# 교차검증

# 1. Holdout 이란?

- 데이터를 두 그룹으로 나누는 것
  - Training set : 모델에 학습을 위해서 사용되는 Dataset
  - Test set : 학습된 모델의 에러율을 측정하기 위해서 사용되는 Dataset

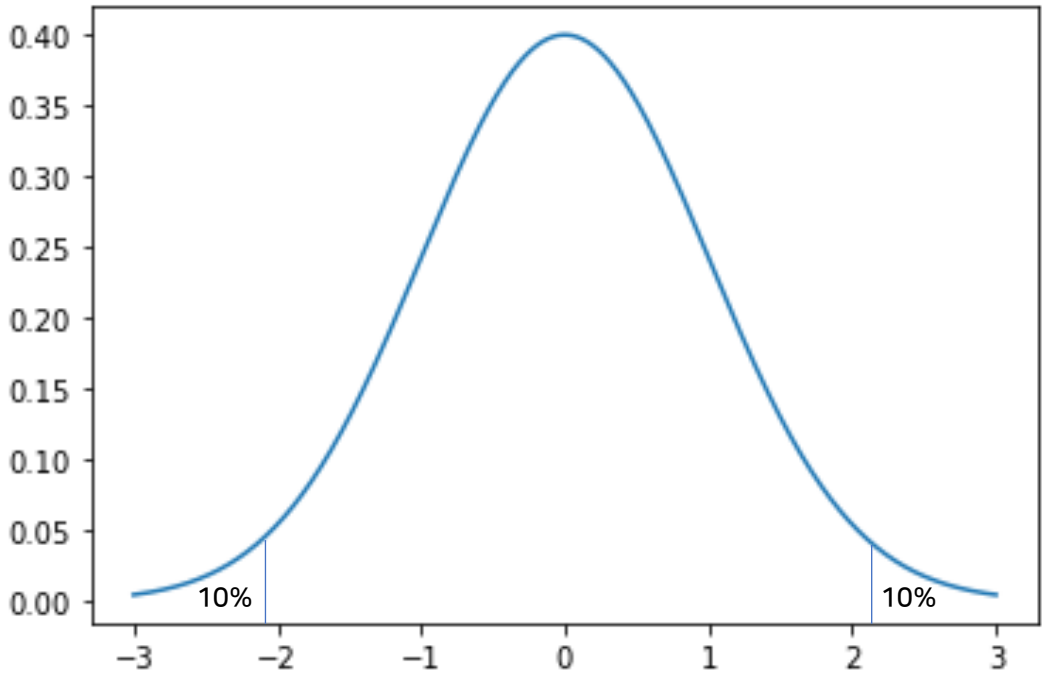
- 필요성
  - 모델의 과적합 정도를 측정 및 예측하는 척도가 됨

# 2. Holdout 적용?

- Scikit-learn의 “train-test-split”을 적용

# 2. K-Fold Cross Validation 이란?

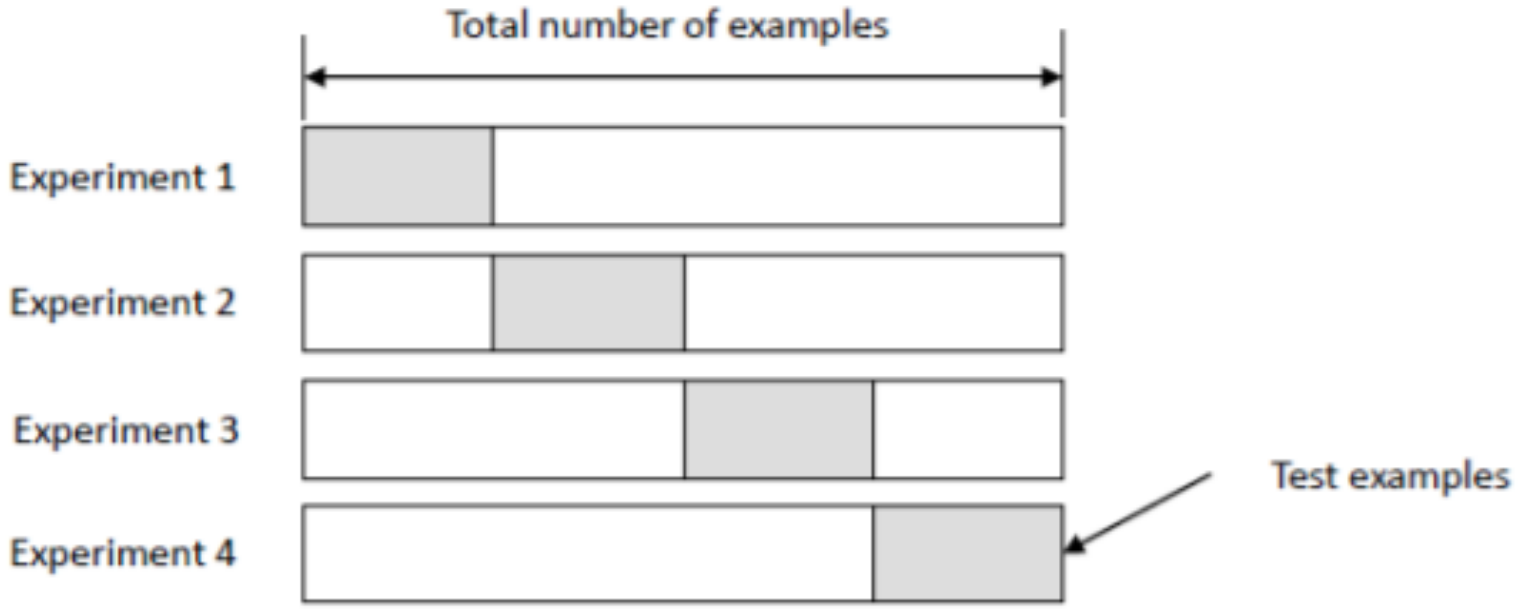
- 기존 Holdout의 한계점?
  - 20% 데이터의 낭비
  - 운이 안좋게도 Test Data가 잘못 분류되면 (극한에 해당하는 값들만 뽑게 되면) 잘못 학습할 수 있음



# 2. K-Fold Cross Validation 이란?

- K-Fold Cross Validation
  - Test 데이터를 중복되지 않게 K개의 Dataset 생성

- K-Fold Cross Validation 장점
  - 모든 데이터를 최소한 한 번씩 사용
  - Error도 평균치로 적용하게 됨

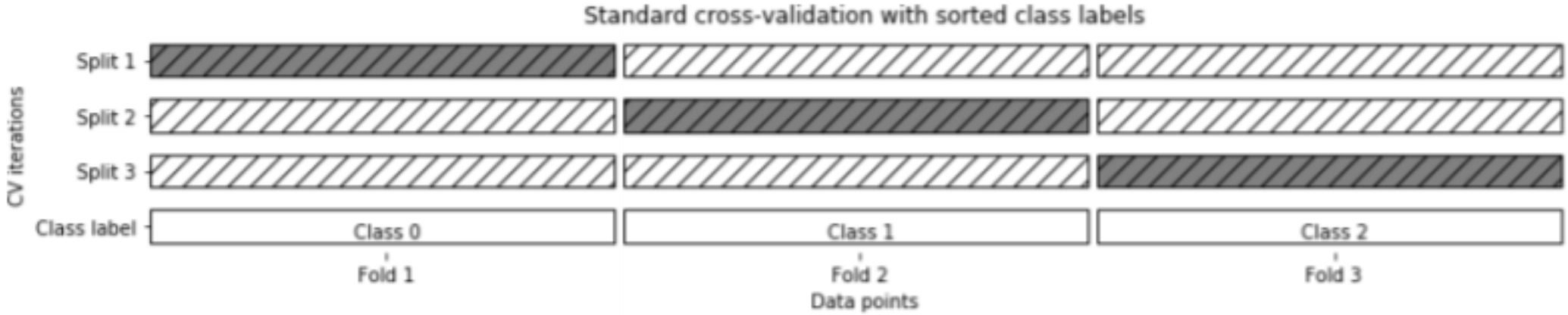



< IF K가 4일 때 >

# 3. Stratified KFCV 이란?

- KFCV의 한계점?
  - 운이 안 좋으면 Target Data가 편향되어 추출될 수 있다.

  - EX) 50명의 남성, 50명의 여성을 Holdout으로 20%의 Test Dataset을 만들어 낸다고 가정하자.
  - 통계적인 확률에 의해서 20%의 Test Dataset에도 남녀의 성비는 1:1이 될 것으로 추측된다.
  - 하지만, 20% Test Dataset이 모두 남성, 혹은 모두 여성이 될 가능성도 존재한다.



# 3. Stratified KFCV 이란?

- Stratified KFCV?
  - KFCV로 Train, Test Dataset을 만들어낼 때, Target Class의 비율도 맞추어 적용