



부산대학교

PUSAN NATIONAL UNIVERSITY

Cross-Validation

One Split Is One Sample

Measuring Error You Can Trust

BX3600087-140 / DS3600087-140 / AB3600087-140

기계학습개론

2026학년도 2학기

오용경

데이터사이언스전공

yongkyung.oh@pusan.ac.kr

Section 1 • Cross-Validation

One Split Is One Sample

한 번 나눈 결과는 우연일 수 있다

한 번 나누어 재기



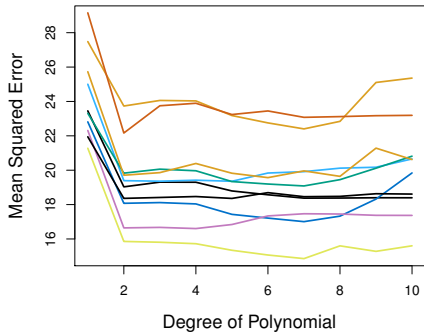
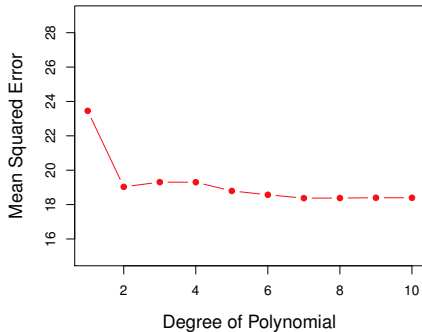
관측을 학습 자료(파란색)와 검증 자료(베이지색)로 나누고, 검증 자료에서 예측 성능을 평가한다

Source: James et al. (2023), Figure 5.1.

자료 분할이 바뀌면 학습과 평가도 달라진다

달라지는 것	검증 점수에 미치는 영향
학습 관측	학습된 계수와 예측값이 달라질 수 있다
검증 관측	평가하는 관측의 난이도와 구성이 달라진다
검증 점수	같은 학습 방법이라도 분할마다 달라질 수 있다

반복해 나눠 보면



오른쪽은 Auto 자료를 열 번 다르게 나눈 결과다. 같은 차수에서도 검증 MSE가 달라진다. 여러 fold를 평균하는 교차검증에서는 이 흔들림이 훨씬 작다

Source: James et al. (2023), Figure 5.2, Section 5.1.3.

한 번의 분할이 갖는 두 한계

문제	평가에 미치는 영향
분할에 따른 변동	학습·검증 자료를 어떻게 나누느냐에 따라 점수가 달라진다
적은 학습 자료	전체 자료로 학습할 때보다 성능을 낮게 추정할 수 있다

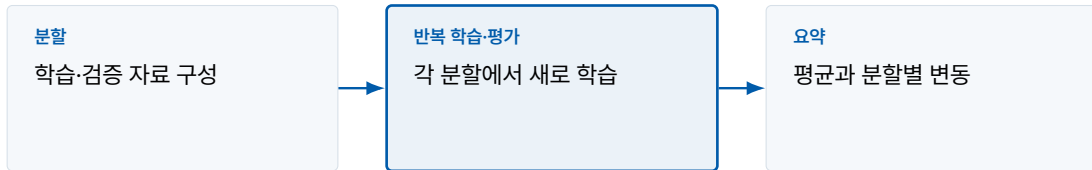
하나의 검증 점수만으로 분할에 따른 변동을 알 수는 없다

여러 분할에서 검증 성능을 요약한다

DEFINITION

교차검증(cross-validation)

학습·검증 자료를 나누는 과정을 반복하여 검증 성능을 요약하는 평가 방법



학습 자료가 적으면 오차를 높게 추정할 수 있다

학습에 쓴 자료	학습 결과	비교할 때의 주의점
일부 자료	적은 관측으로 학습한 모델	자료 부족으로 예측 성능이 낮아질 수 있다
개발 자료 전체	최종적으로 사용할 모델	일부 자료로 학습한 모델과 성능이 다를 수 있다

개발 자료는 테스트 자료를 제외한 학습·검증 자료를 뜻한다

Checkpoint

One Split Is One Sample

- 같은 자료를 다르게 나누면 점수가 왜 달라지는가
- 학습 자료가 적으면 검증 오차가 높아질 수 있는 이유는 무엇인가
- 한 번의 검증 점수로 분할에 따른 변동을 알 수 있는가

Section 2 • Cross-Validation

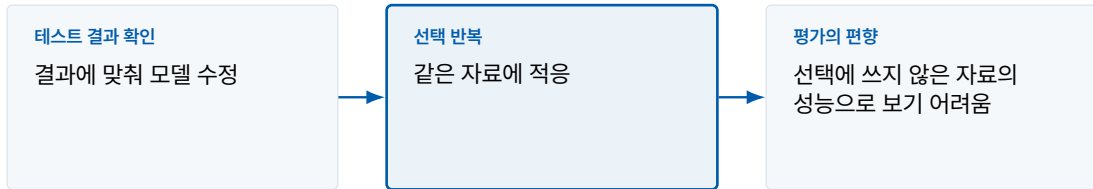
Training, Validation, Test

세 몫이 각각 하는 일이 다르다

학습·검증·테스트 자료의 역할

자료	사용 목적	사용 시점
학습	계수 등 모델의 매개변수 추정	모델을 학습할 때
검증	후보 모델과 설정 선택	선택 기준을 비교할 때
테스트	선택을 마친 모델의 성능 평가	선택에 쓰지 않고 마지막에

테스트 결과로 모델을 고르면 평가가 편향된다



Caution 테스트 자료는 모델 선택에 사용하지 않는다. 선택을 마친 뒤 최종 평가에 사용한다

자료 수와 평가의 안정성을 함께 고려한다

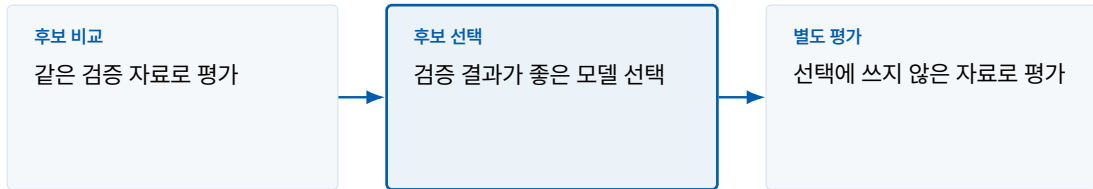
자료 상황	분할 시 고려점
관측이 충분함	학습 자료를 충분히 남기면서 검증에도 필요한 관측 수 확보
관측이 적음	교차검증으로 여러 관측을 번갈아 검증에 사용
반복 측정·시간 자료	관측 수뿐 아니라 사람·시간의 의존 관계 반영

모든 문제에 공통인 분할 비율은 없으며, 테스트 자료는 별도로 유지한다

학습·검증 성능으로 원인을 좁혀 간다

관찰	가능한 원인	추가로 확인할 것
두 성능 모두 낮음	과소적합, 최적화 문제, 잡음	학습 과정·특성·모델 복잡도
학습만 높음	과적합 또는 자료 분포 차이	학습·검증 구성과 복잡도
두 성능 모두 높음	후보 모델이 유용할 가능성	선택 후 테스트 성능

모델 선택도 검증 자료의 영향을 받는다



선택에 사용한 검증 점수는 최종 성능을 낙관적으로 추정할 수 있다

Checkpoint

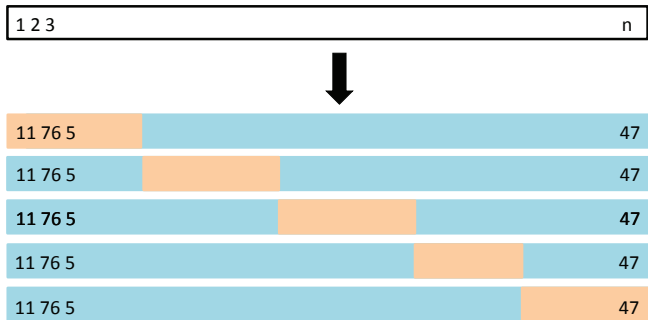
Training, Validation, Test

- 검증 자료와 테스트 자료는 각각 언제 쓰는가
- 테스트 점수를 보고 모델 설정을 바꾸면 무엇이 문제인가
- 학습만 좋고 검증이 나쁘면 어떤 원인을 확인해야 하는가

k-Fold Cross-Validation

모든 관측을 한 번씩 검증에 쓴다

k개의 fold로 나누는 교차검증



fold는 교차검증에서 나눈 한 몫이다. 다섯 fold 중 하나(베이지색)로 검증하고 나머지 넷(파란색)으로 학습한다

Source: James et al. (2023), Figure 5.5.

각 분할에서 학습·검증 자료를 구성한다

반복	학습에 쓰는 fold	점수를 재는 fold
1	2, 3, 4, 5	1
2	1, 3, 4, 5	2
5	1, 2, 3, 4	5

각 분할에서 모델을 새로 학습한다. 여기서는 5-fold 중 세 번의 학습·검증 구성을 표시했다

k 개 검증 오차를 평균한다

$$CV_{(k)} = \frac{1}{k} \sum_{i=1}^k MSE_i$$

관측 하나의 사용	k 번의 학습·검증 전체에서
학습	$k - 1$ 번 사용
검증	1번 사용

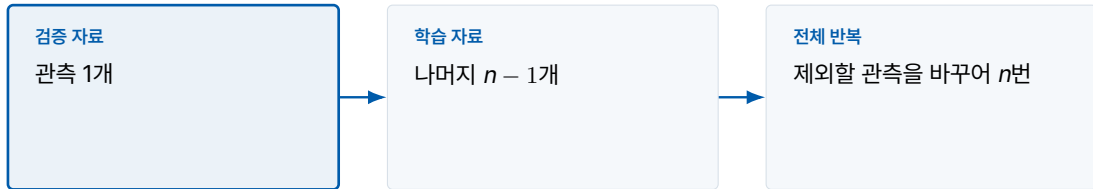
MSE_i 는 i 번째 fold의 검증 MSE다. fold별 점수를 같은 비중으로 평균한다. 관측 392개의 5-fold 크기는 79·79·78·78·78이며, 매번 학습·검증 자료는 겹치지 않는다

평균이 같아도 분할별 변동은 다르다

모델	다섯 fold의 정확도	평균
A	0.79, 0.80, 0.80, 0.80, 0.81	0.80
B	0.60, 0.70, 0.80, 0.90, 1.00	0.80

분할별 변동은 안정성을 살피는 정보이며, 그 자체가 평균의 신뢰구간은 아니다

관측 하나씩 제외하는 LOOCV



LOOCV(leave-one-out cross-validation)는 $k = n$ 인 교차검증이다

검증 MSE가 가장 낮은 다항 차수를 고른다

Auto 자료에서 horsepower로 mpg를 예측하는 1·2·5·10차 네 후보를 같은 5-fold로 비교한다

다항 차수	평균 검증 MSE	분할 간 표준편차	선택
1	24.13	2.26	선택
2	19.16	3.99	
5	18.98	3.66	
10	19.70	2.75	

같은 fold끼리 비교하면 5차가 다섯 번 중 네 번 낮지만 2차–5차는 -0.67 에서 0.64 까지 흔들린다. 평균 차이 0.18 이 작으니 2차도 함께 본다

같은 분할에서 후보를 비교한다. MSE는 낮을수록 좋고 정확도는 높을수록 좋다

Adapted from James et al. (2023), Chapter 5 lab, Auto.

Checkpoint

k-Fold Cross-Validation

- 한 관측은 검증에 몇 번 쓰이는가
- 평균과 함께 분할별 변동을 보는 까닭은 무엇인가
- k 를 n 으로 두면 무엇이 되는가

What k Buys and Costs

k 에 따른 추정 오차와 계산 비용

k 는 학습 자료의 양과 반복 횟수를 바꾼다

k	학습에 쓰는 관측 수	모델 학습 횟수
5	약 $0.8n$	5
10	약 $0.9n$	10
n	$n - 1$	n

매번 쓰는 학습 자료가 많아져도 검증 오차 추정치의 분산이 항상 작아지는 것은 아니다

LOOCV의 추정 분산이 작다고 단정할 수 없다

특징	오차 추정에 미치는 영향
학습 자료가 거의 같음	학습된 모델의 예측들이 서로 강하게 상관될 수 있음
상관된 결과의 평균	독립적인 결과를 평균할 때보다 변동이 덜 줄 수 있음

Caution LOOCV의 n 개 검증 오차는 거의 같은 자료로 학습한 모델에서 나와 함께 움직인다. 값이 서로 비슷하다는 뜻은 아니다

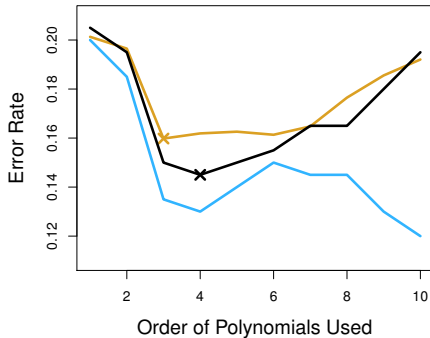
$k = 5$ 또는 10부터 검토할 수 있다

k	장점	고려할 점
5	학습 반복이 비교적 적음	학습 자료 약 80%, 테스트 오차를 조금 높게 추정할 수 있음
10	각 실행에서 학습 자료 약 90% 사용	5-fold보다 반복 횟수가 많음
n	거의 모든 관측으로 학습해 편향이 가장 작음	반복이 많고 추정 분산이 더 클 수 있음

계산 비용과 오차 추정의 편향(치우침)·분산(흔들림)을 함께 고려한다

Source: James et al. (2023), Section 5.1.4.

교차검증 오차와 테스트 오차를 비교한다



모의 분류 자료에 다항식 차수별 로지스틱 회귀를 적합했을 때의 오류율이다. 갈색은 테스트, 파란색은 학습, 검은색은 10-fold 교차검증이며, 교차검증은 4차, 테스트는 3차에서 가장 작다. 실제 자료에서는 테스트 오류율을 알 수 없어 교차검증 곡선으로 차수를 고른다

Source: James et al. (2023), Figure 5.8.

후보 선택과 최종 평가를 구분한다

목적	점수를 사용하는 방법
후보 선택	같은 교차검증 분할에서 후보들의 검증 성능을 비교한다
최종 평가	고른 후보를 선택에 쓰지 않은 테스트 자료로 평가한다

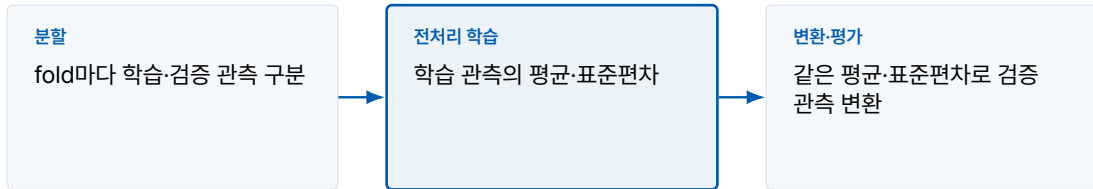
교차검증으로 후보를 고를 수 있지만, 고른 후보의 점수에는 선택의 영향이 남는다

fold를 나눌 때 조심할 것

자료의 성질	그냥 나누면	할 일
한 범주가 드물다	어떤 fold에는 없을 수 있다	비율을 지켜 나눈다
한 사람의 반복 관측	같은 사람이 양쪽에 든다	사람 단위로 나눈다
시간 순서가 있다	미래로 과거를 맞힌다	시간 순으로 나눈다

Caution 무작위로 나누는 것이 언제나 옳지는 않다

전처리의 통계량도 학습 fold에서 구한다



Caution 평균·표준편차를 전체 자료에서 미리 구하면 검증 정보가 학습 과정에 들어간다

코드에서는 fold 분할과 교차검증을 따로 정한다

KFold가 fold를 나누고, cross_validate가 fold마다 파이프라인 전체를 새로 학습한다

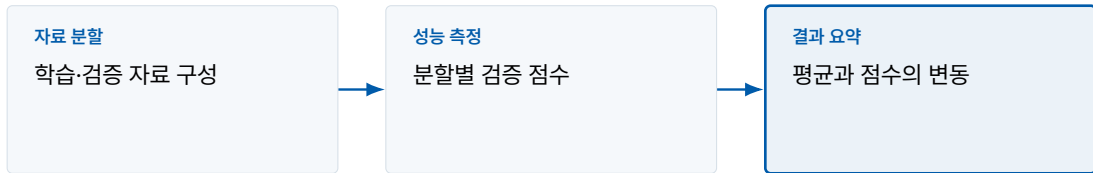
```
1 from sklearn.model_selection import KFold, cross_validate
2 from sklearn.pipeline import make_pipeline
3 from sklearn.preprocessing import StandardScaler
4 from sklearn.linear_model import LinearRegression
5
6 cv = KFold(n_splits=5, shuffle=True, random_state=0)
7 model = make_pipeline(StandardScaler(), LinearRegression())
8 scores = cross_validate(model, X, y, cv=cv,
9                          scoring="neg_mean_squared_error")
10 mse = -scores["test_score"]
```

test_score는 각 fold의 검증 점수다. 이 지표는 MSE에 음수 부호를 붙인 값이라 부호를 다시 바꾼다

용어	무엇을 가리키는가
검증 자료	후보를 고르는 데 쓰는 몫
k -fold 교차검증	자료를 k 개 fold로 나눠 k 번 재고 평균하는 방법
LOOCV	k 를 관측 수만큼 크게 잡은 경우

점수와 함께 변동을 살펴본다

검증 점수는 어떤 자료로 학습하고 평가했는지에 따라 달라진다



평균 점수가 비슷해도 분할별 점수의 변동은 다를 수 있다

지표에 따라 좋은 점수의 방향이 다르다

평가지표	좋은 방향	지표가 재는 값
MSE	낮을수록 좋음	관측값과 예측값 사이의 제곱 오차
정확도	높을수록 좋음	전체 관측 중 올바르게 분류한 비율

검증 결과를 비교하기 전에 지표와 점수의 방향을 정한다

교차검증의 학습 횟수는 후보 수와 k 의 곱이다

하는 일	후보 수	k	학습 횟수
한 번 나눠 재기	1	없음	1
교차검증	1	5	5
후보 열 개 비교	10	5	50

표는 후보 비교에 드는 학습 횟수다. 선택한 설정으로 개발 자료 전체를 다시 학습하는 1회는 별도로 센다

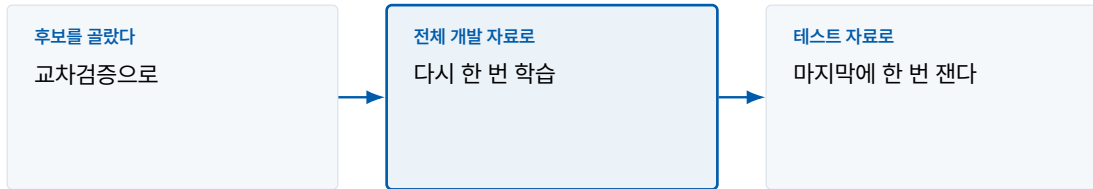
Caution 후보 수와 k 가 곱해지므로 모델 설정을 늘리면 시간이 빠르게 늘어난다

평균 차이가 작으면 분할별 결과도 비교한다

확인할 것	판단에 사용하는 방법
같은 fold의 점수 차이	일관되게 나은 후보가 있는지 확인
모델의 복잡도·비용	성능 차이가 작으면 단순한 후보도 고려
분할 간 표준편차	변동은 보여 주지만, 두 후보의 차이가 우연인지 판단하는 근거는 아님

평균 차이와 표준편차만 비교해 어느 후보가 더 낫다고 단정하지 않는다

고른 뒤에 하는 일



선택한 설정으로 개발 자료 전체에서 다시 학습한다. 테스트 자료는 여기에 포함하지 않는다

Checkpoint

What k Buys and Costs

- LOOCV로 학습한 모델들의 예측이 서로 상관될 수 있는 까닭은 무엇인가
- 후보 선택에 사용한 교차검증 점수에는 어떤 영향이 남는가
- 한 사람을 반복 측정 한 자료는 어떻게 나누는가

Which Number Tells You Which Mistake?

Next Lecture: Evaluation Metrics and Project Launch

- 회귀와 분류의 지표
- 혼동행렬에서 나오는 두 비율
- 임계값을 옮기며 그리는 곡선