23 하계 학부연구생 프로그램

ch02. 데이터 다루기

인공지능공학과 12223547 박혜민

생선 이진 분류 모델

01

알고리즘 선정

- k-Nearest Neighbors

02

test set의 필요성

- train set 으로 정확도 평가가 가능할까?

03

샘플링 편향

- 기존 클래스 비율을 최대한 유지하는 방법

04

표준점수

- 단위가 다른 두 특성 데이터 다루기

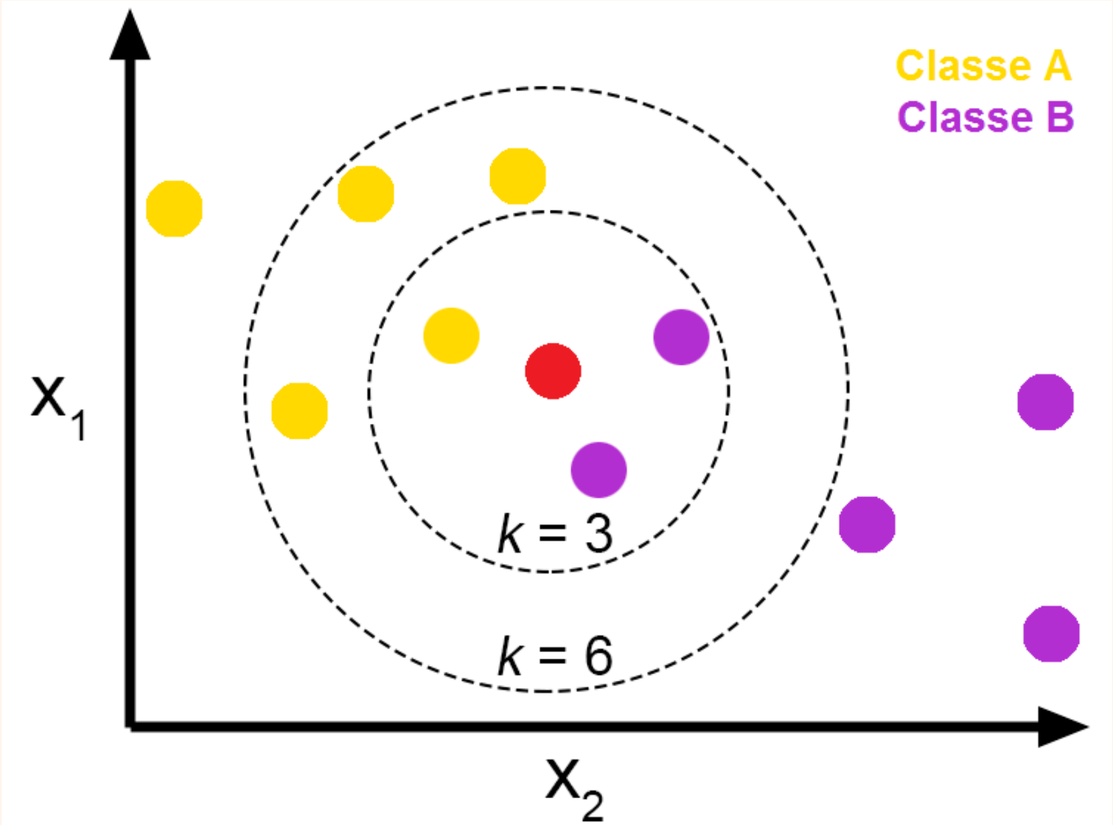
k-Nearest Neighbors

주변의 가장 가까운 k개의 데이터를 참고하여 다수의 레이블을 정답으로 사용

대표 패키지 : sklearn

특징

- 훈련과정 불필요
- 참고 데이터 수에 따라 예측값 변동
- 데이터 많은 경우 적합하지 않음



거리 계산

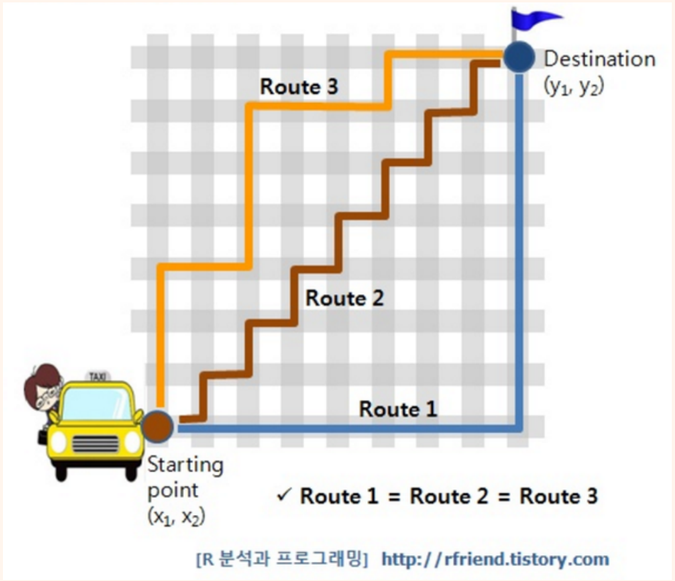
유클리드 거리 vs 맨해튼 거리

유클리드 거리

$$d(\mathbf{p}, \mathbf{q}) = d(\mathbf{q}, \mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

멘헤튼 거리

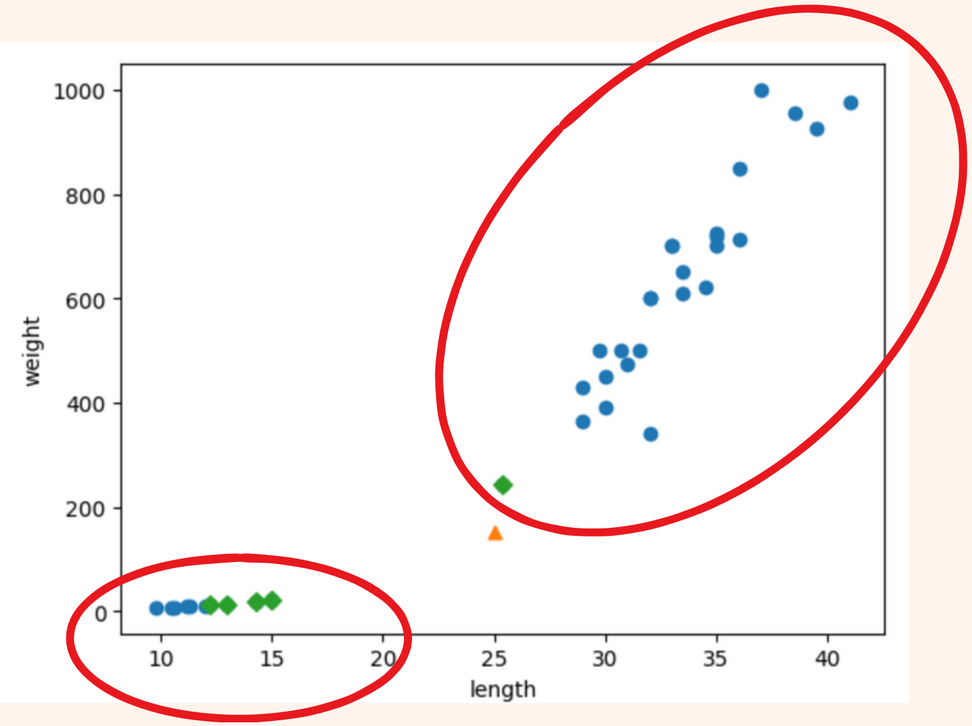


k 값 설정

적절한 k값 찾기

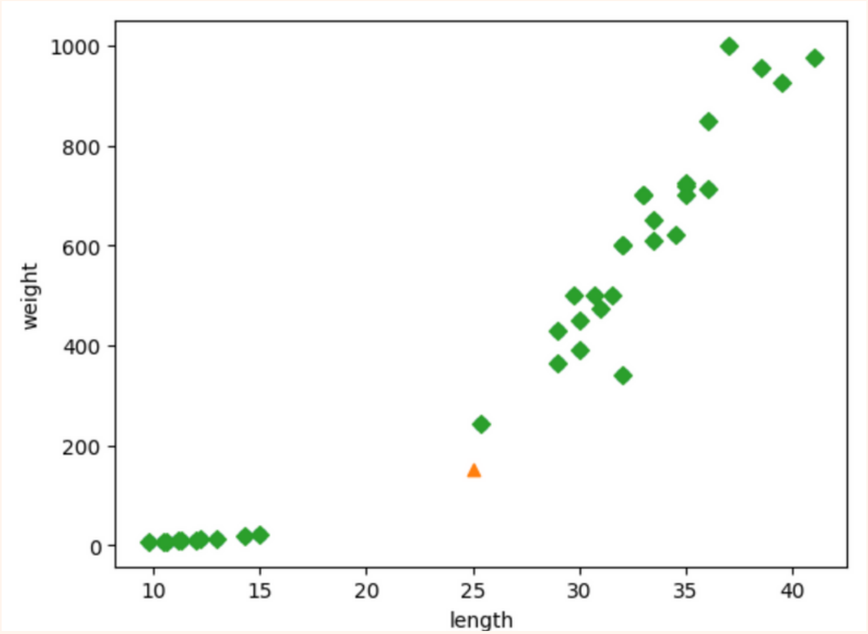
k = 5

class 2



class 1

k = 49



어떠한 데이터가 들어와도 class2로 분류하는 문제 발생

test set의 필요성

연습 문제와 시험문제가 같다면?

샘플링 편향

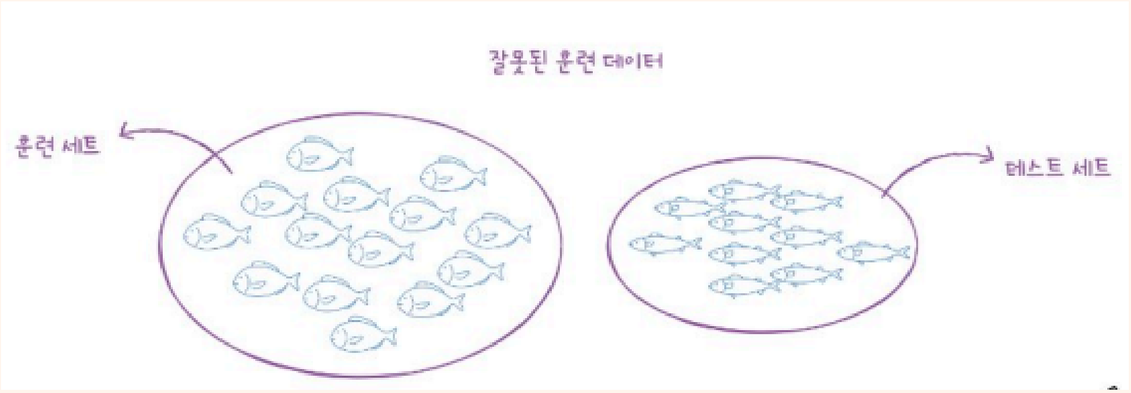
train set와 test set 에 샘플이 골고루 섞여 있지 않는 경우

문제점

- 테스트 샘플에 대하여 훈련 세트의 다수 클래스로 잘못 예측

기존 데이터의 클래스 비율을 유지하는 방법

- numpy 라이브러리
- train_test_split() 함수

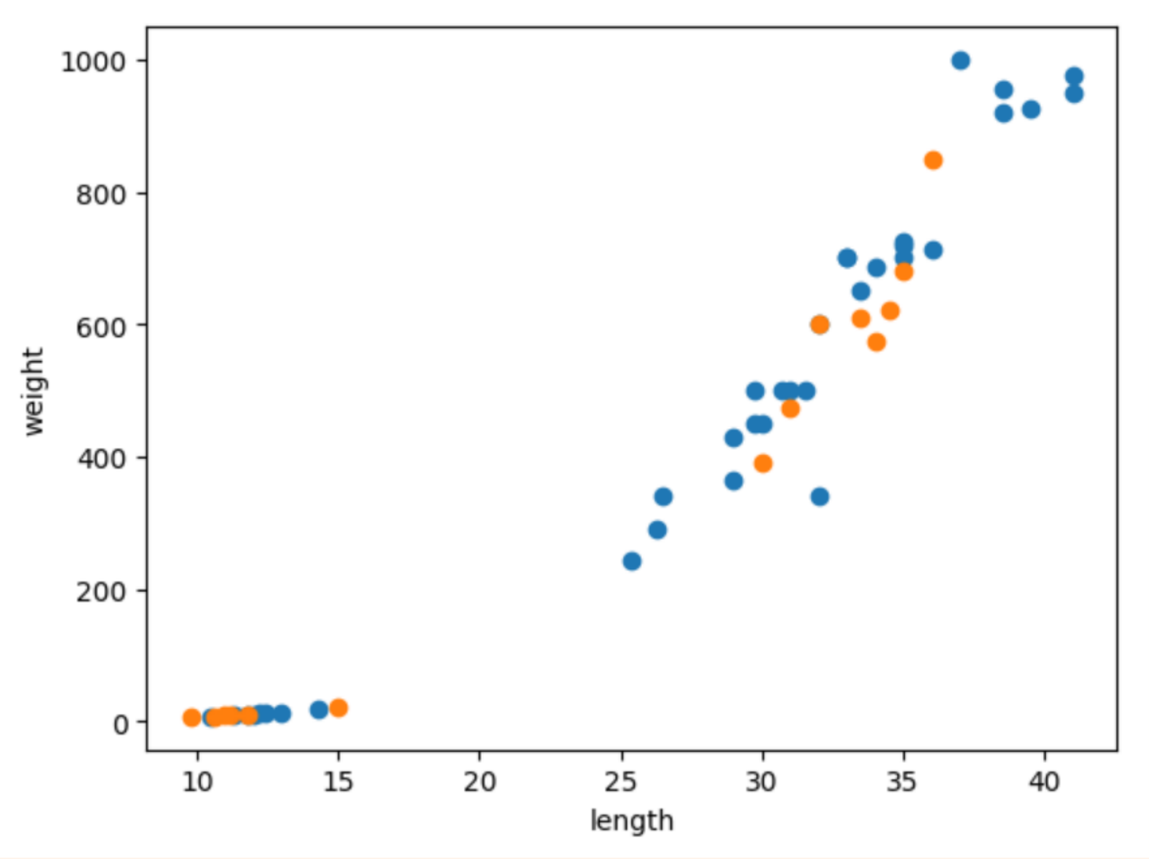


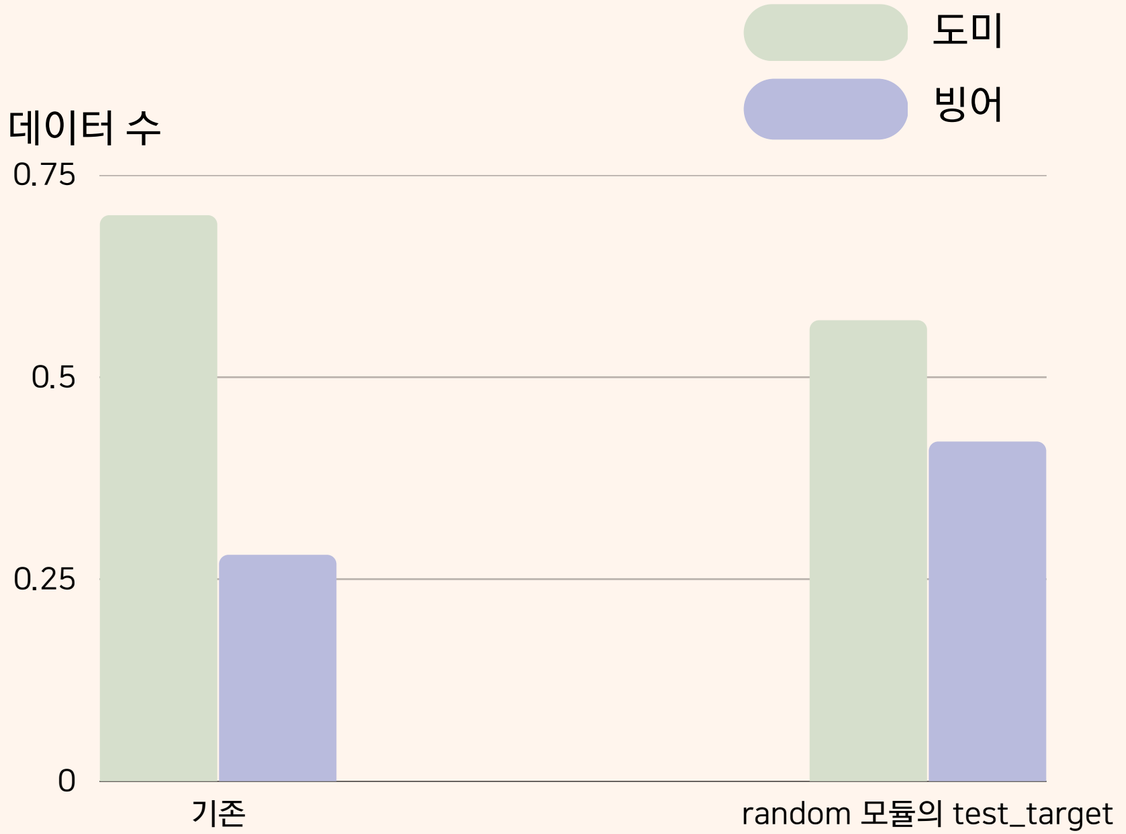

```
kn = kn.fit(train_input, train_target)
print(kn.score(test_input, test_target))
print(train_target)
```

```
0.0
[1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.
 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.]
```

클래스 비율 유지 - numpy

1. random 모듈의 shuffle 함수로 index 랜덤하게 배열
2. train set, test set 각각에 위 배열을 인덱스로 전달

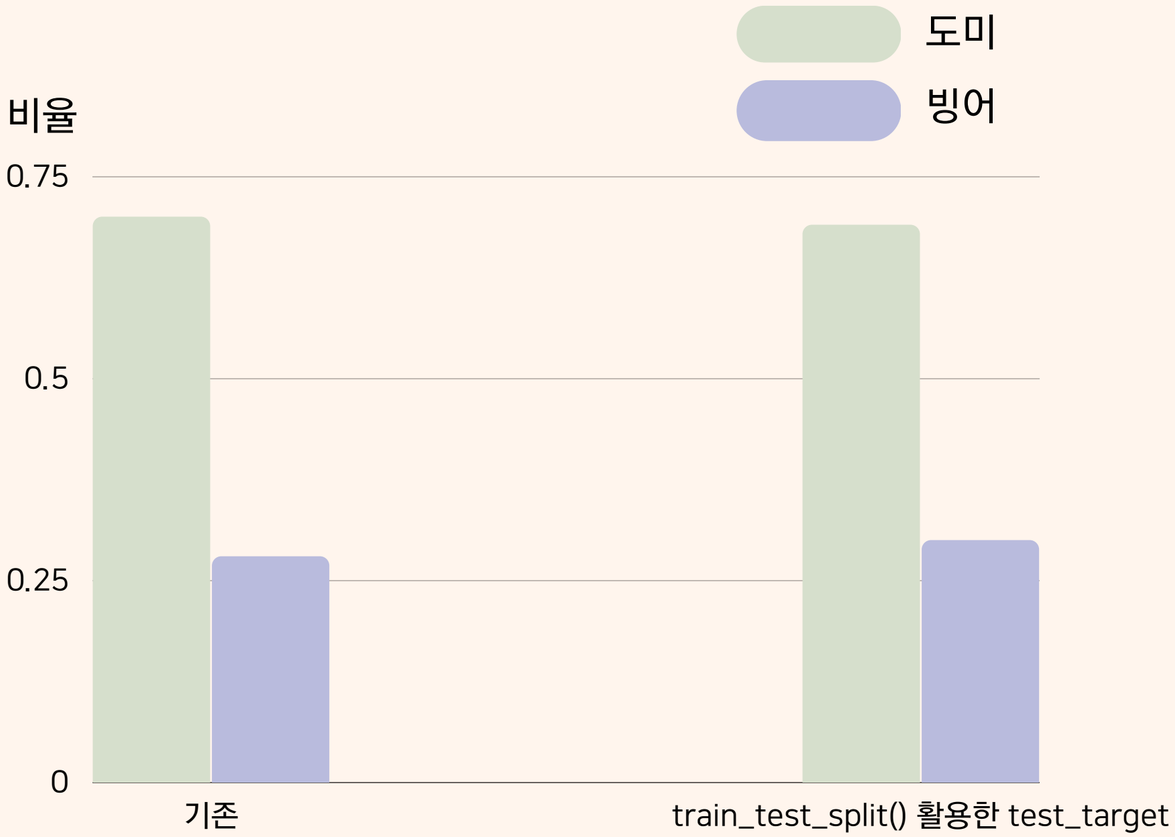




한계 : 기존 입력 데이터 수가 적어서 random 모듈 만으로는 비율 유지 어려움

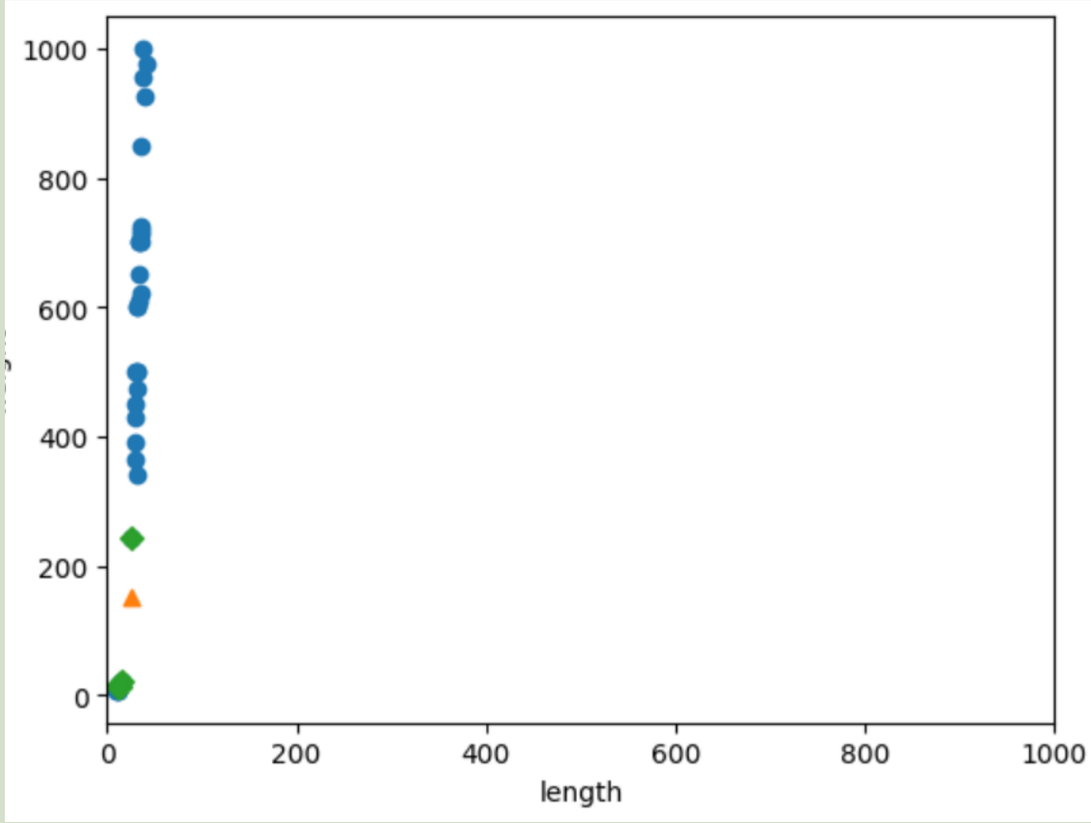
클래스 비율 유지 – train_test_split()

train_test_split() 의 stratify 매개변수에 타깃 데이터 전달



데이터 수 적어도 기존 클래스 비율 잘 반영

각 특성의 스케일이 다른 경우 문제점



도미 데이터의 최근점 이웃 5개 중 4개가 빙어

↓ 왜?

각 특성의 값들을 같은 범위의 값이라고 가정하고
거리 계산하기 때문

↓

특성값의 크기에 영향받지 않는 값 필요

표준 점수(z점수)

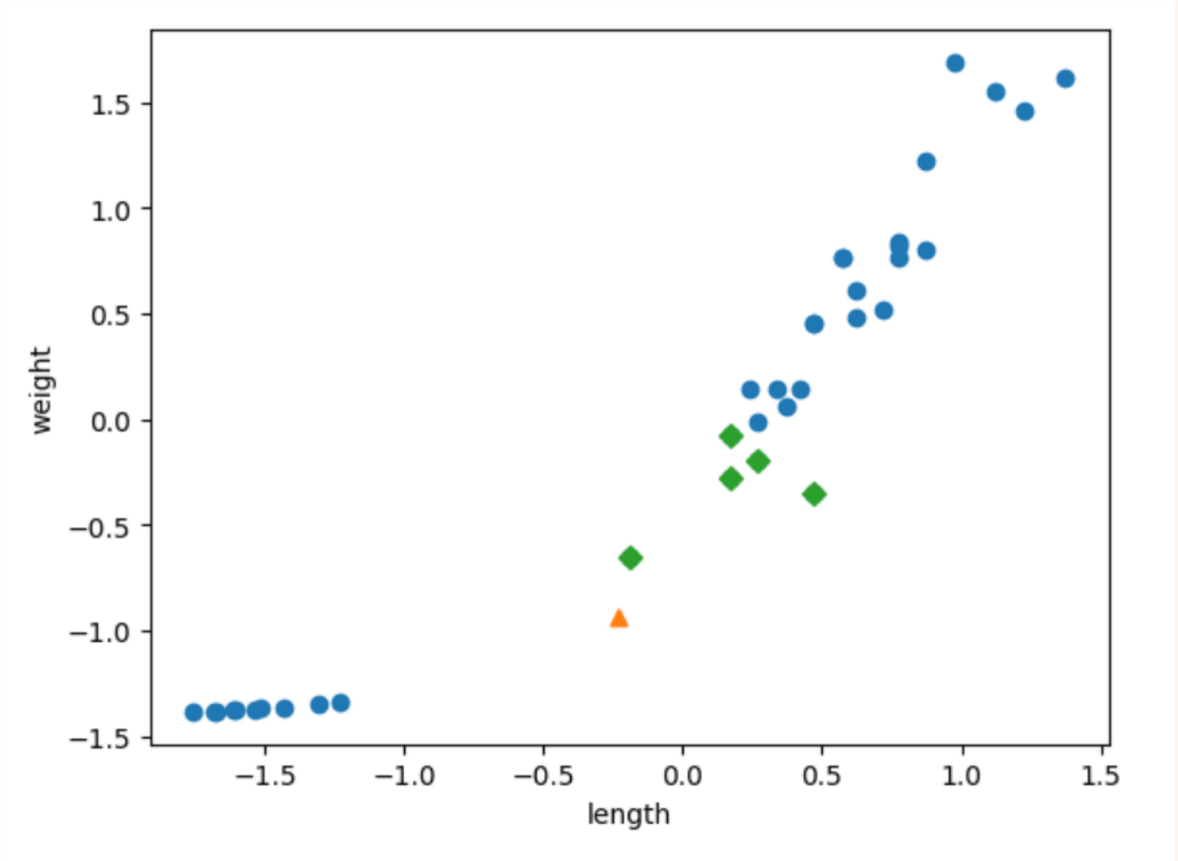
각 특성 값이 평균에서 표준편차의 몇 배 만큼 떨어져 있는지 나타내는 값

공식

- (데이터 - 평균)/표준편차

적용

- 훈련 셋 뿐만 아니라 예측 데이터, 테스트 데이터도 훈련 셋의 평균과 표준편차로 변환



https://bkshin.tistory.com/entry/%EB%A8%B8%EC%8B%A0%EB%9F%AC%EB%8B%9D-6-K-%EC%B5%9C%EA%B7%BC%EC%A0%91%EC%9D%B4%EC%9B%83KNN

http://news.bizwatch.co.kr/article/tax/2018/06/18/0023

http://www.fornurse.co.kr/news/articleView.html?idxno=11384