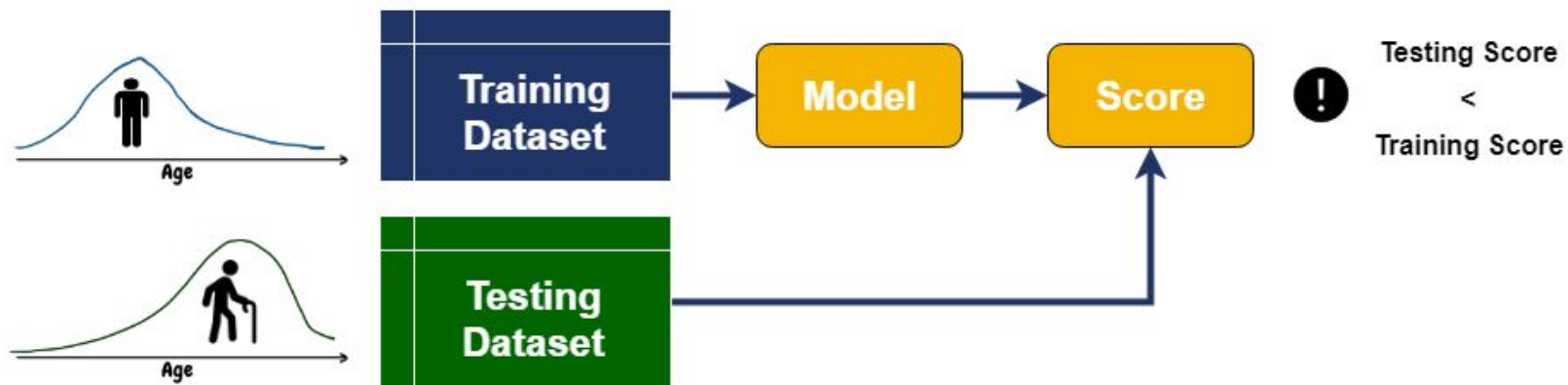


Dive into Deep Learning

3.16 환경

김 주 현

What is Dataset Shift?



Dataset Shift (Dataset Drift)

Covariate
Shift

Prior
Probability
Shift

Concept
Shift

공변량 변화(covariate shift)

공변량(covariates)(학습 데이터)의 분포가 테스트 데이터의 분포가 다른 상황을 의미

수학적으로 말하자면,

$p(x)$ 는 변화하는데, $p(y|x)$ 는 그대로 있는 경우를 의미

Covariate Shift는 cross validation을 수행할 때

많은 문제를 일으킨다

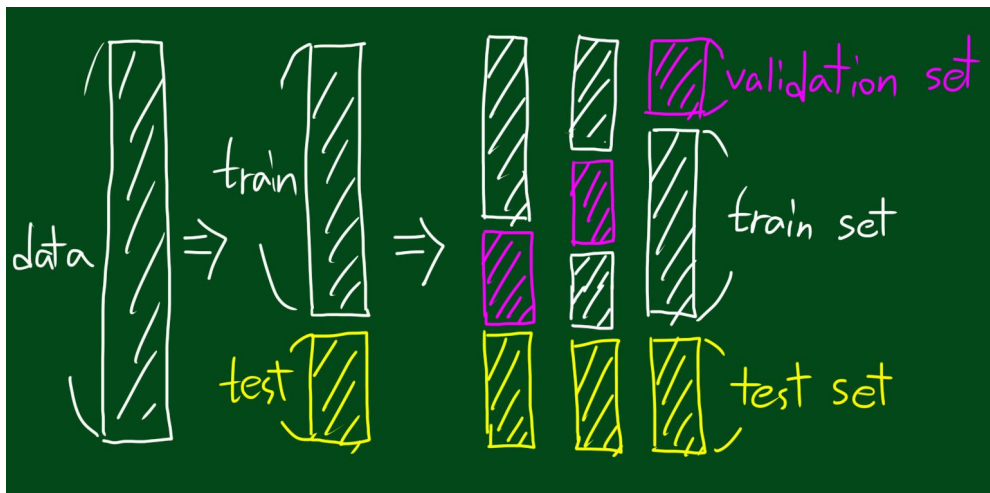
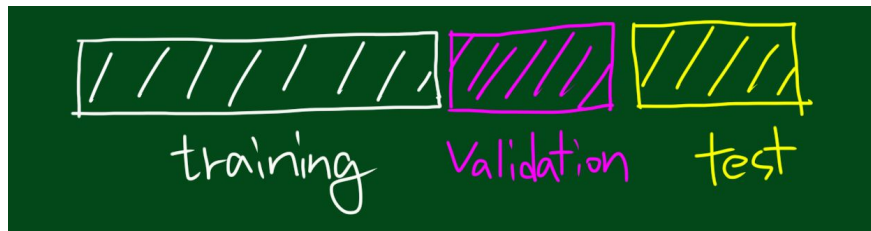
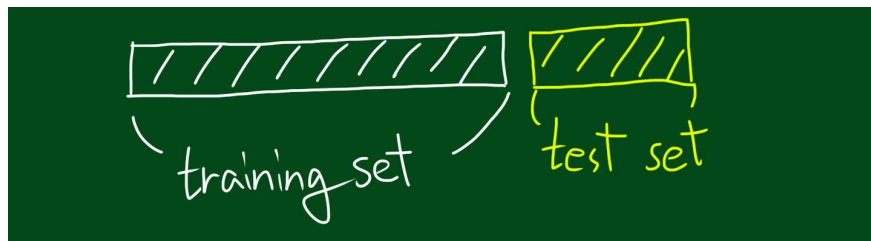
Why?

Cross Validation은 공변량 이동이 없으면 거의 **biased** 되지 않지만,

공변량 이동 시에 매우 **biased 된다!**

Cross validation

training data로 학습시키고,
validation set으로 parameter 튜닝 하고,
test set으로 최종 테스트



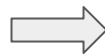
개념 변화(concept shift)

시간에 따라 시시각각 데이터 값이 달라지므로,

대상이 되는 분류모델 (target classification model)이 시간에 따라 계속 달라지는 현상

교통 카메라 렌즈

환경의 영향으로 서서히 성능 떨어짐

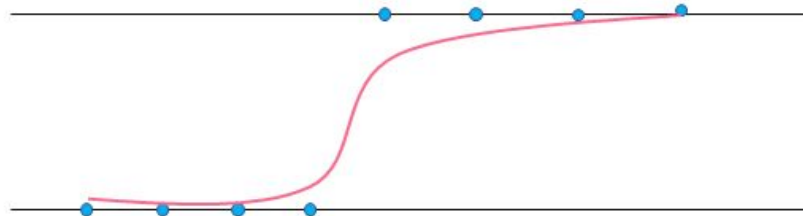


이미지 품질에 영향을 미침

공변량 (covariate shift) 교정

Importance Reweighting

로지스틱 회귀



$$f(x) = \frac{1}{1 + \exp(-f(x))}$$

로지스틱 회귀분석(logistic regression analysis)이란?

관찰값이 2개의 모집단 중 1개에 포함된 경우

테스트 대상이 어떤 모집단에 포함되어야 하는지 예측하는 통계적 분석방법

로지스틱 회귀를 이용하면 **확률 비율**을 계산해낼 수 있다!

$$\begin{aligned}
 p(z = 1|x) &= \frac{p(x)}{p(x) + q(x)} \\
 p(z = -1|x) &= \frac{q(x)}{p(x) + q(x)}
 \end{aligned}
 \quad \Rightarrow \quad
 \begin{aligned}
 \frac{p(z = 1|x)}{p(z = -1|x)} &= \frac{q(x)}{p(x) + q(x)}
 \end{aligned}$$

$$\begin{aligned}
 1) \quad p(z = 1|x) &= \frac{1}{1 + \exp(-f(x))} & \text{----- } p(x) \\
 2) \quad p(z = -1|x) &= \exp\left(\frac{-f(x)}{1 + \exp(-f(x))}\right) & \text{----- } q(x)
 \end{aligned}
 \quad \boxed{\frac{p(x)}{q(x)} = \exp(f(x))}$$

1. 두 분포에서 추출된 데이터를 구분하는 것
2. 두 번째는 가중치를 다시 적용한 최소화 문제

Prototype

1. 학습셋 $\{(x_i, -1) \dots (z_j, 1)\}$ 을 생성
2. 로지스틱 회귀(Logistic regression)를 이용해서 **이진(binary) 분류기**를 학습시킴
-> 함수 f 라 칭함
3. $\beta_i = \exp(f(x_i))$ 또는 $\beta_i = \min(\exp(f(x_i)), c)$ 를 이용해서 학습 데이터에 가중치를 적용
4. 데이터 X 와 이에 대한 레이블 Y 에 대한 학습을 수행할 때, **가중치 β_i** 를 이용

Summary

많은 경우에 학습셋과 테스트셋은 같은 분포로부터 얻어지지 않는다

이런 상황을 우리는 **공변량 변화(covariate shift)**라고 한다

공변량 변화(covariate shift)는 변화가 아주 심하지 않을 경우에 탐지하고 이를 교정할 수 있다

만약 그렇게 하지 못하면,

테스트 시점에 예상과 다른 결과가 나오므로 모델을 만들 때 이 점을 고려하여야 한다