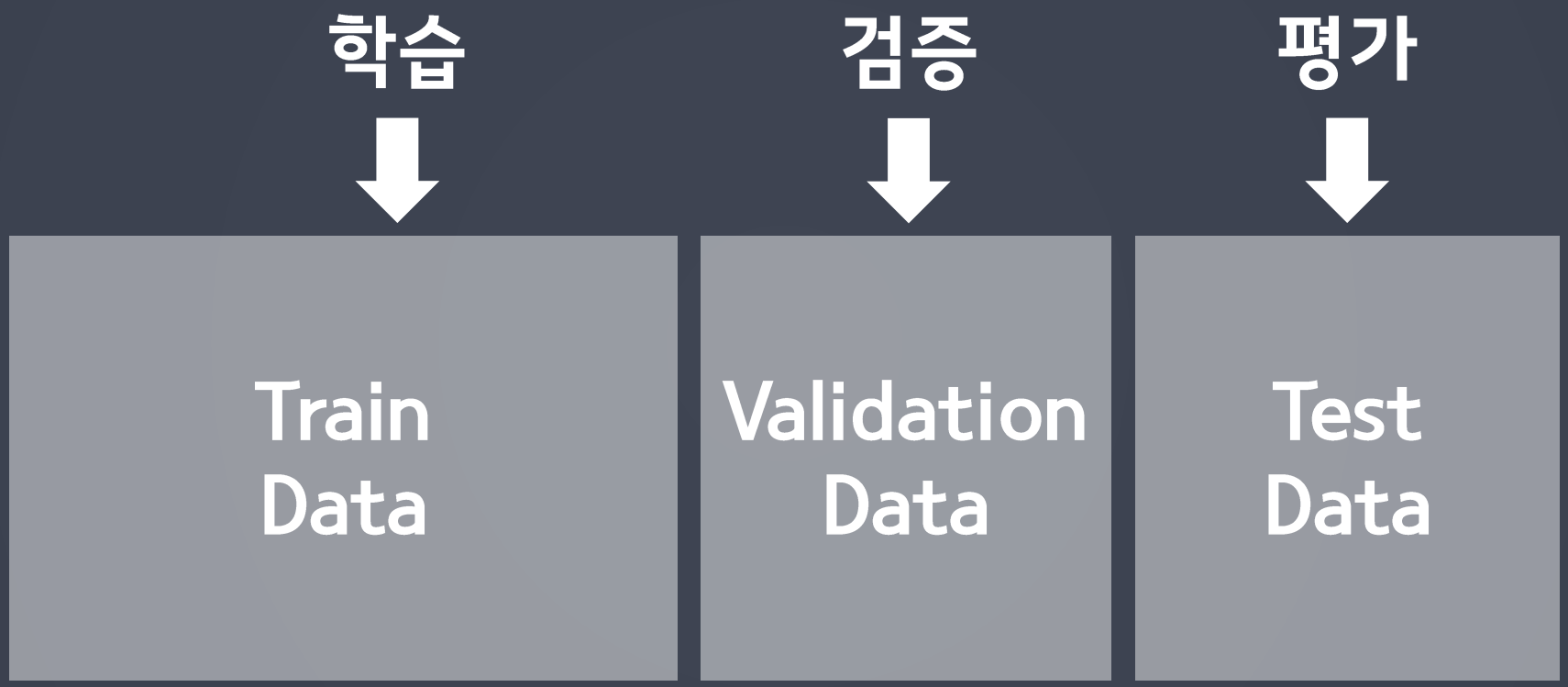
M	N	O
Embarked_S	Embarked_C	Embarked_Q
1	0	0
0	1	0
1	0	0
1	0	0
1	0	0
0	0	1
1	0	0
1	0	0

Cross validation
(교차검증)

Cross validation(교차검증)

테스트 세트에 맞게 학습 될 수 있다.

Cross validation(교차검증)



Cross validation(교차검증)

Cross validation
(교차검증)

학습-평가 데이터 나누기를 여러 번 반복하여
일반화 에러를 평가하는 방법

Cross validation(교차검증)

K-fold cross-validation 동작 방법

1. 데이터 셋을 k개로 나눈다.

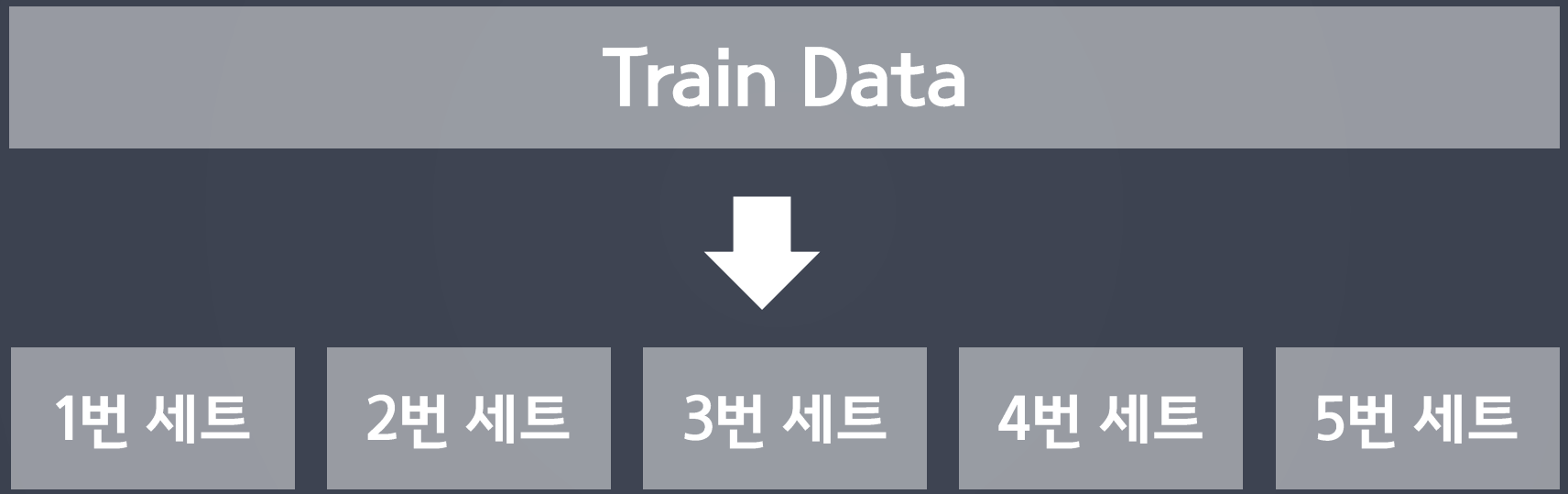
2. 첫 번째 세트를 제외하고 나머지에 대해 모델을 학습한다.
   그리고 첫 번째 세트를 이용해서 평가를 수행한다.

3. 2번 과정을 마지막 세트까지 반복한다.

4. 각 세트에 대해 구했던 평가 결과의 평균을 구한다.

Cross validation(교차검증)

K-fold cross-validation 동작 방법



Cross validation(교차검증)

K-fold cross-validation 동작 방법

Cross validation(교차검증)

cross-validation 장/단점

- 데이터의 여러 부분을 학습하고 평가해서 일반화 성능을 측정하기 때문에 안정적이고 정확하다. (샘플링 차이 최소화)
- 모델이 훈련 데이터에 대해 얼마나 민감한지 파악가능 (점수 대역 폭이 넓으면 민감)
- 데이터 세트 크기가 충분하지 않은 경우에도 유용하게 사용 가능하다.
- 여러 번 학습하고 평가하는 과정을 거치기 때문에 계산량이 많아진다

Decision Tree 활용 Titanic 실습 - ex05

Decision Tree를 활용해 Titanic 데이터를 학습하고 교차검증을 적용해보자.