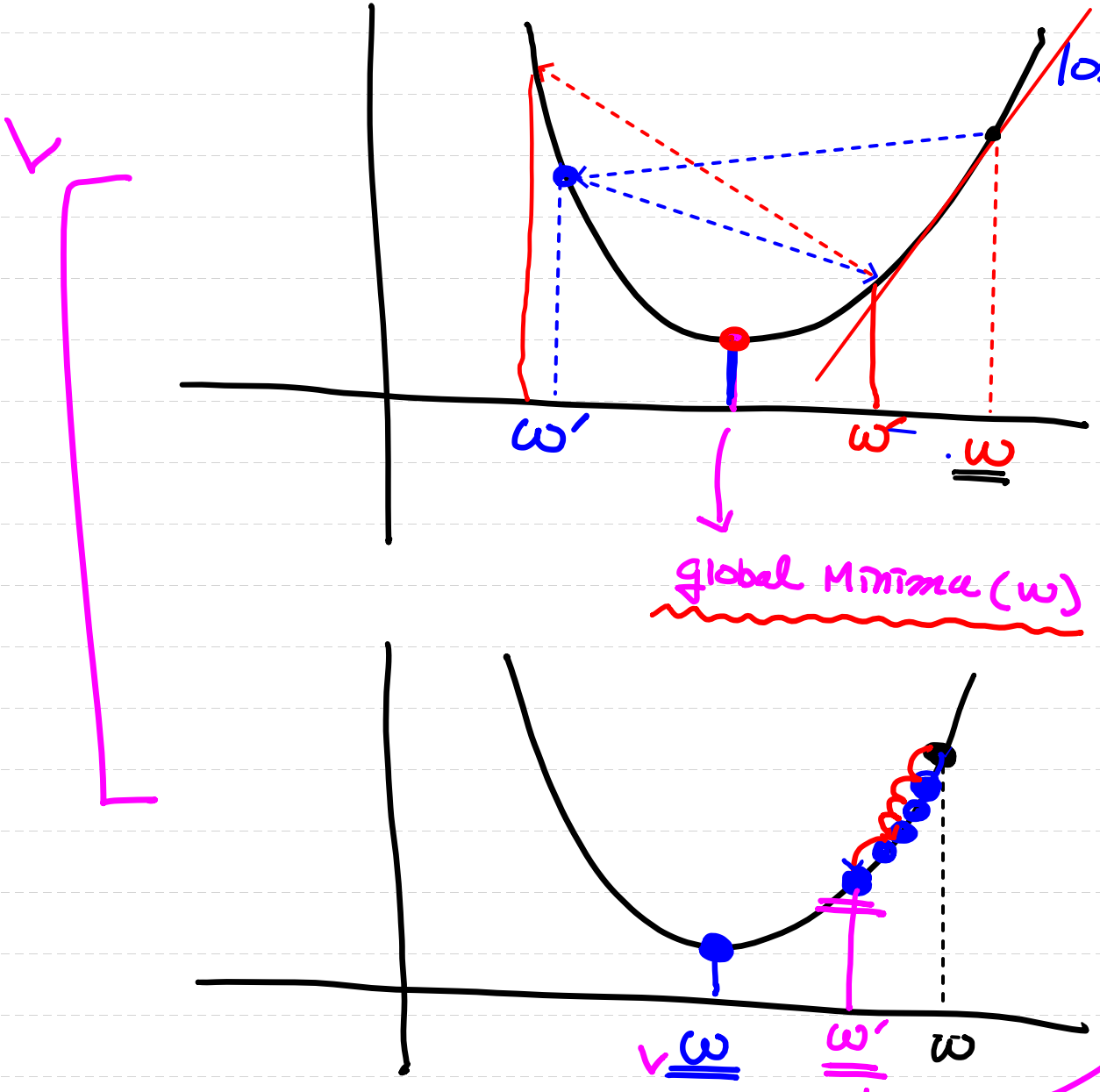
· 04/05

① learning rate (학습율) → $1e^{-4}$ (초기) ★

loss의 값을 보면서 learning rate를 조절



learning rate를 크게 잡으면

$$\omega' = \omega - \alpha \frac{\partial E(w,b)}{\partial \omega}$$

α 크다면

⇒ Overshooting 발생

α값이 아주 작다면 ??

② Normalization (정규화)

- MinMax Scaling (0~1)
- Standardization (표준화)

이상치의 상대적으로 덜 민감해요

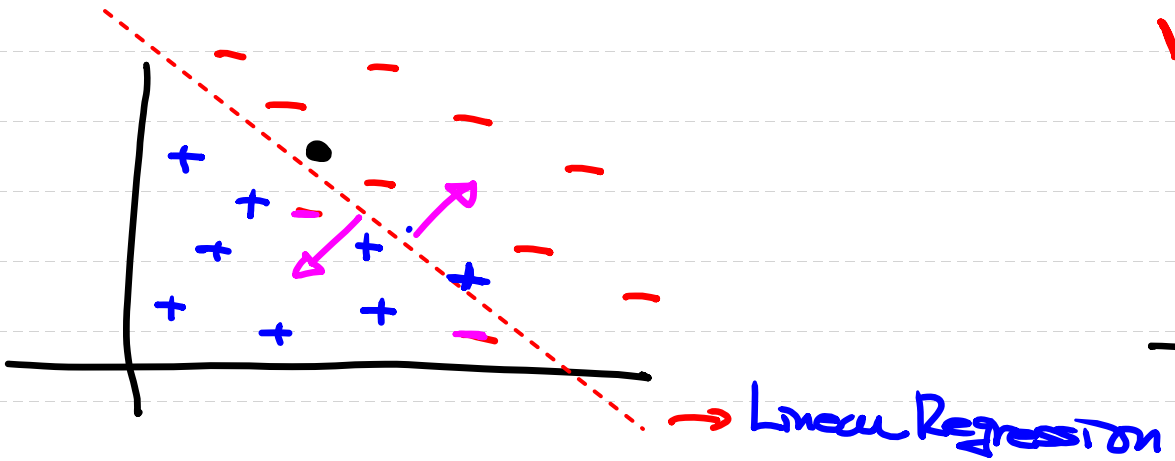
Scale이 통일하지 않아요.

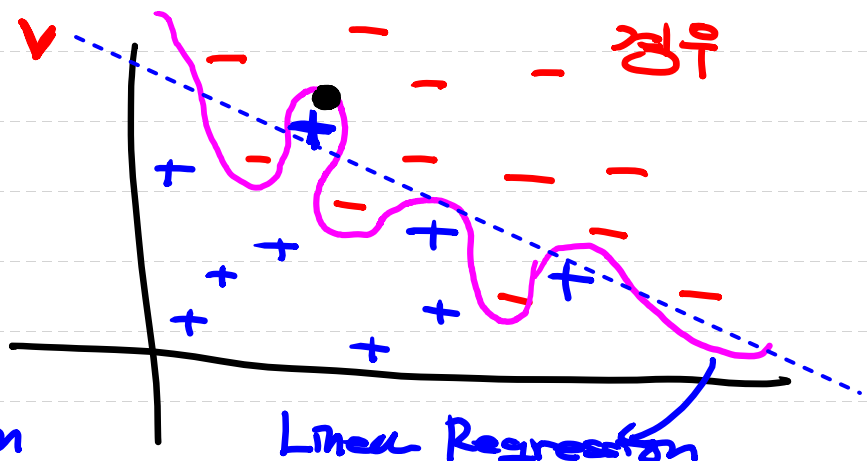
③ Overfitting (과대적합)

v Underfitting (과소적합)

과적합 → 이렇게 되지 않도록 처리

Overfitting은 Model이 Training Data Set에 너무 잘 맞춰져서 오히려 Test Data Set의 대해서는 정확도가 떨어지는 경우





Overfitting을 방지하려면 어떻게 해야 하나요?

① 많은 Training Data를 사용해서 학습
- 내가 가진 Data가 작다. → 데이터 증식(Augmentation)
- 하지만 Data가 충분하면 → resource를 많이 사용하고 학습시간이 상당히 길어져요

② Feature의 개수를 줄일 필요가 있어요.
독립변수
- 종속변수와 상관관계가 적은거.
- 중복된 feature을 삭제 → "데이터분석"

③ Weight (w)의 값이 너무 크지 않도록 조절 → 어떻게요??
- 값이크면 그래프가 구불구불하게 그려져요!
- → loss함수를 변화시켜서 w값을 조절
  - loss값을 의미적으로 크게해서 w값을 작게조절

④ Deep Learning에서
Dropout이라는 방법을 이용

Regularization (정제)
- L1 → Lasso Regression
- L2 → Ridge Regression

수식이 달라요!!

ⓐ Logistic Regression (Binary classification)

$$y = \frac{1}{1+e^{-(wx+b)}}$$

Code

$$\text{Cross Entropy (loss)} = -\sum_{i=1}^{n}\{t_i \log y_i + (1-t_i)\log(1-y_i)\} + \frac{1}{2}\alpha\sum_{i=1}^{n}|w_i|$$

L1 Regularization

$\frac{1}{2}\alpha\sum_{i=1}^{n}|w_i|$ ✓

Regularization ✓

Strength (내가 설정하는 parameter)

L2 Regularization

$\frac{1}{2}\alpha\sum_{i=1}^{n}|w_i|^2$ ✓

code로 표현!

tf.nn._l2_loss(w)

④ Over Sampling

데이터의 불균형을 어떻게 해소할 것인가?

↳ "Class에 속한 데이터의 개수에 차이가 있을때" (데이터의 비대칭문제)

→ "imbalanced Data problem"

[0 / 1] Class가 2개 → Class는 분류를 지칭

- Under Sampling ↳ Data가 개수가 충분히 많다면
- ★ Over Sampling

종속변수 (class) 2

0	1000000	→ 0 50000
1	100000	→ 1 50000

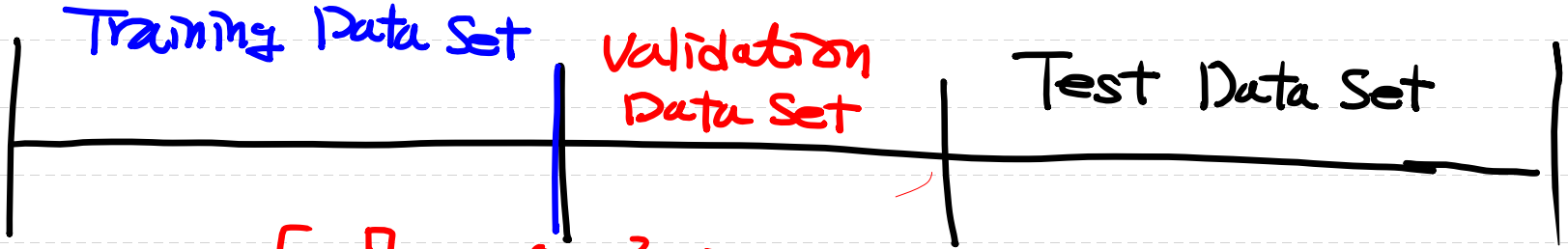
(데이터가 이렇게 많은 경우가 흔치않고 정상적인 데이터를 잃어버릴수 있어요)

↓

★★★ 데이터도 부족하고 편향되어 있을때 소수의 데이터를 복제

↳ "SMOTE 기법" ✓ 관련된 module도 있어요!!

⑤ k-Fold Cross Validation (vs. Hold-out validation)



[7 : 3]
[8 : 2]

↳ 보통 랜덤

[0 0 0 0 0 0 0 0 1 1 1] → 전처리 Data

X [0 0 0 0 0 0 0 0] [1 1 1]

↓ Training Data Set

↓ Validation Data Set

O [0 0 0 0 0 1 1] [0 0 1]

이런 처리는 Sklearn 이 담당!!

k-Fold Cross validation ⇒ 시간이 오래 걸린다는 단점

(5-Fold)

Sklearn