Data Science

교차검증 (Cross Validation)

n-fold cross validation(n 겹 교차 검증)

목적

- 분류 모델, 모델 생성 알고리즘, 데이터 셋의 일반화(generalized) 성능 측정
- 가용 데이터 양이 적을 경우를 극복하고자 함.

방법

- 데이터 셋을 n 개의 partition으로 무작위 추출 방법으로 분할
  (partition 간에 사례의 중복이 없도록 분리함)
- Partition 중 n – 1 개를 학습용 데이터 셋으로 사용하여 모델 생성
  나머지 하나의 partition 을 시험용 데이터 셋으로 사용하여 모델 시험
- 위 방법으로 총 n 개의 다른 모델을 생성하여 시험할 수 있음.
- 모델의 최종적인 성능은 n 개의 모델의 성능을 평균하여 구함.

예) 10-fold CV:

9개/1개의 partition 을 학습/시험용 데이터로 사용, 총 10개의 모델 시험

n-fold cross validation(n 겹 교차 검증)

Leave-one-out method

- N 개의 사례를 가지는 데이터 셋에서
- N – 1 개의 사례를 학습용 데이터로 모델 생성
  나머지 1개 사례를 시험용 데이터로 사용 모델 시험
- 총 N 개의 모델을 생성하여 일반화된 성능을 측정