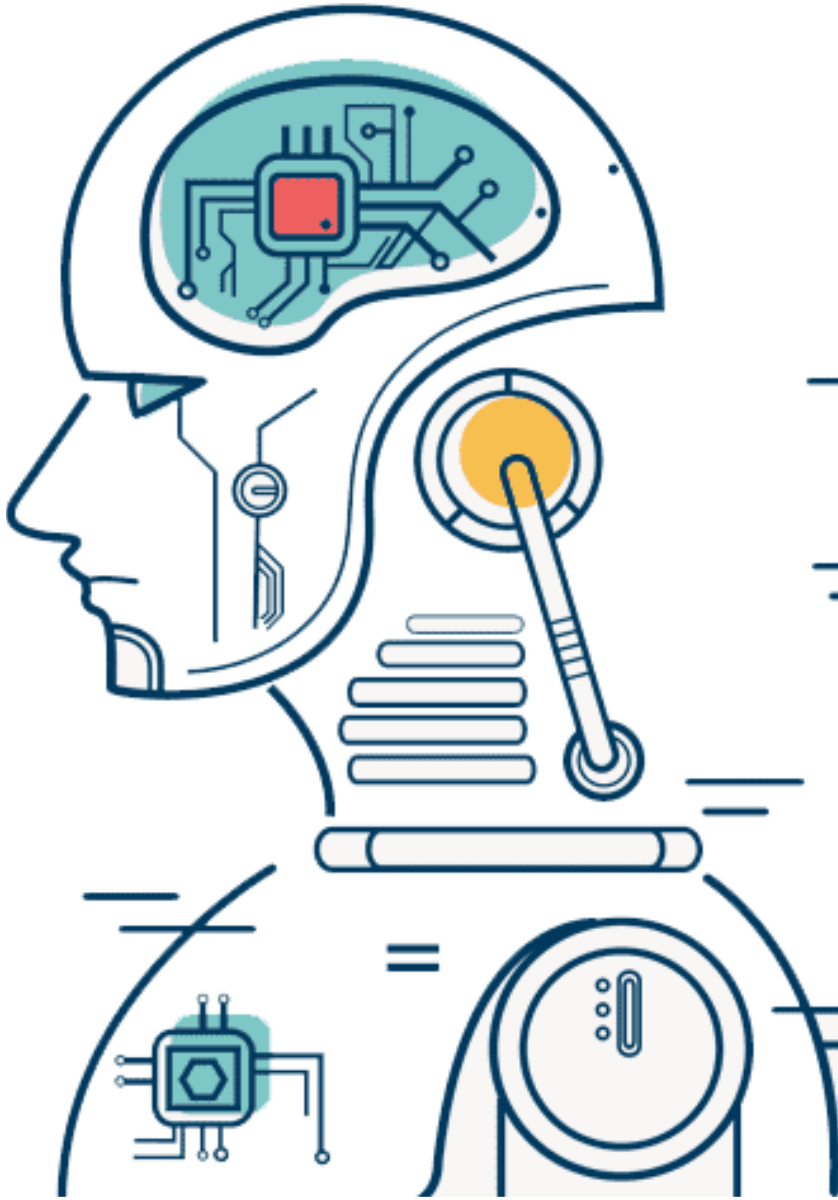
스마트인재개발원
Smart Human Resources Development

Machine Learning

Chapter02 일반화, 과대, 과소,KNN



강사 손지영

학습목표

스마트인재개발원
Smart Human Resources Development

- 일반화, 과대적합, 과소적합을 이해할 수 있다.
- KNN 알고리즘을 이해할 수 있다.
- 하이퍼 파라미터 튜닝을 할 수 있다.

일반화, 과대적합, 과소적합

모델의 신뢰도를 측정하고, 성능을 확인하기 위한 개념

어떤 게 공인지 아이에게 설명해보자

공에 대해 설명1 - 과대적합

공의 종류는 다양하지만, 축구공만 설명

- 둥글다.
- 오각형과 육각형으로 이어졌다.
- 검은색과 흰색으로 구성된다.
- 반짝반짝 광이 난다.

공에 대해 설명2 - 과소적합

공의 특징은 다양하지만, 단순하게 설명

- 둥글다.

과대적합, 과소적합

- **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

마케팅 대상 선별 조건

보트 회사 고객

45세 이상, 자녀 셋 미만, 이혼하지 않은 고객

규칙 복잡, 너무 상세
적절한 고객 선별 어려움

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

과소적합

마케팅 대상 선별 조건

보트 회사 고객

집이 있는 고객

규칙 단순, 너무 간단
적절한 고객 선별 어려움

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

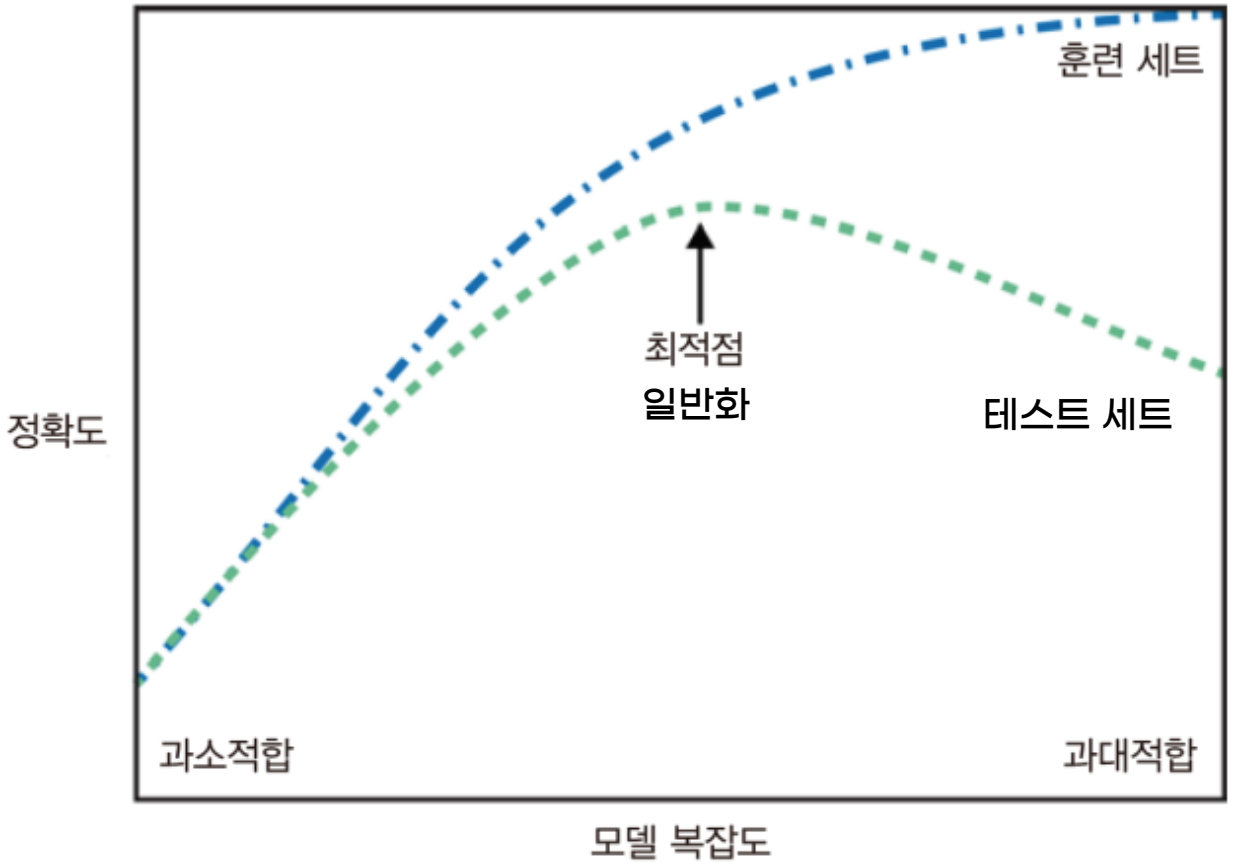
일반화, 과대적합, 과소적합

머신러닝 궁극적인 목표

일반화 성능이 최대가 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

모델 복잡도 곡선



모델의 복잡도 해결

- 일반적으로 데이터 양이 많으면 일반화에 도움이 됨
- 주어진 훈련데이터의 다양성 보장되어야 함
- 편중된 데이터를 많이 모으는 것보다 다양한 데이터포인트를 골고루 나타내기
- 규제(Regularization)을 통해 모델의 복잡도를 적정선으로 설정

K-Nearest Neighbors (KNN)

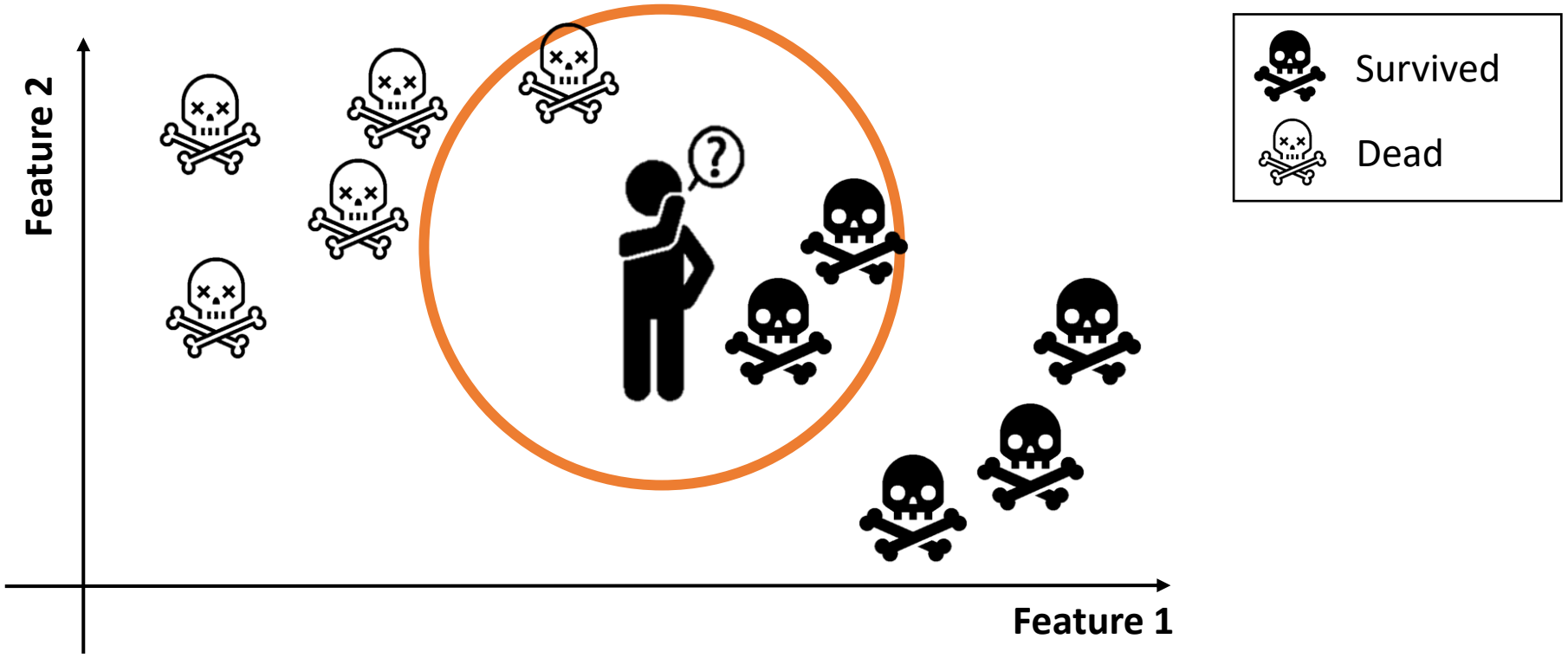
K-Nearest Neighbors(knn)이란?
k-최근접 이웃 알고리즘

- 유유상종의 개념과 유사
- 유사한 점이 서로 가까이에서 발견될 수 있다는 가정
- 분류와 회귀에 모두 사용 가능
- 특정 데이터 포인트와 가장 가까운 이웃 데이터 포인트를 찾아 이웃과 동일한 클래스로 분류
- 특정 데이터 포인트와 가장 가까운 이웃 데이터 포인트를 찾아 이웃의 값을 평균 내어 회귀 예측

K-Nearest Neighbors(knn)의 k란?

k-최근접 이웃 알고리즘

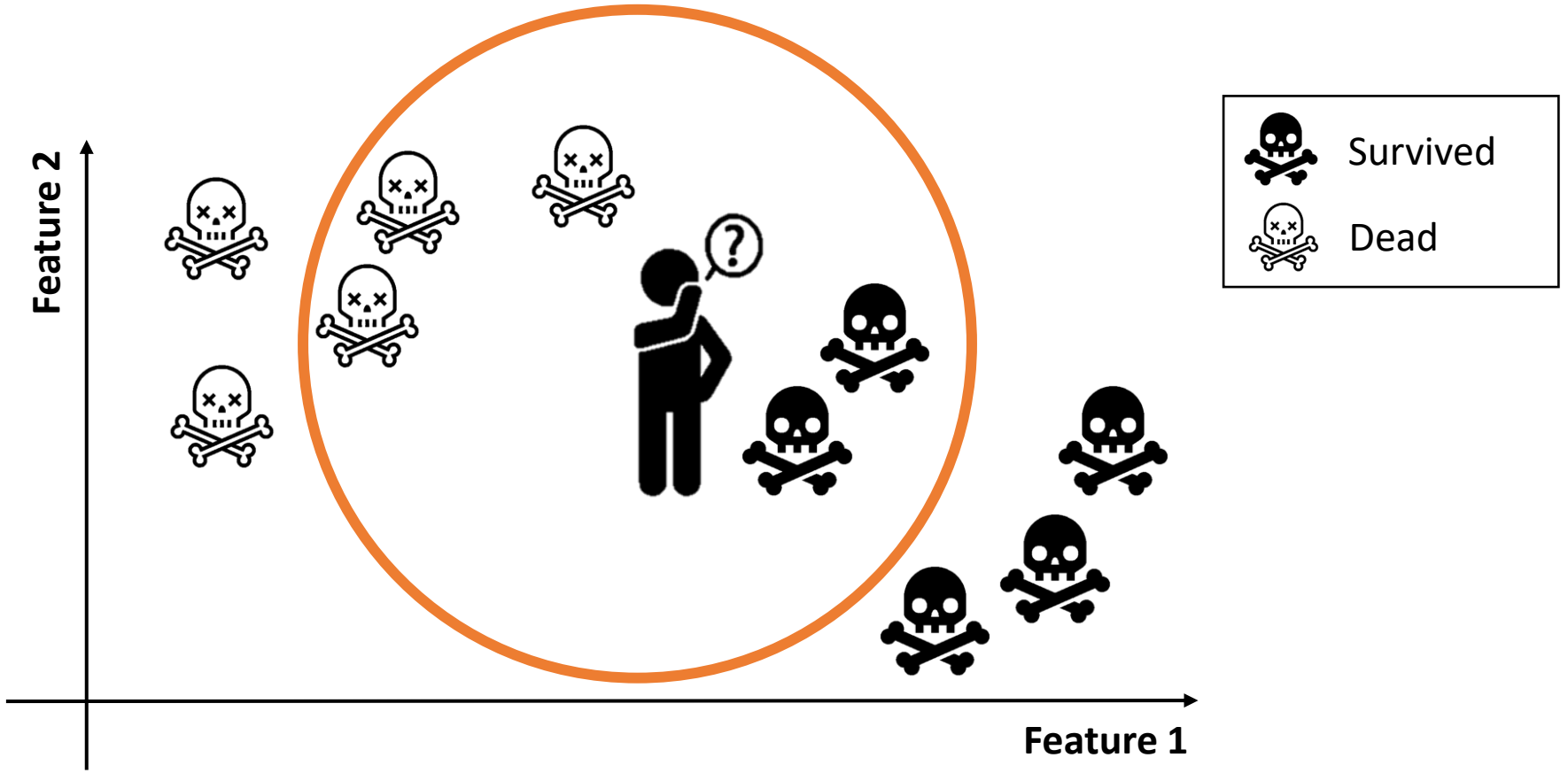
KNN (k=3)



K-Nearest Neighbors(knn)의 k란?

k-최근접 이웃 알고리즘

KNN (k=5)



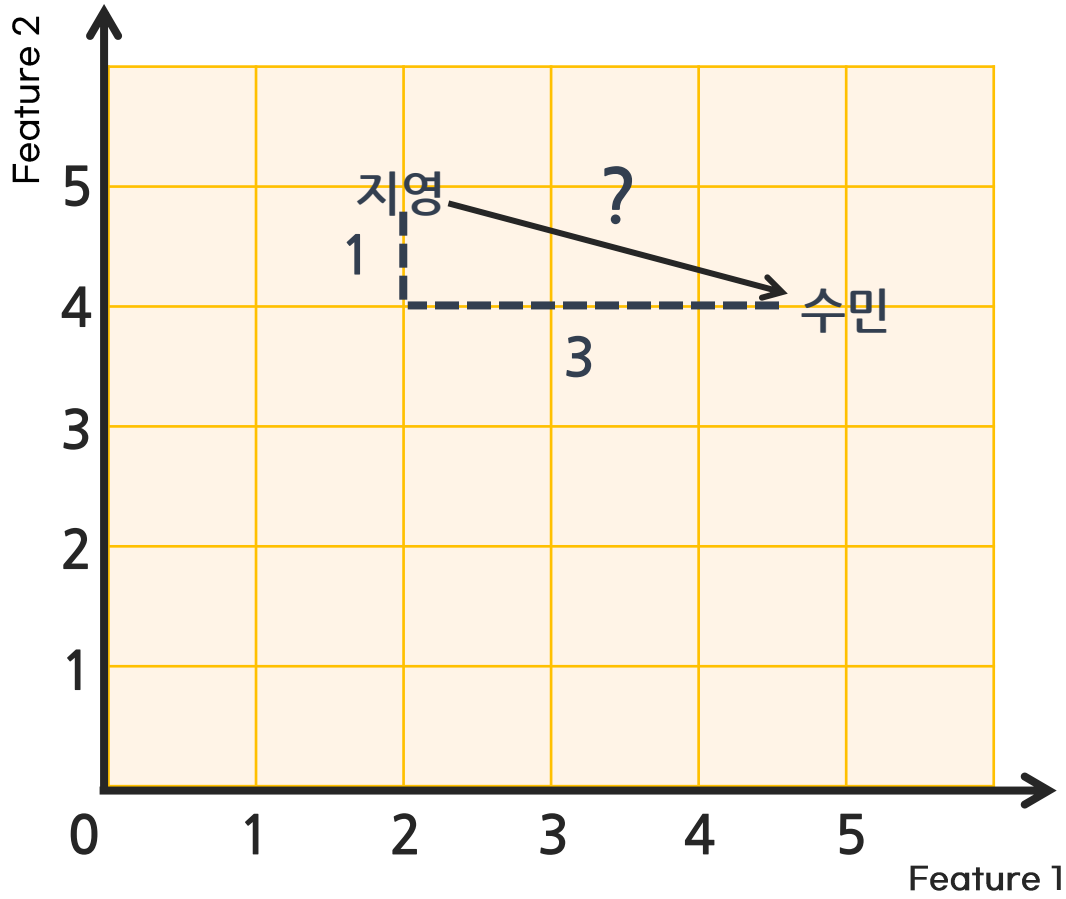
K-Nearest Neighbors(knn)의 k란?

k-최근접 이웃 알고리즘

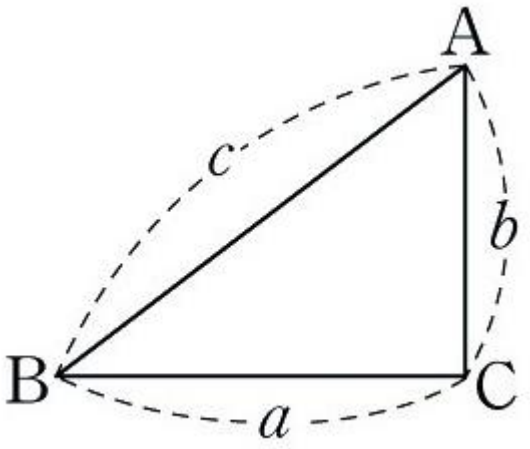
- k 값이 작을 수록 모델의 복잡도가 상대적으로 증가(noise 값에 민감)
- 반대로 k 값이 커질수록 모델의 복잡도가 낮아짐
- k는 이웃의 수를 의미
- k의 수는 학습데이터가 100개라면 100까지 설정 가능
- 예측하면 빈도가 가장 많은 클래스 레이블로 분류

K-Nearest Neighbors(knn) 거리공식

k-최근접 이웃 알고리즘



$a^2 + b^2 = c^2$ (피타고라스의 정리)



K-Nearest Neighbors(knn) 거리공식

k-최근접 이웃 알고리즘

유클리디언 거리공식 (Euclidean Distance)

데이터 포인트(sample) 사이 거리 값 측정 방법

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

	Feature1	feature2	feature3	·····
지영	4	3	1	·····
수민	5	2	0	·····

K-Nearest Neighbors(knn) 장단점 및 키워드

k-최근접 이웃 알고리즘

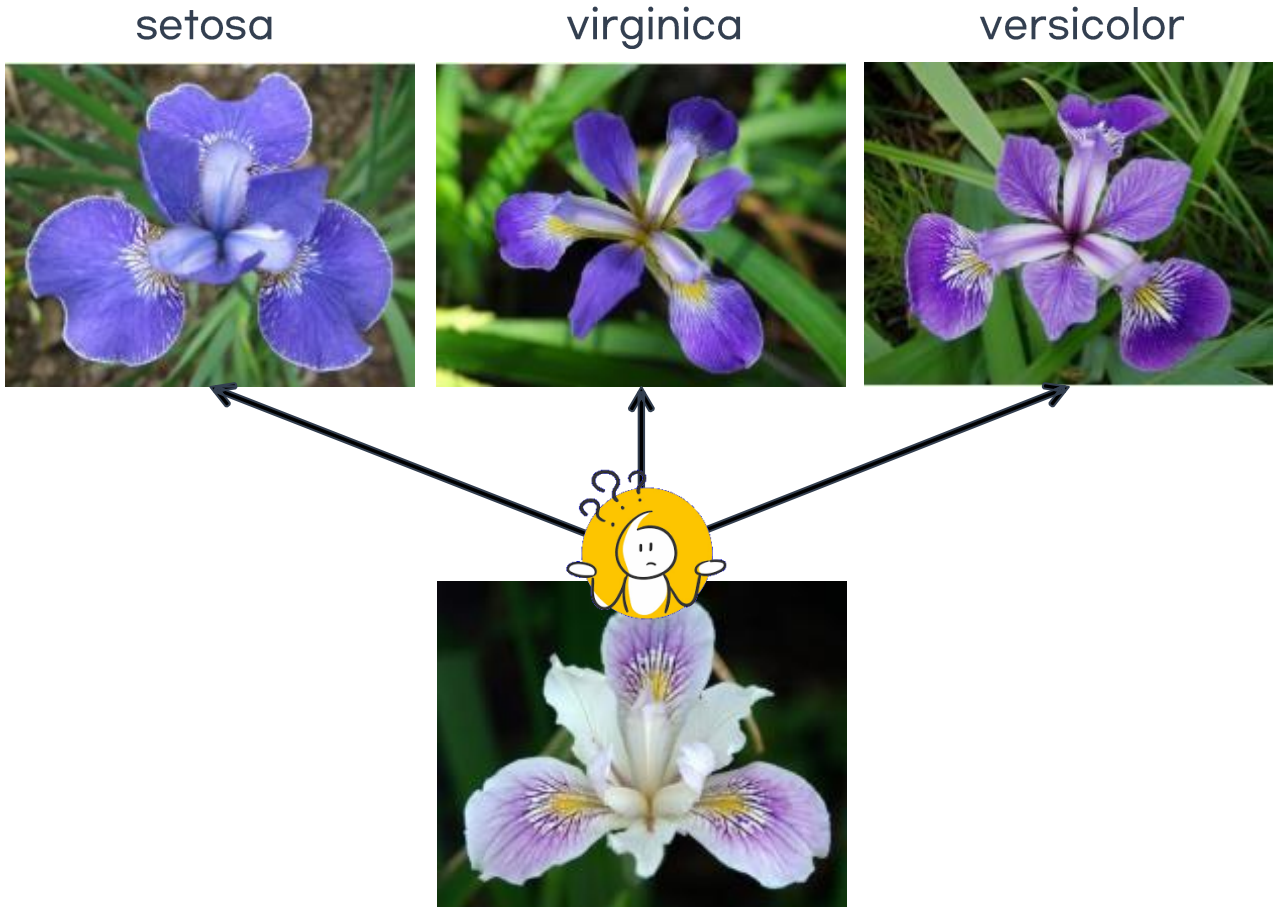
- n_neighbors : 이웃의 수 , metrics: 유클리디안 거리 방식
- 새로운 테스트 데이터 세트가 들어오면 훈련데이터 세트와의 거리를 계산
- 오류 데이터가 결과값에 크게 영향을 미치지 않음
- 훈련 데이터 세트가 크면(특성,샘플의 수) 예측이 느려짐

K-Nearest Neighbors(knn) 실습
k-최근접 이웃 알고리즘

Iris(붓꽃) 데이터를 이용한 붓꽃 품종 분류

K-Nearest Neighbors(knn) 실습

붓꽃(iris) 데이터 셋(scikit-learn 제공)



K-Nearest Neighbors(knn) 실습

붓꽃(iris) 데이터 셋(scikit-learn 제공)

- 150개의 샘플 데이터, 4개의 특성과 1개의 정답(3개의 품종)로 구성

	sepal_length	sepal_width	petal_length	petal_width	species
	꽃받침 길이	꽃받침 넓이	꽃잎 길이	꽃잎 넓이	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-setosa
...					
150	5.9	3.0	5.1	1.8	Iris-virginica

K-Nearest Neighbors(knn) 실습
k-최근접 이웃 알고리즘

유방암(Breast Cancer) 데이터를 이용한 유방암 악성/양성 분류 실습

K-Nearest Neighbors(knn) 실습

유방암(Breast Cancer) 데이터 셋(scikit-learn 제공)

- 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.

다음 시간에 만나요!

Decision Tree,
Label Encoding,
One-hot Encoding,
Cross validation