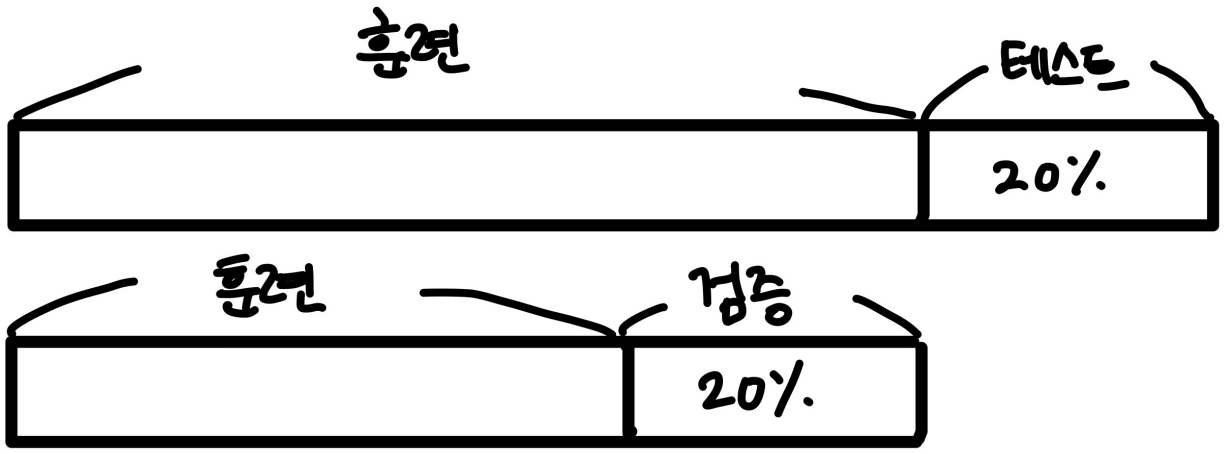
Cross Validation

Why?

훈련시킬 데이터의 개수가 부족하면 머신러닝 모델이 다양한 패턴을 학습하지 못하기 때문에 성능 높은 모델을 만들기 어렵다. 현실에서는 원하는 데이터를 충분하게 확보하기 어려운 경우가 많기 때문에 데이터의 특성은 유지하면서 데이터의 양을 늘리는 기법이 필수적이다. **교차검증**도 그러한 방법 중 하나이다.

How?

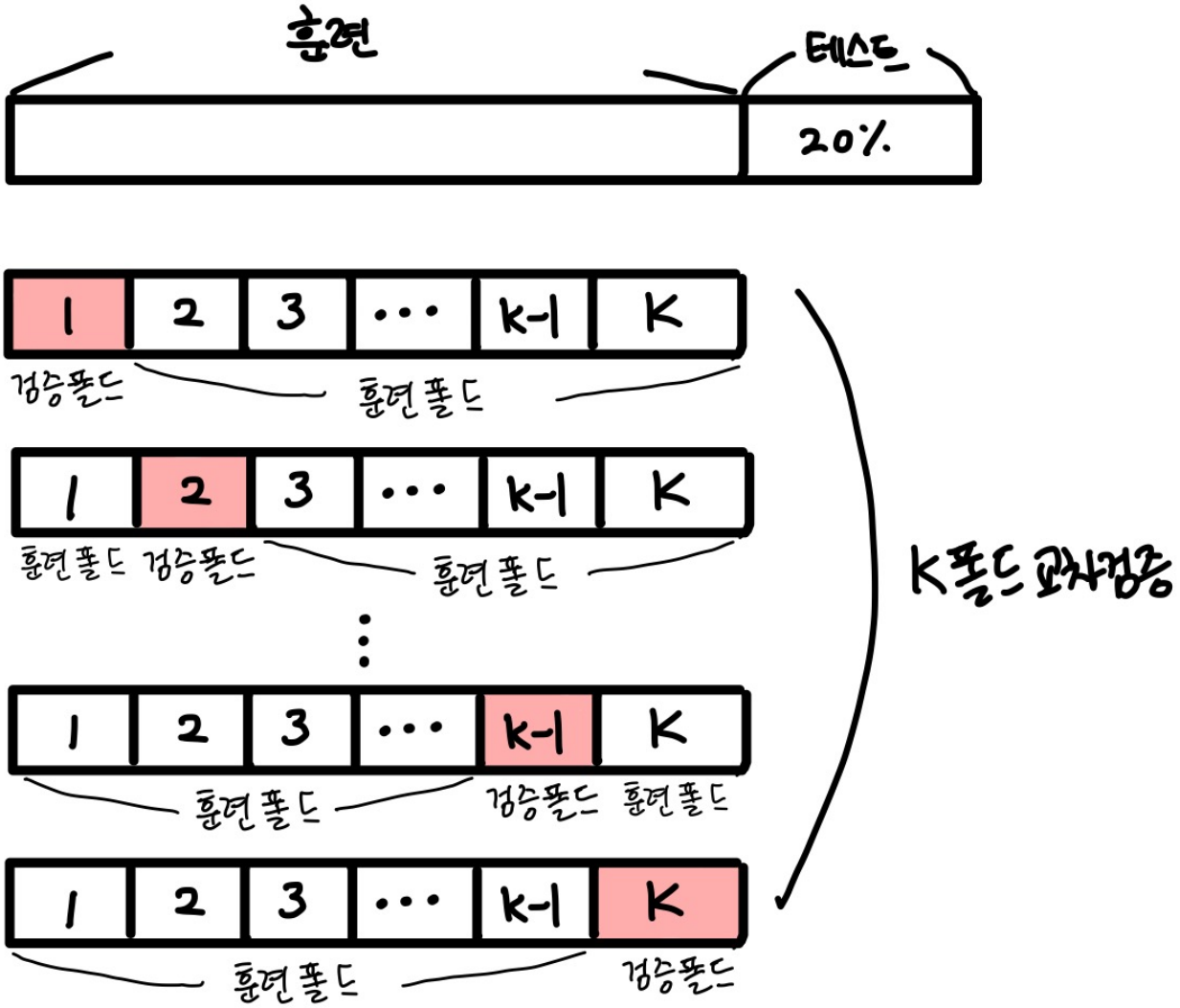
기존방식



기존 데이터 분할 방식

이전 포스팅까지 사용했던 방식은 전체 데이터를 8:2로 나눠 훈련세트를 얻은 후, 이 훈련세트를 다시 8:2로 분할해 검증세트로 사용했다. 전체 데이터가 100개라고 가정하면 60개의 데이터만 훈련에 사용할 수 있던 셈이다.

K-폴드 교차검증 방식



K-폴드 교차검증 방식

교차검증은 전체 데이터 세트를 8:2로 나눈 다음 8에 해당하는 훈련세트를 다시 k개의 작은 덩어리로 나누다. 그런 다음 작은 덩어리 1번씩 검증에 사용하고 나머지 덩어리를 훈련에 사용한다. 이때 한 덩어리를 폴드라고 하고, k개의 폴드로 나눈다고 하여 k-폴드 교차검증이라고도 한다. 전체 데이터 개수가 100개, K가 10이라고 가정하면 10개의 폴드가 생기므로 90개의 샘플을 훈련할 수 있다.

원리

> \1. 훈련세트를 K개의 폴드로 나눈다.
> \2. 첫 번째 폴드를 검증 세트로 사용하고 나머지 폴드(K-1)를 훈련세트로 사용한다.
> \3. 모델을 훈련한 후 검증세트로 평가한다.
> \4. 차례대로 다음 폴드를 검증세트로 사용하여 반복한다.
> \5. K개의 검증세트로 K번 성능을 평가한 후, 계산된 성능의 평균을 내어 최종 성능을 계산한다.