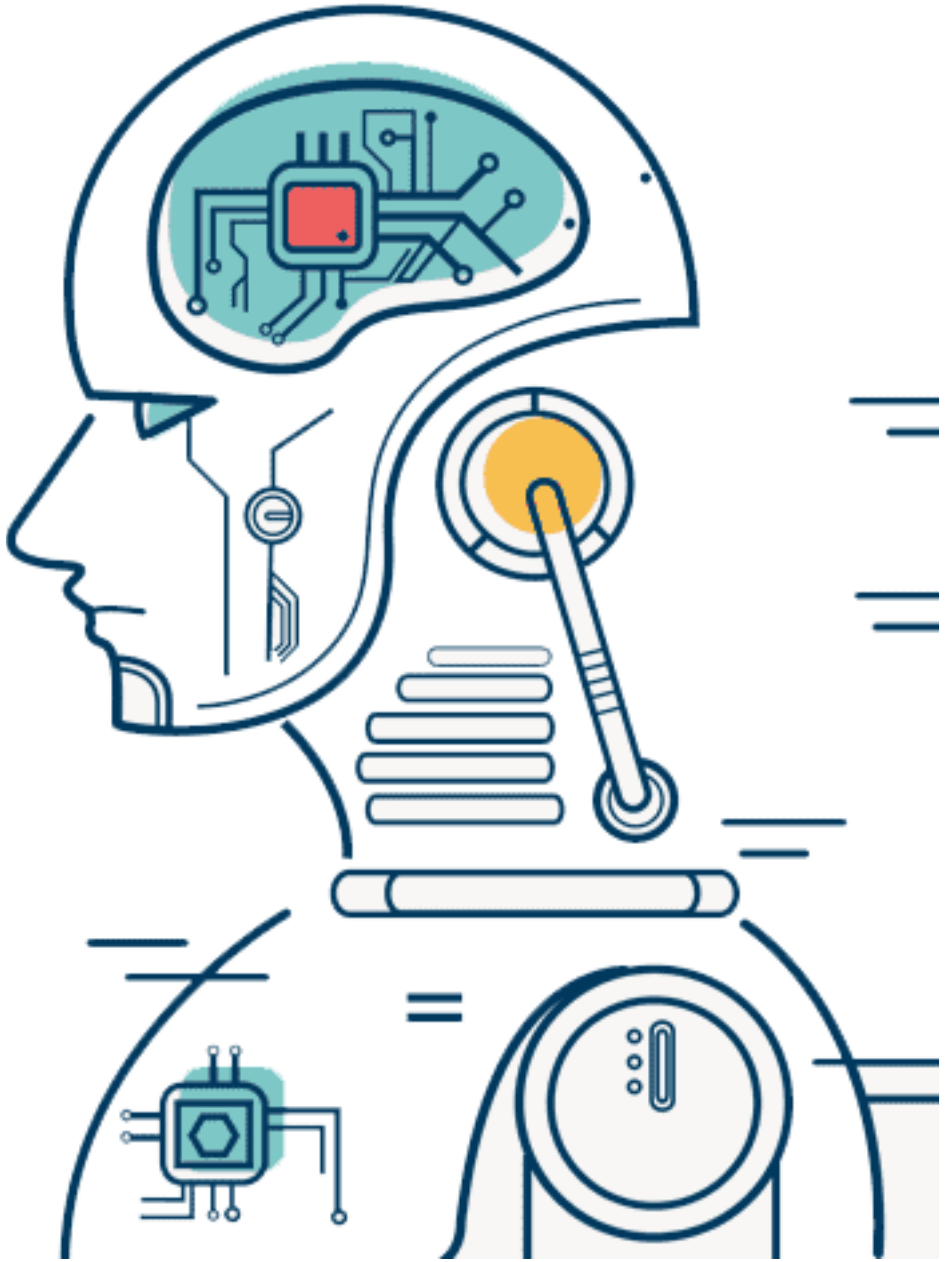
Smart Media
스마트미디어인재개발원

Machine Learning



Chapter 2 지도학습

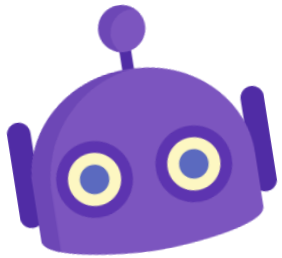
(일반화, 과대적합, 과소적합, KNN, Iris/유방암 실습, KNN회귀)

학습목표

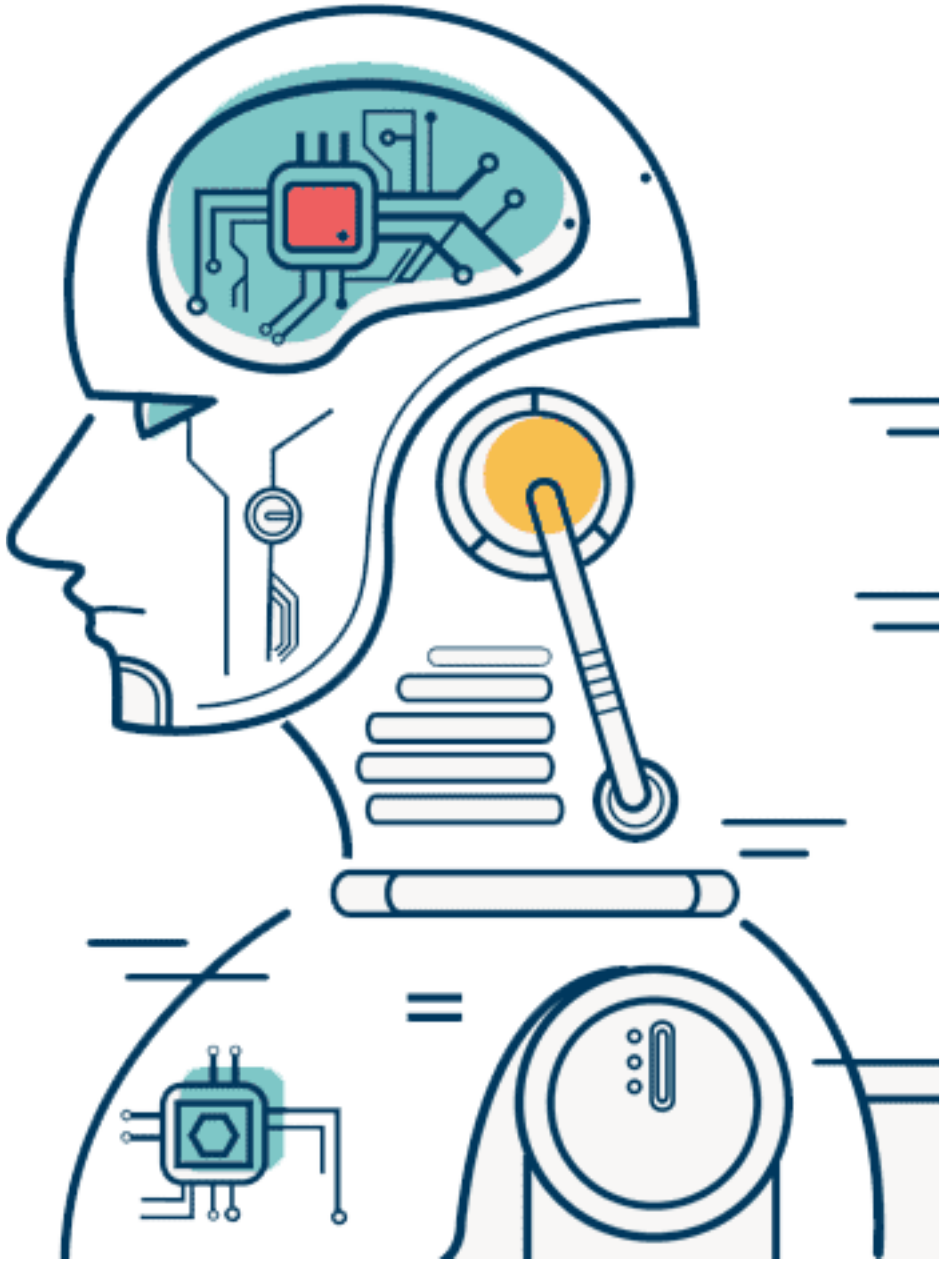
Smart Media
스마트미디어인재개발원

- 일반화, 과대적합, 과소적합을 이해 할 수 있다.
- KNN 알고리즘을 이해 할 수 있다.
- 하이퍼파라미터 튜닝을 할 수 있다.
- KNN 회귀에 대해 이해할 수 있다

Smart Media
스마트미디어인재개발원



일반화,과대적합,과소적합



일반화, 과대적합, 과소적합

아이에게 공이 무엇인지 알려주자

Smart Media
스마트미디어인재개발원

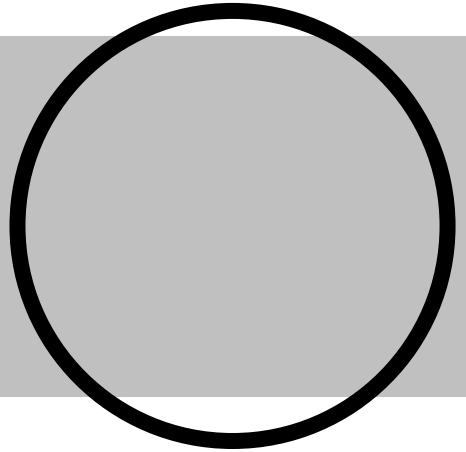
공이라는 것은..

둥글게 생겼다.
오각형이 여러 개 붙어있다.
검은색과 흰색으로 구성된다.
반짝반짝 광이 난다.

과대적합

일반화, 과대적합, 과소적합

공이라는 것은..



둥글게 생겼다.

과소적합

일반화, 과대적합, 과소적합

일반화 (Generalization)

- 훈련 세트로 학습한 모델이 테스트 세트에 대해 정확히 예측하도록 하는 것 .

과대적합 (Overfitting)

- 훈련 세트에 너무 맞추어져 있어 테스트 세트의 성능 저하.

과소적합 (Underfitting)

- 훈련 세트를 충분히 반영하지 못해 훈련 세트, 테스트 세트에서 모두 성능이 저하.

과대적합 (Overfitting)

요트 회사의 고객

45세 이상, 자녀 셋 미만,
이혼하지 않은 고객

나이	보유차량수	주택보유	자녀수	혼인상태	애완견	보트구매
66	1	yes	2	사별	no	yes
52	2	yes	3	기혼	no	yes
22	0	no	0	기혼	yes	no
25	1	no	1	미혼	no	no
44	0	no	2	이혼	yes	no
39	1	yes	2	기혼	yes	no
26	1	no	2	미혼	no	no
40	3	yes	1	기혼	yes	no
53	2	yes	2	이혼	no	yes
64	2	yes	3	이혼	no	no
58	2	yes	2	기혼	yes	yes
33	1	no	1	미혼	no	no

Smart Media
스마트미디어인재개발원

과소적합 (Underfitting)

요트 회사의 고객

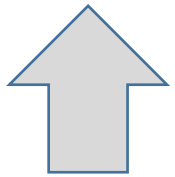
집이 있는 고객

나이	보유차량수	주택보유	자녀수	혼인상태	애완견	보트구매
66	1	yes	2	사별	no	yes
52	2	yes	3	기혼	no	yes
22	0	no	0	기혼	yes	no
25	1	no	1	미혼	no	no
44	0	no	2	이혼	yes	no
39	1	yes	2	기혼	yes	no
26	1	no	2	미혼	no	no
40	3	yes	1	기혼	yes	no
53	2	yes	2	이혼	no	yes
64	2	yes	3	이혼	no	no
58	2	yes	2	기혼	yes	yes
33	1	no	1	미혼	no	no

일반화, 과대적합, 과소적합

Smart Media
스마트미디어인재개발원

일반화 성능이 최대화 되는 모델을 찾는 것이 목표



과대적합 (Overfitting)

- 너무 상세하고 복잡한 모델링을 하여 훈련데이터에만 과도하게 정확히 동작하는 모델.

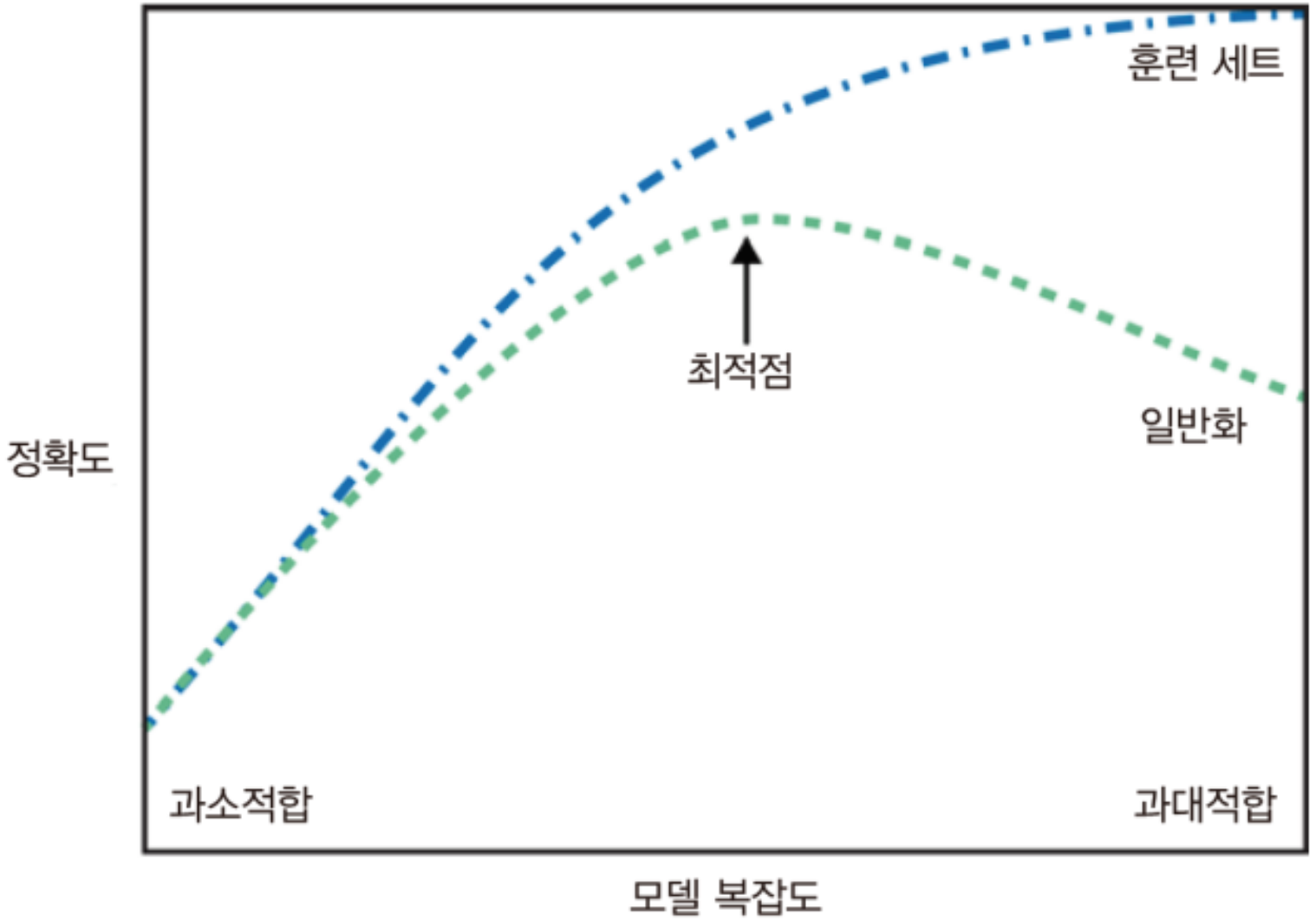
과소적합 (Underfitting)

- 모델링을 너무 간단하게 하여 성능이 제대로 나오지 않는 모델.

일반화, 과대적합, 과소적합

Smart Media
스마트미디어인재개발원

모델 복잡도 곡선



일반화, 과대적합, 과소적합

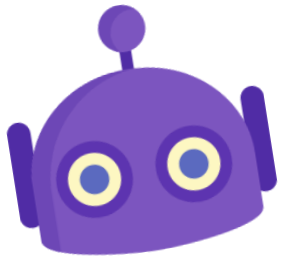
해결방법

- 주어진 훈련데이터의 다양성 보장 → 다양한 데이터포인트를 골고루 나타내야 한다.
- 일반적으로 데이터 양이 많으면 일반화에 도움이 된다.
- 하지만 편중된 데이터를 많이 모으는 것은 도움이 되지 않는다.
- 규제(Regularization)을 통해 모델의 복잡도를 적정선으로 설정한다.

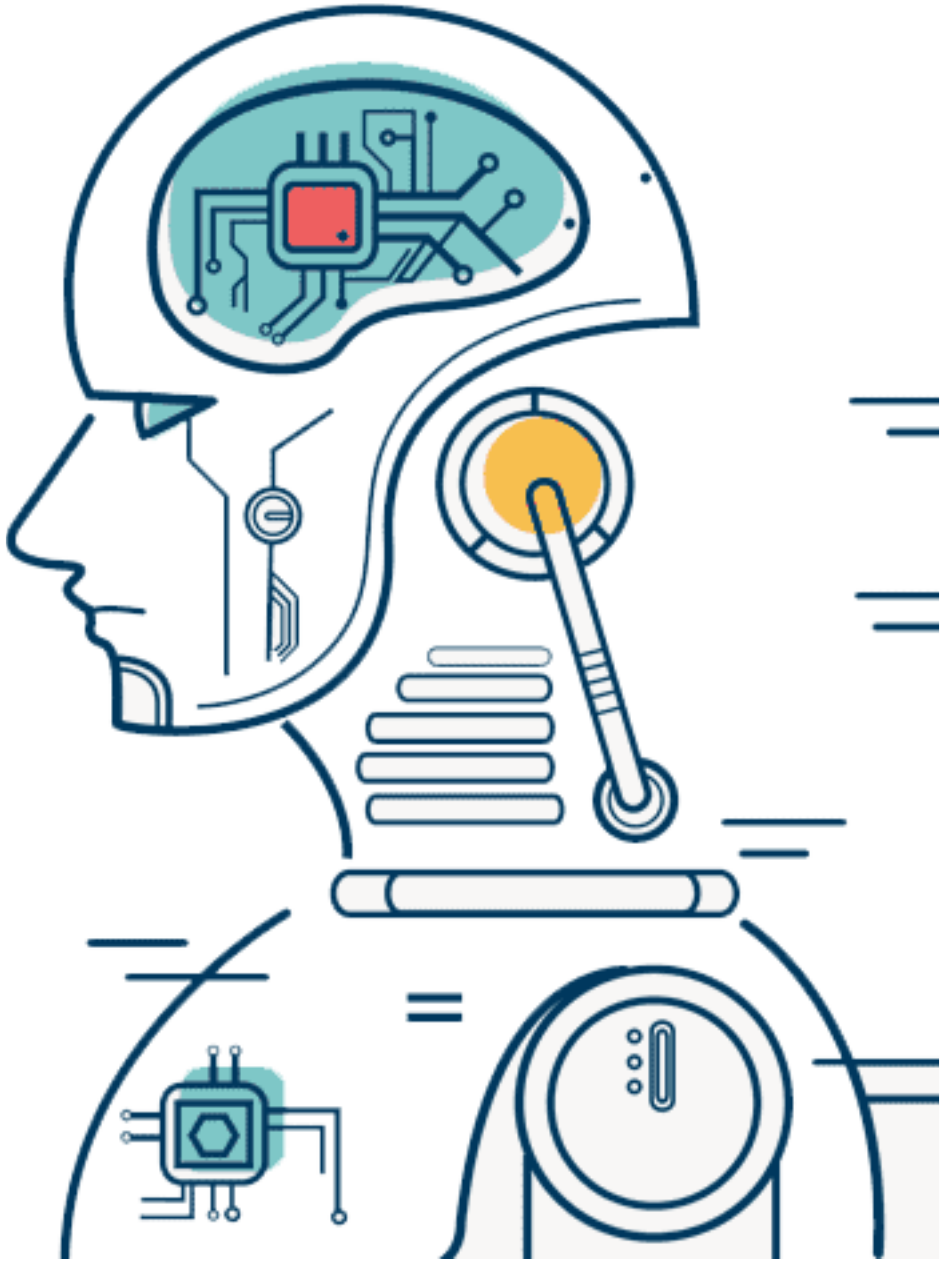
BMI 데이터에서 특성을 추가하여 과소적합문제를 해결해보자

	Gender	Height	Weight	GenderxGender	GenderxHeight	GenderxWeight	HeightxHeight	HeightxWeight	WeightxWeight
0	0	174	96	0	0	0	30276	16704	9216
1	0	189	87	0	0	0	35721	16443	7569
2	1	185	110	1	185	110	34225	20350	12100
3	1	195	104	1	195	104	38025	20280	10816
4	0	149	61	0	0	0	22201	9089	3721

Smart Media
스마트미디어인재개발원



K-Nearest Neighbors (KNN)



k-최근접 이웃 알고리즘

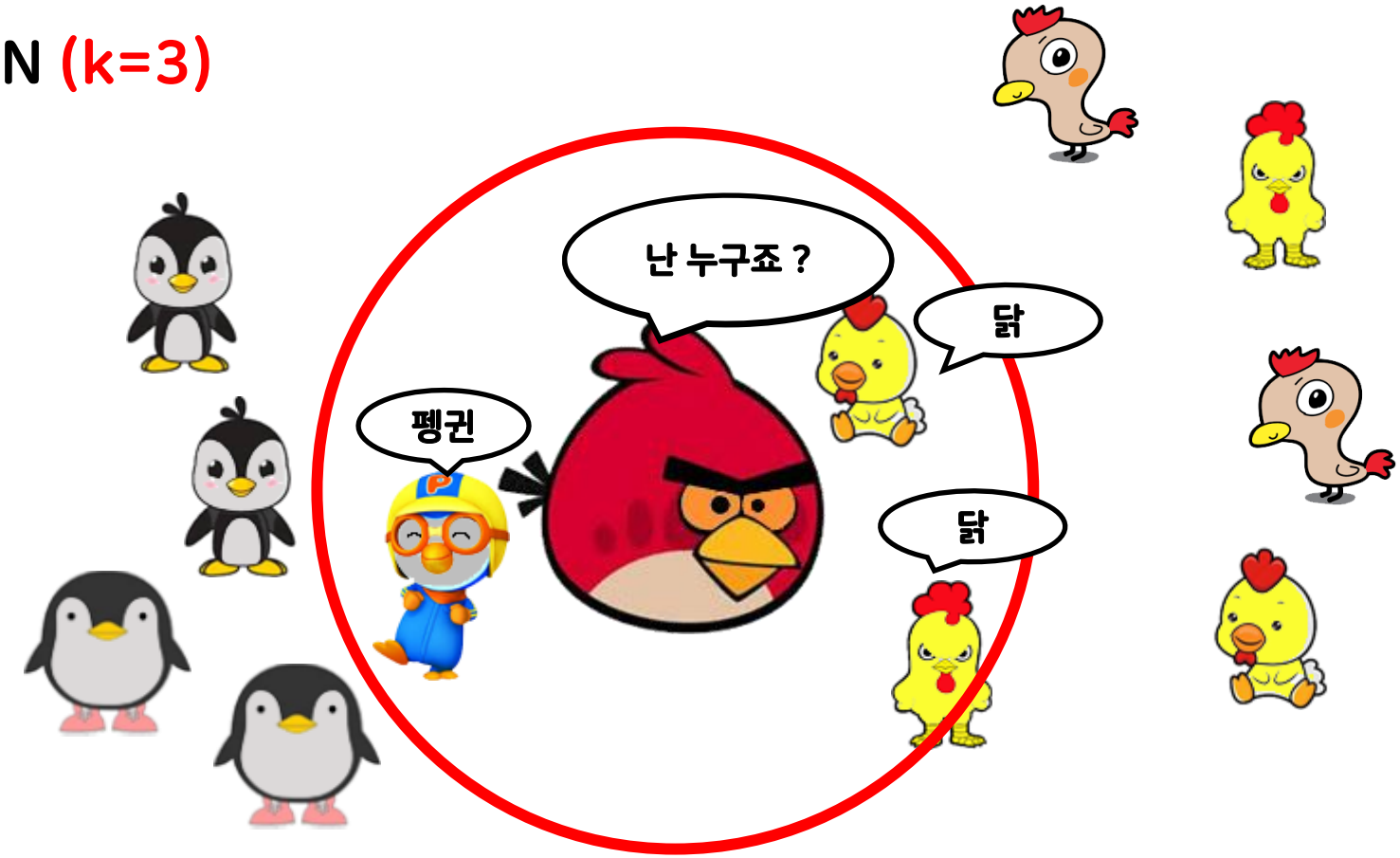
- 새로운 데이터 포인트와 가장 가까운 훈련 데이터셋의 데이터 포인트를 찾아 예측
- k 값에 따라 가까운 이웃의 수가 결정
- 분류와 회귀에 모두 사용 가능

K-Nearest Neighbors (KNN)

Smart Media
스마트미디어인재개발원

k-최근접 이웃 알고리즘

KNN (k=3)

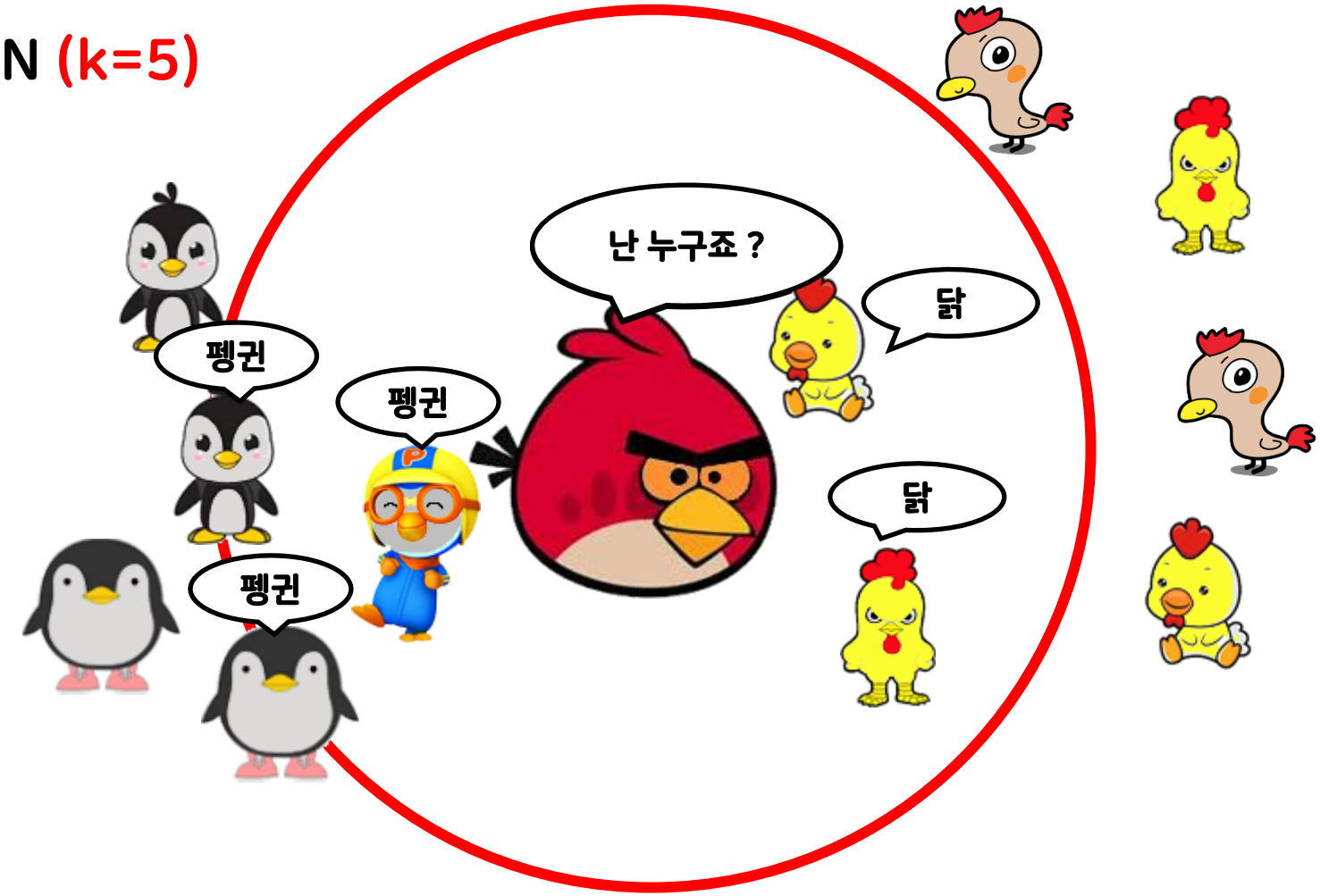


K-Nearest Neighbors (KNN)

Smart Media
스마트미디어인재개발원

k-최근접 이웃 알고리즘

KNN (k=5)

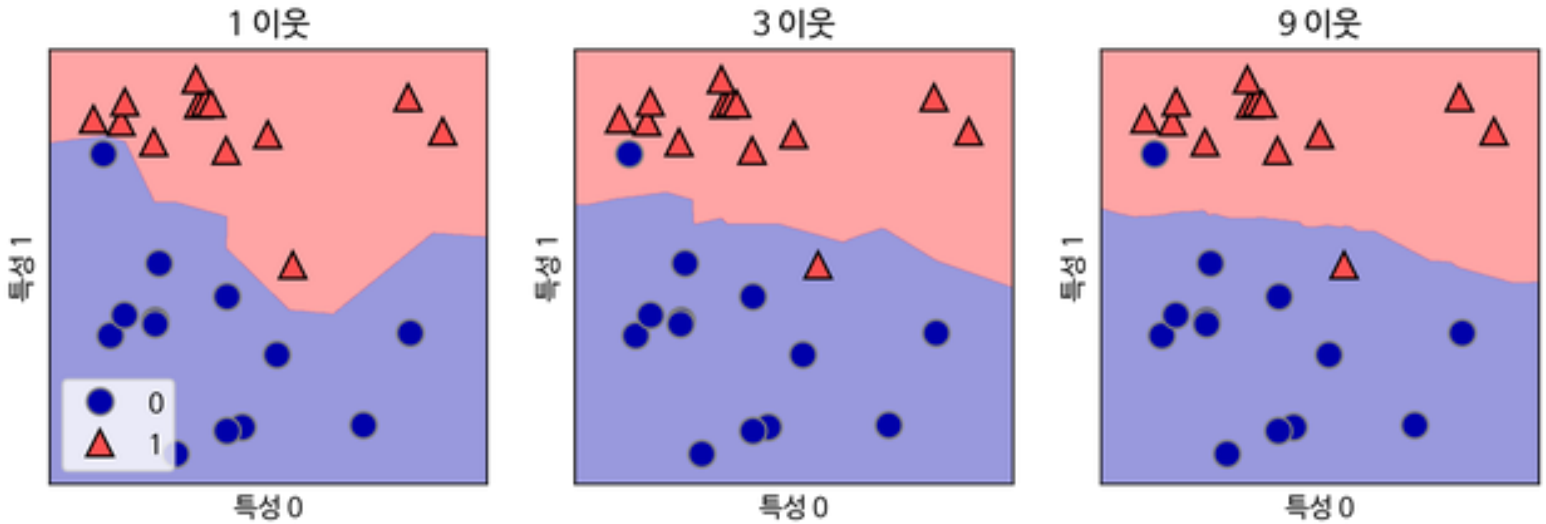


k-최근접 이웃 알고리즘

- **결정경계 (Decision Boundary)** : 클래스 분류하는 경계

  - 이웃이 적을수록 모델의 복잡도 상승 → **과대적합**

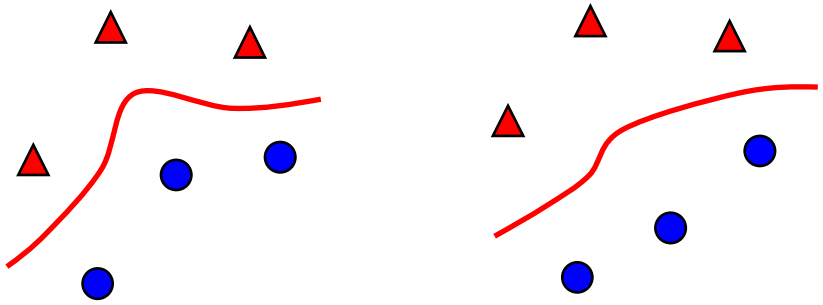
  - 이웃이 전체 데이터 개수와 같아지면 항상 가장 많은 클래스로 예측 → **과소적합**



Smart Media
스마트미디어인재개발원

k-최근접 이웃 알고리즘

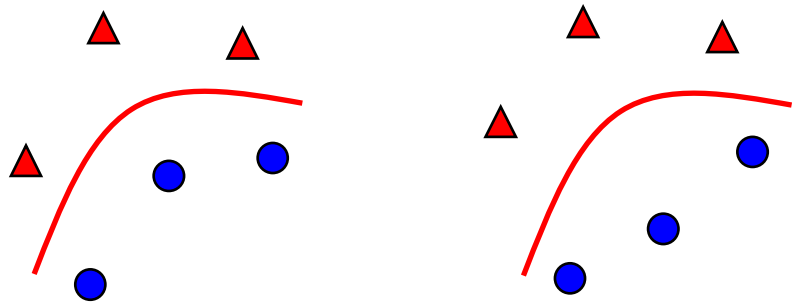
적은 이웃 → 과대 적합



train 데이터 test 데이터

데이터가 살짝만 달라져도 결정 경계가
달라져서 정확도가 떨어짐

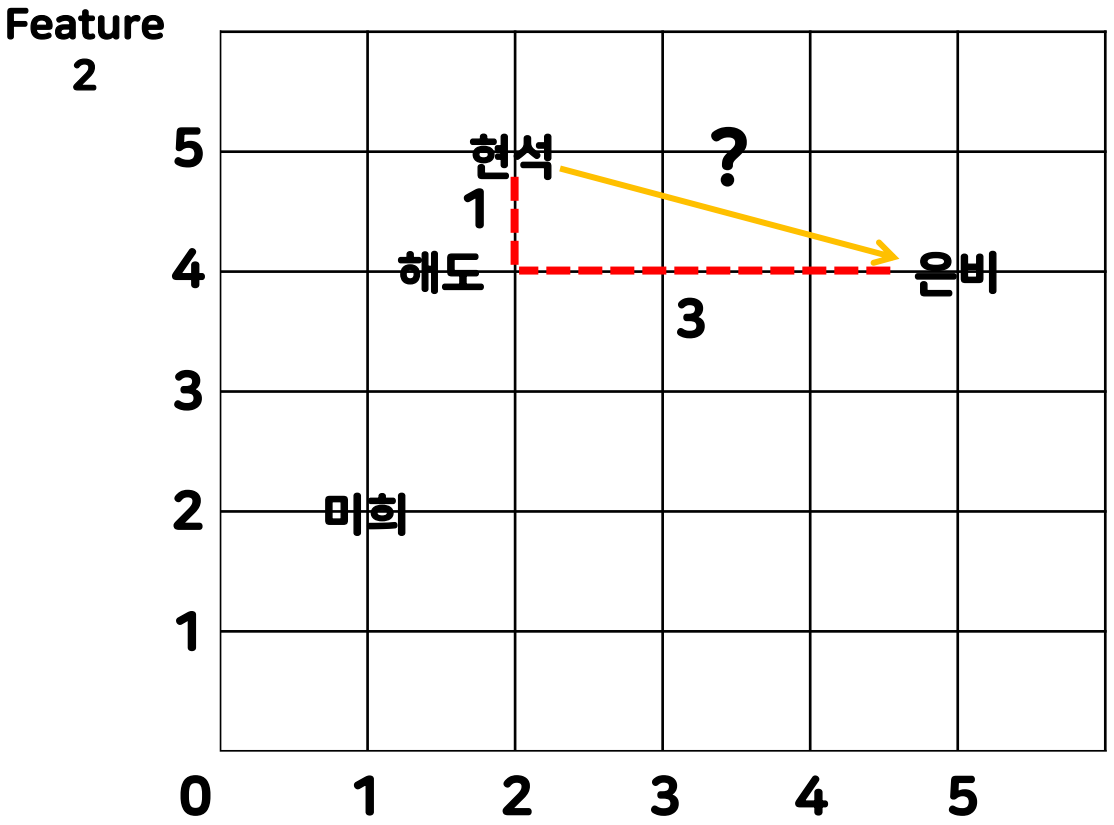
많은 이웃 → 과소 적합



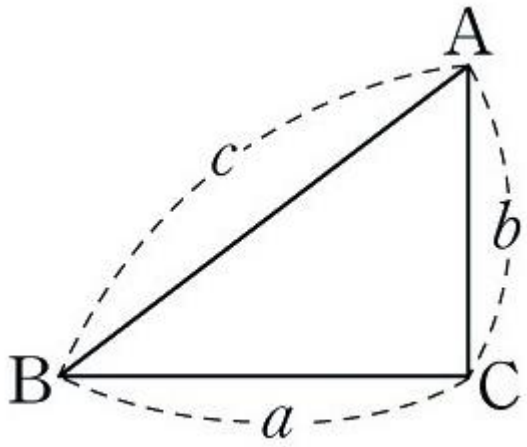
train 데이터 test 데이터

데이터가 달라져도 결정 경계가
변하지 않아서 정확도가 올라감

데이터 포인트(sample) 사이 거리 값 측정 방법



$a^2 + b^2 = c^2$ (피타고라스의 정리)



데이터 포인트(sample) 사이 거리 값 측정 방법

$$\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

유클리디언 거리공식 (Euclidean Distance)

주요 매개변수(Hyperparameter)

scikit-learn의 경우

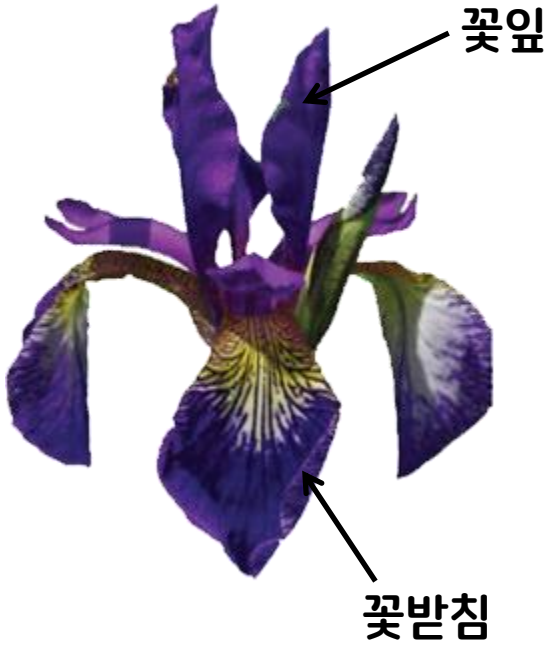
```
KNeighborsClassifier(n_neighbors=이웃의 수)
```

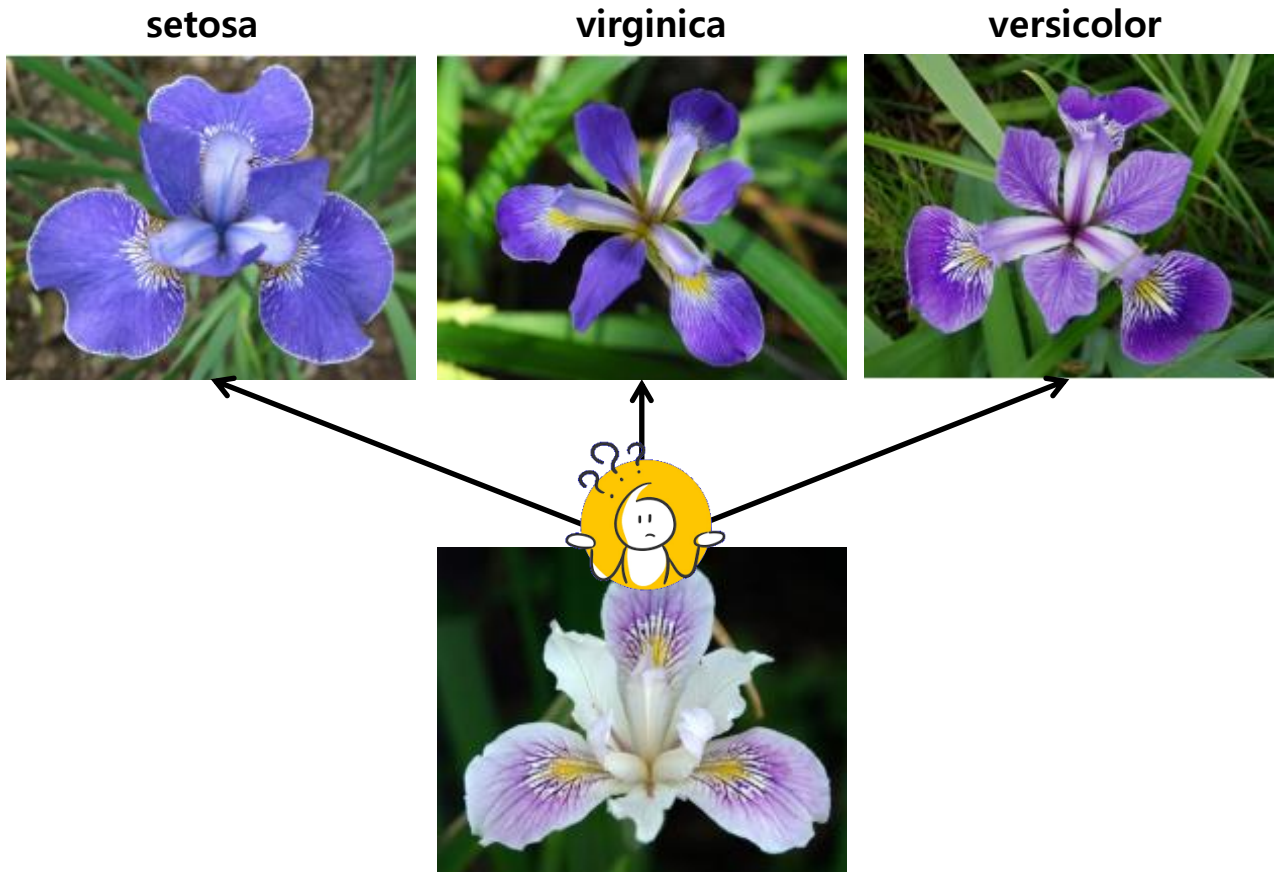
장단점

- 이해하기 매우 쉽고 조정 없이도 좋은 성능을 발휘하는 모델
- 훈련 데이터 세트가 크면(특성,샘플의 수) 예측이 느려진다
- 수백 개 이상의 많은 특성을 가진 데이터 세트와 특성 값 대부분이 0인 희소(sparse)한 데이터 세트에는 잘 동작하지 않는다
- 거리를 측정하기 때문에 같은 scale을 같도록 정규화 필요
- 전처리 과정이 중요, 잘 쓰이지는 않음

iris 데이터를 이용한 KNN 분류 실습

붓꽃(iris) 데이터셋





붓꽃(iris) 데이터셋

- 150개의 데이터
- 4개의 특성과 1개의 클래스(3개의 품종)로 구성

	sepal_length	sepal_width	petal_length	petal_width	species
	꽃받침 길이	꽃받침 너비	꽃잎 길이	꽃잎 너비	품종
1	5.1	3.5	1.4	0.2	Iris-setosa
2	4.9	3.0	1.4	0.2	Iris-setosa
3	4.7	3.2	1.3	0.2	Iris-versicolor
...					
150	5.9	3.0	5.1	1.8	Iris-virginica

train_test_split() 함수

- 데이터 셋에서 훈련데이터와 테스트데이터로 분리하는 기능

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y,
                                                    test_size = 0.3,
                                                    random_state=0)
```

X : 특성 데이터
y : 라벨 데이터
test_size ： 테스트 셋의 비율
random_state : 선택할 데이터 시드
weight ：가중치 함수

유방암 데이터를 이용한 KNN 분류 실습

K-Nearest Neighbors (KNN) 실습

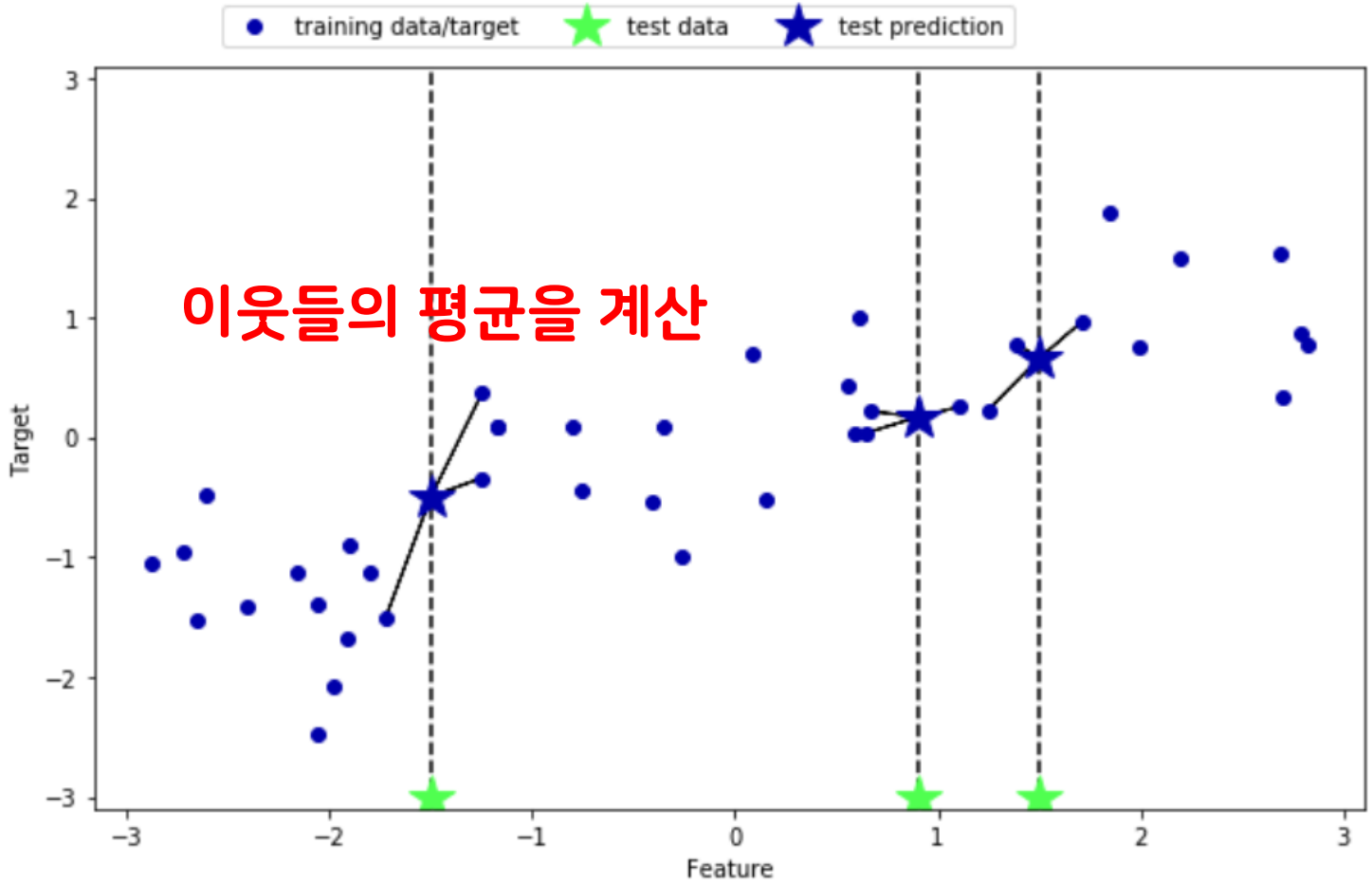
Smart Media
스마트미디어인재개발원

유방암 데이터셋

- wisconsin의 유방암 데이터셋
- 총 569건의 데이터
  (악성(212), 양성 (357)으로 구성)

id	환자 식별 번호
diagnosis	양성 여부 (M = 악성, B = 양성)
각 세포에 대한 정보들	
radius	반경 (중심에서 외벽까지 거리들의 평균값)
texture	질감 (Gray-Scale 값들의 표준편차) #gray-scale 값은 광도의 정보를 전달할 수
perimeter	둘레
area	면적
smoothness	매끄러움(반경길이의 국소적 변화)
compactness	조그만 정도(둘레^2/면적 - 1)
concavity	오목함(윤곽의 오목한 부분의 정도)
points	오목한 점의 수
symmetry	대칭
dimension	프랙탈 차원(해안선근사 -1)
_mean	3 ~ 12 번까지는 평균값을 의미합니다.
_se	13 ~ 22 번까지는 표준오차(Standard Error) 를 의미합니다.
_worst	23 ~ 32 번까지는 각 세포별 구분들에서 제일 큰 3개의 값을 평균낸 값입니다.

K-Nearest Neighbors (KNN) 회귀



주요 매개변수(Hyperparameter)

scikit-learn의 경우

KNeighborsRegressor(n_neighbors=이웃의 수)

K-Nearest Neighbors (KNN) 회귀 실습

Smart Media
스마트미디어인재개발원

보스톤 집값 데이터셋

- 506개의 데이터
- 13개의 정보와 1개의 클래스로 구성

0	CRIM : 인구 1인당 범죄 발생 수
1	ZN : 25,000평방 피트 이상의 주거 구역 비중
2	INDUS : 소매업 외 상업이 차지하는 면적 비율
3	CHAS : 찰스강 위치 변수 (1: 강 주변, 0: 이외)
4	NOX : 일산화질소 농도
5	RM : 집의 평균 방 수
6	AGE : 1940년 이전 지어진 비율
7	DIS : 5가지 보스턴 시 고용 시설까지의 거리
8	RAD : 순환고속도로의 접근 용이성
9	TAX : $10,000당 부동산 세율 총계
10	PTRATIO : 지역별 학생과 교사 비율
11	B : 지역별 흑인 비율
12	LSTAT : 급여가 낮은 직업에 종사하는 인구 비율 (%)
13	가격 (단위 : $1,000)