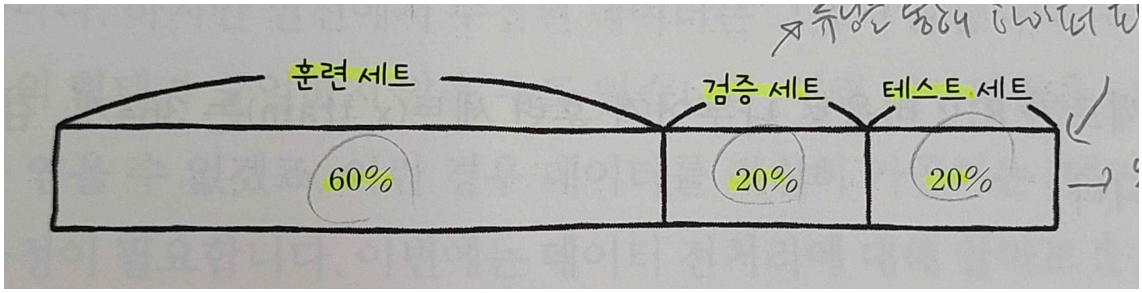
디지털 영상처리 연구실 연구보고서

정지우

#훈련 노하우

검증 세트를 준비하자

모델을 튜닝할 때, 테스트 세트를 사용하지 않고 검증세트를 사용한다.



=> 훈련세트를 80% -> 60%으로 줄이고 검증세트를 만든다.

```
cancer = load_breast_cancer()
x = cancer.data
y = cancer.target
x_train_all, x_test, y_train_all, y_test = train_test_split(x, y, stratify=y,
                                                    test_size=0.2, random_state=42)

[62] x_train, x_val, y_train, y_val = train_test_split(x_train_all, y_train_all, stratify=y_train_all,
                                                test_size=0.2, random_state=42)

print(len(x_train), len(x_val))

364 91

[65] sgd = SGDClassifier(loss='log_loss', random_state=42)
sgd.fit(x_train, y_train)
sgd.score(x_val, y_val)

0.6923076923076923
```

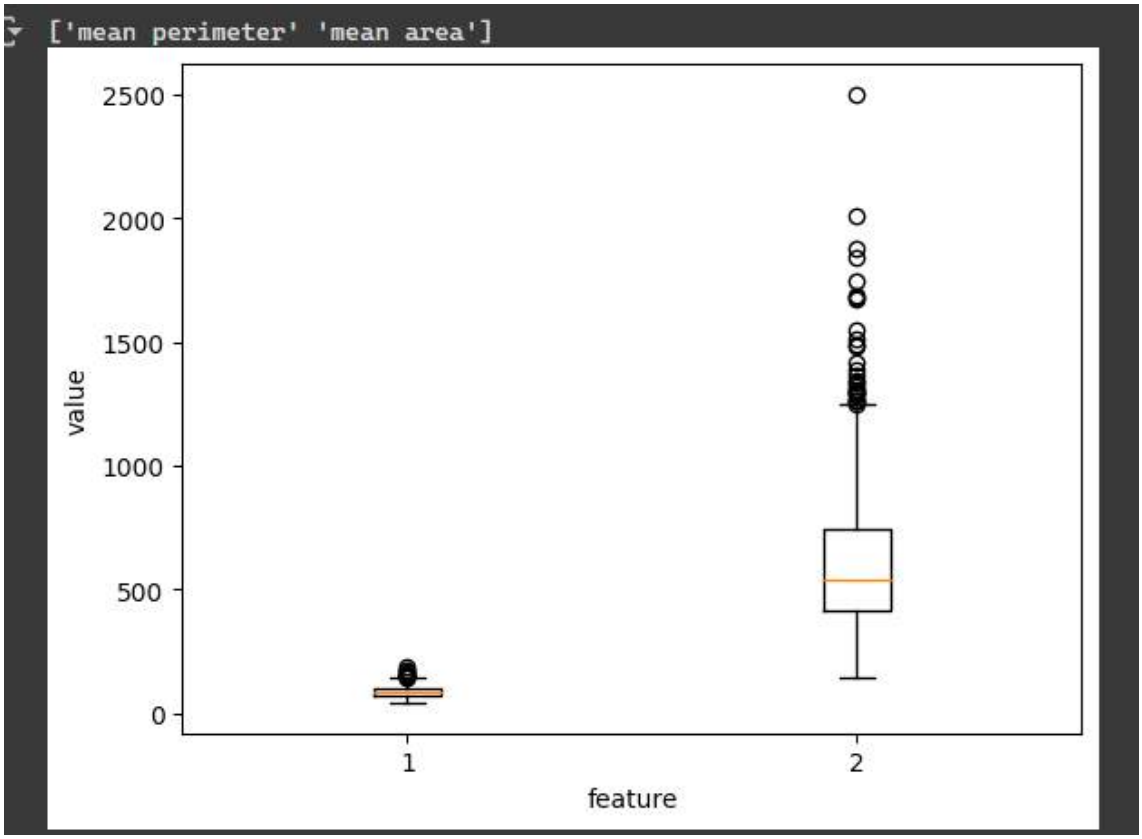
=> 첨에 8:2로 train과 test를 나누고 다시 train에서 8:2로 나눠서 검증세트(val)를 만들고 이를 통해 정확도를 측정함

=> 0.69%로 오히려 이전보다 낮아짐.

이유: 데이터 양이 작아서 그렇다. (정확히 모든 데이터량이 569개)

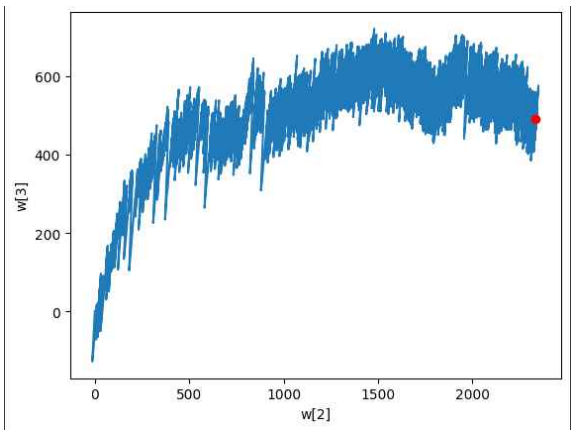
- 일반적으로 10만개 정도 있으면 8:1:1이 적당하다.

스케일을 조정해야한다.



2가지 x에 들어가는 파라메타가 존재.
근데 1번 특징은 200~400 사이인데 2번 특징은 200~2500까지 존재
=> 즉, 두 값의 크기의 범위가 너무 다름

- 100번의 에포크 동안의 변경된 가중치를 나타낸 그래프

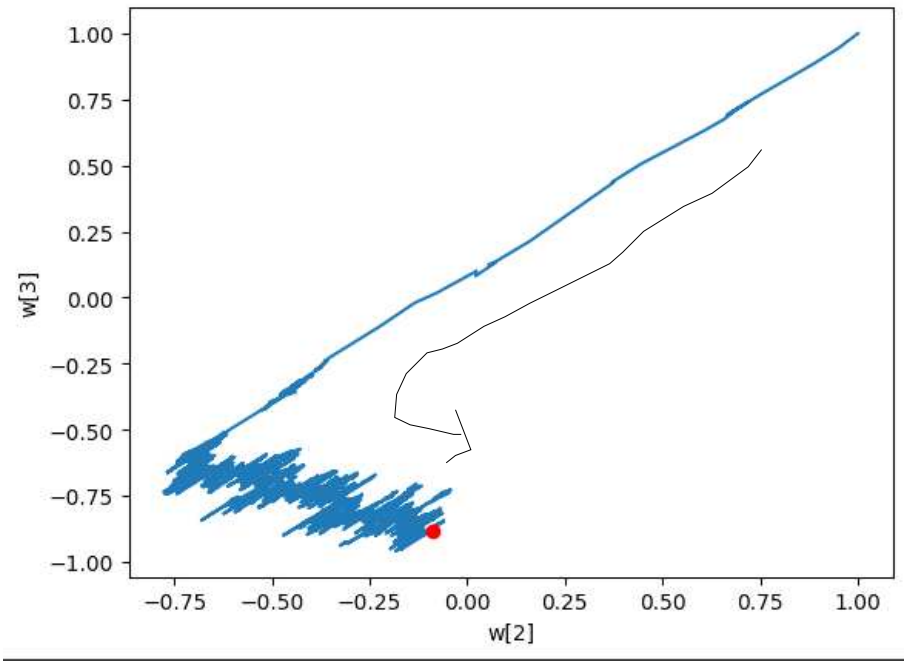


=> 2번 w 값은 스케일이 커서 조금씩 최적값에 가까워지지만, 1번 w값은 요동침.

표준화

$$Z = \frac{X - m}{\sigma}$$

이러한 공식을 통해 평균이 0이고 분산이 1인 특성이된다.



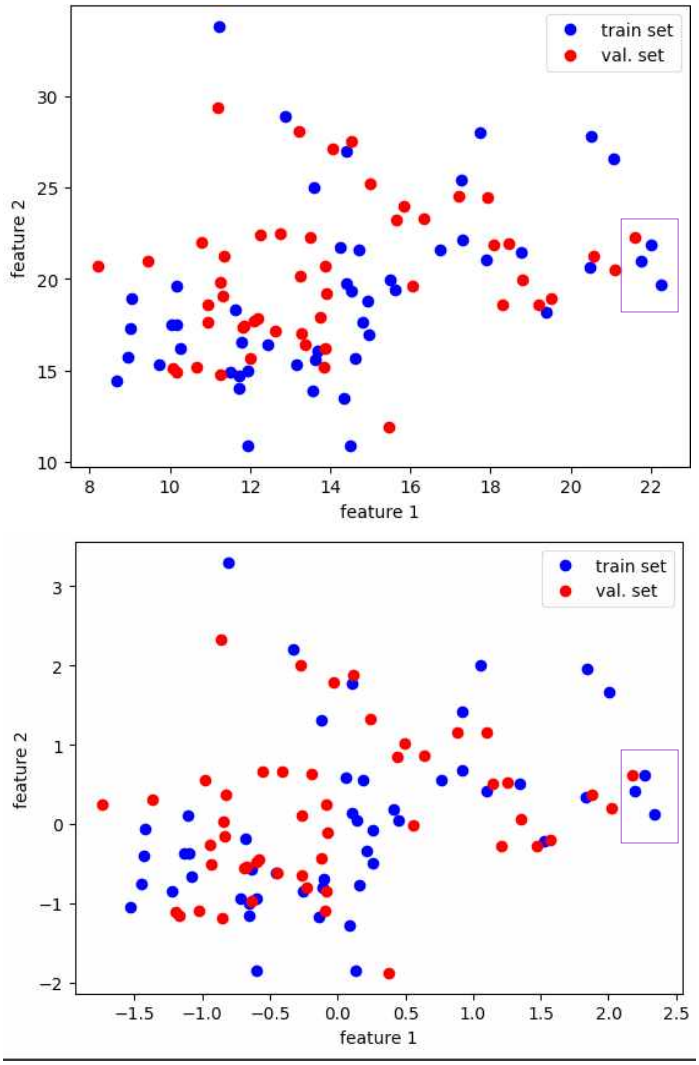
두 특성의 변화비율이 비슷해서 대각성 방향으로 가중치가 이동되어 최적값에 근접하는걸 알 수 있다.

```
[ ] val_mean = np.mean(x_val, axis=0)
    val_std = np.std(x_val, axis=0)
    x_val_scaled = (x_val - val_mean) / val_std

[75] layer2.score(x_val_scaled, y_val)

    0.967032967032967
```

(검증 세트도 표준화 시키고 테스트함)

=> 향상완료

##맹점



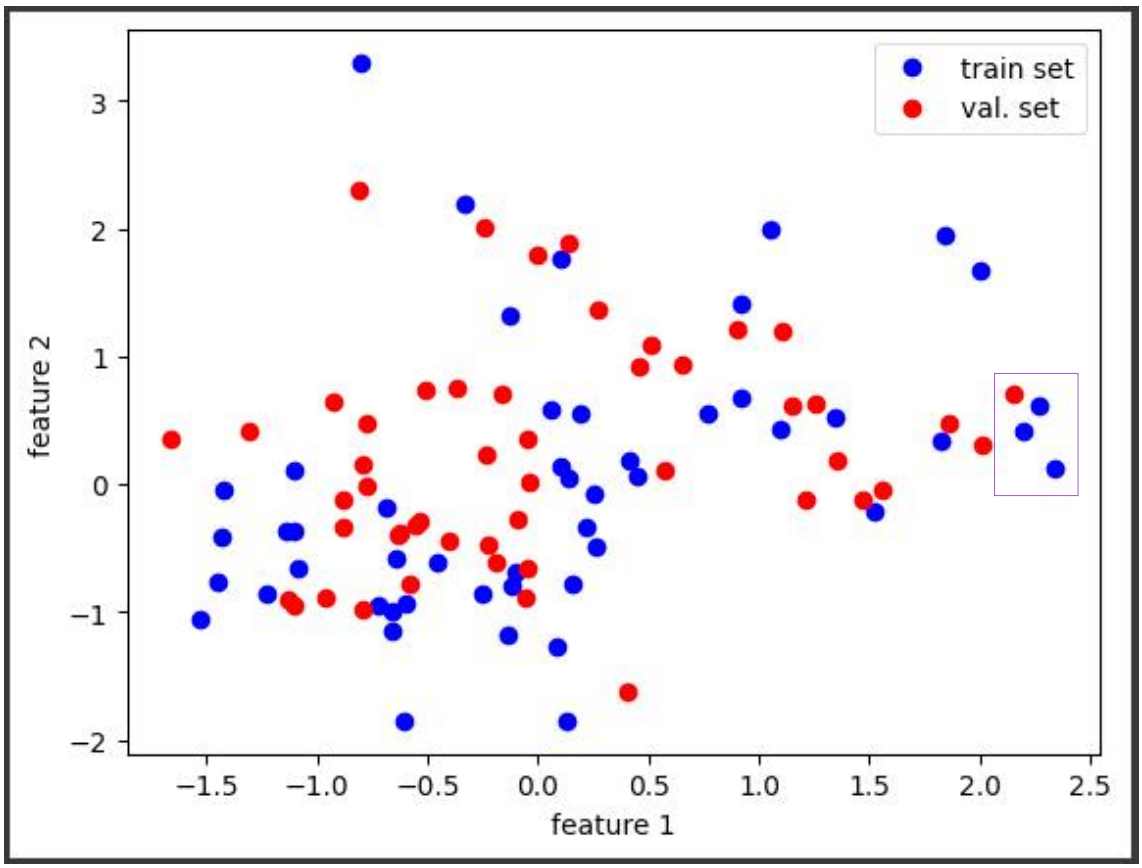
스케일 변화 전과 후의 특징들 사이의 거리가 유지 안됨

-> 즉, 다른 비율로 변환됨.

해결법

```
x_val_scaled = (x_val - train_mean) / train_std
```

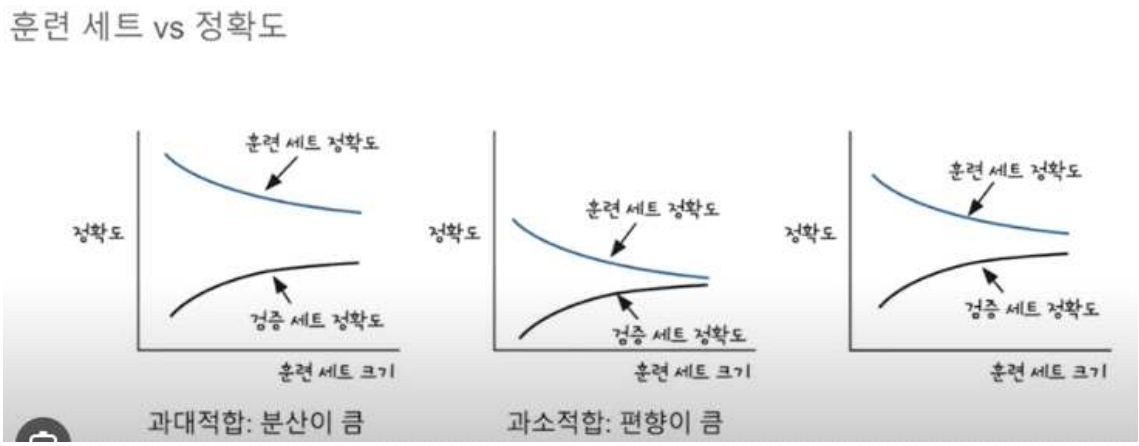
검증세트를 표준화 할 때, 훈련세트의 평균과 표준편차를 이용해서 계산한다.



```
layer2.score(x_val_scaled, y_val)
0.9670329670329670
```

=> 검증 값은 이렇게 하나 안하나 차이는 없지만 데이터 수가 적기 때문이고,
만일 데이터량이 많은데 전처리 잘못하면 성능차이 크게남

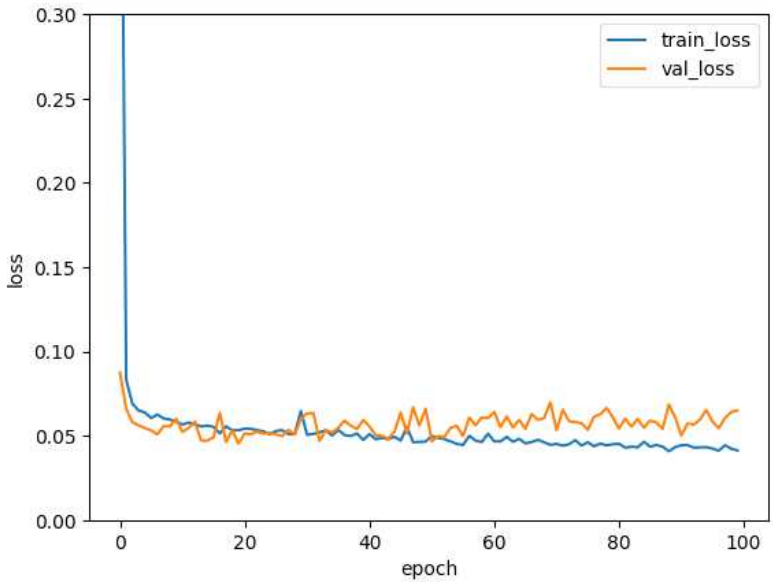
##과대적합 과소적합



=> 간단히 이야기해서 훈련세트에만 잘 맞는 형태로 만들어지면 훈련세트에서만 성능이 좋아지기 때문에 이런 이유로 과대적합 발생

과대적합: 훈련세트는 좋은데 검증세트하면 정확도 떨어지는거

과소적합은 모델이 복잡하지 않아(최적화가 제대로 수행되지 않아) 학습 데이터의 구조/패턴을 정확히 반영하지 못하는 문제



대략 20번의 에포크 이후 훈련세트에 더 알맞게 모델이 바뀌는거라서 검증세트에는 멀어지게 되는거다.

이러한 에포크를 수정하는것도 중요하다.

```
layer4 = SingleLayer()
layer4.fit(x_train_scaled, y_train, epochs=20)
layer4.score(x_val_scaled, y_val)
```

```
0.989010989010989
```

규제방법