모두의 딥러닝 (Deep learning)

Nambeom Kim (nbumkim@gmail.com)

Overfitting problem

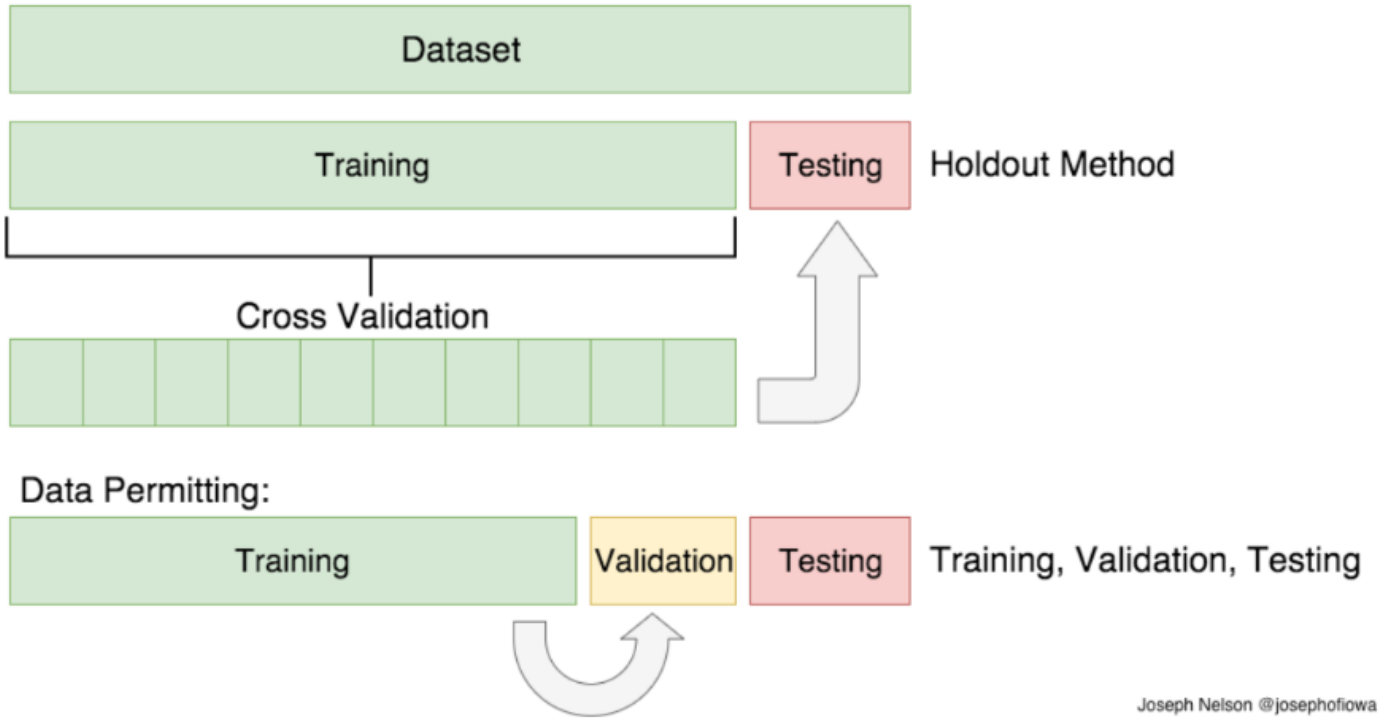
Training set and test set

- ML 모델의 성능평가를 위해서 자료를 분할

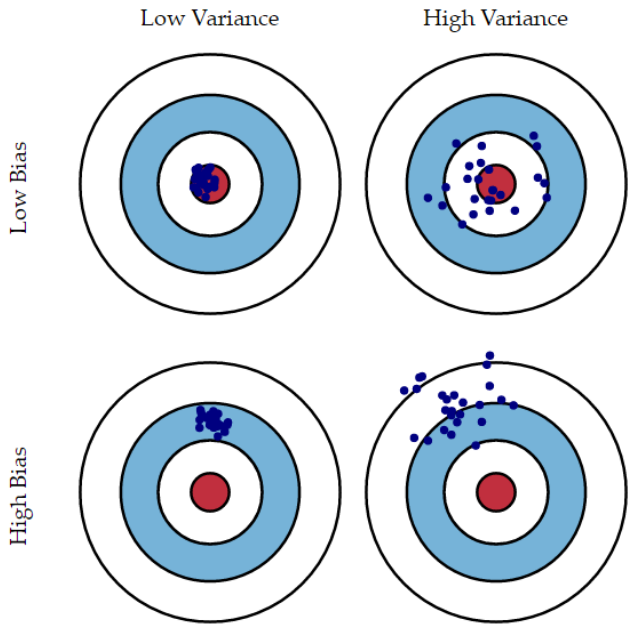
- **Training set**: 모델의 알고리즘 learning, 모델에 사용될 feature들을 결정, 초매개변수 조절 (약 전체 자료수의 70% 로 설정)
  - **Training set**: 모델의 알고리즘 learning
  - **Validation set**: 모델에 사용될 feature들을 결정, 초매개변수 조절, 과적합 (Over-fitting) 방지
- **Test set**: 최종 선택된 모델의 성능평가 (약 전체 자료수의 30% 로 설정), 자료의 수가 적을 경우 생략 가능

일반적인 ML 예측 과정

- 학습세트 (Training set): 머신러닝 모델을 학습할 때 사용
- 검증세트 (Validation set): 하이퍼 파라메터 결정할 때
- 테스트세트 (Test set): 학습된 모델을 평가할 때

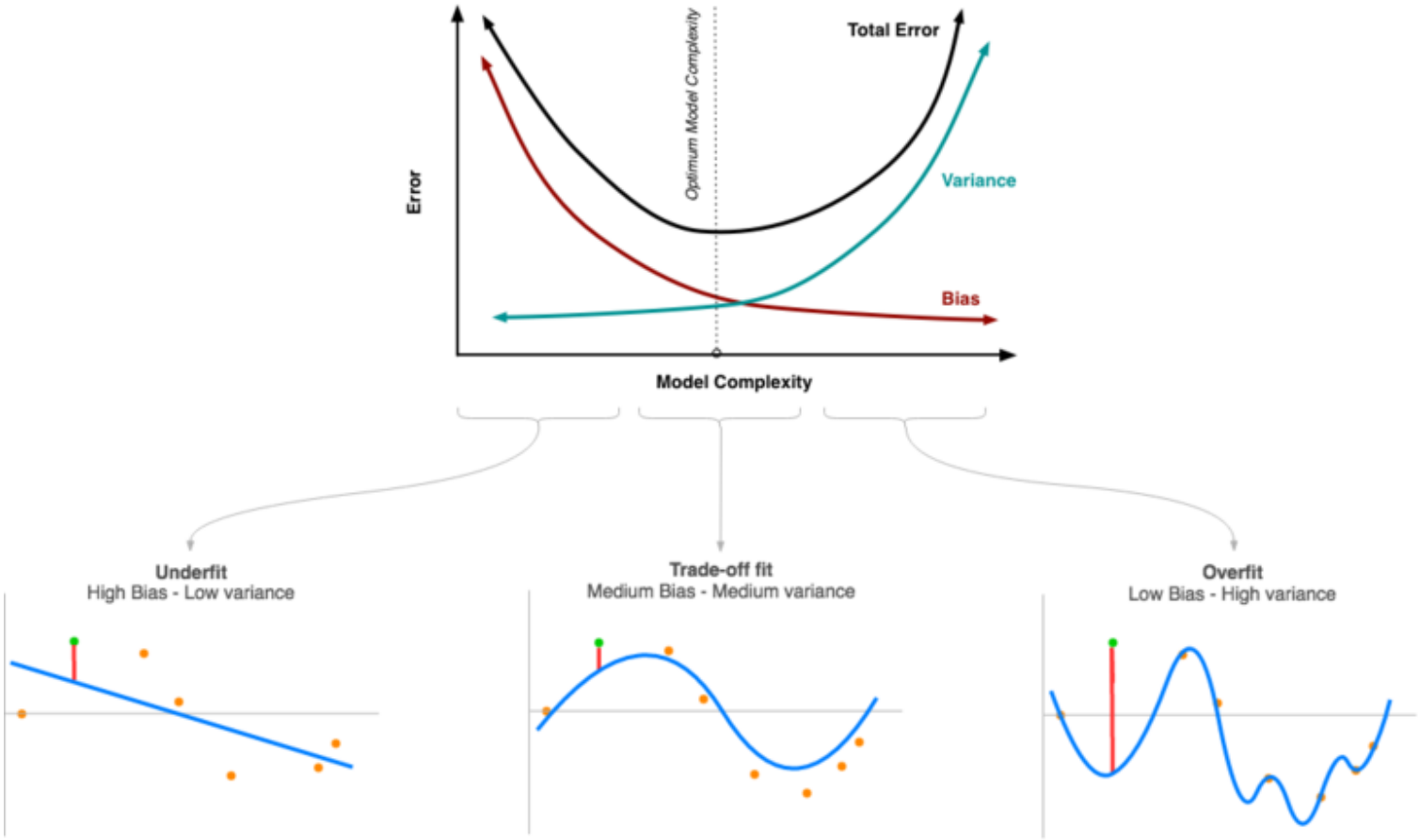


ML 모델의 치우침 (Bias) 과 분산(variance)



https://miro.medium.com

ML 모델의 치우침 (Bias) 과 분산(variance)



https://blog.naver.com/PostView.nhn?blogId=ckdgus1433&logNo=221594203319

치우침(Bias) - 분산(variance) trade-off

치우침(Bias)

- 치우침은 모델의 실제값 (또는 평균) 과 예측치 간의 차이를 의미.
- 과소적합 (underfitting)은 치우침이 높은 모델이 발생하기 쉬움.

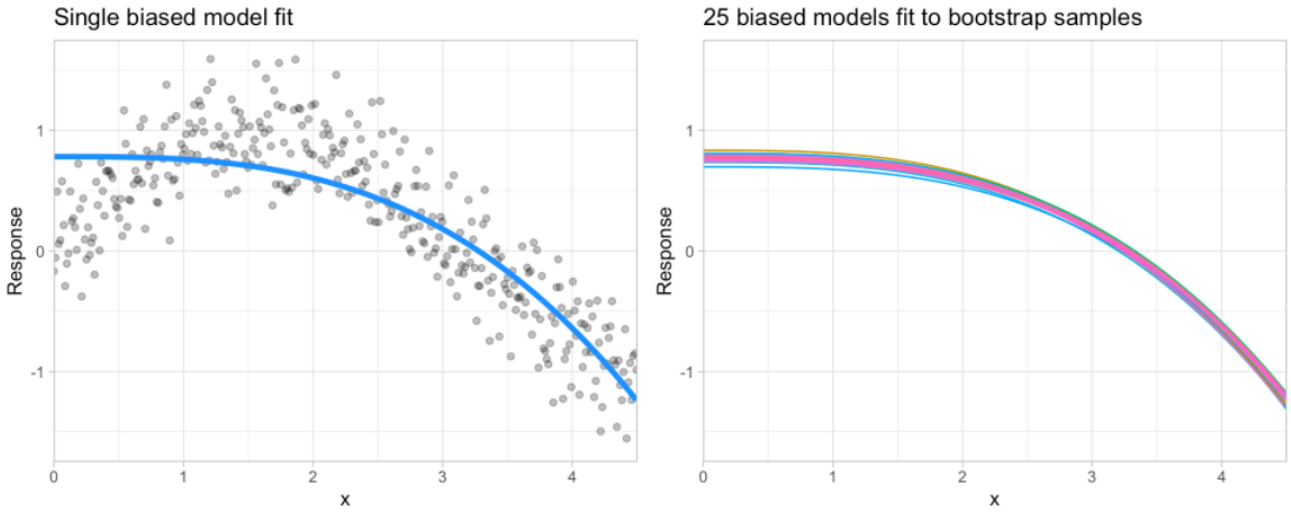


Figure 2.8: A biased polynomial model fit to a single data set does not capture the underlying non-linear, non-monotonic data structure (left). Models fit to 25 bootstrapped replicates of the data are underterred by the noise and generates similar, yet still biased, predictions (right).

https://bradleyboehmke.github.io/HOML

치우침(Bias) - 분산(variance) trade-off

분산 (Variance)

- 분산 (Variance)은 주어진 데이터에서 모델 예측은 변이로 정의.
- 과적합 (overfitting)은 분산이 높은 모델이 발생하기 쉬움.

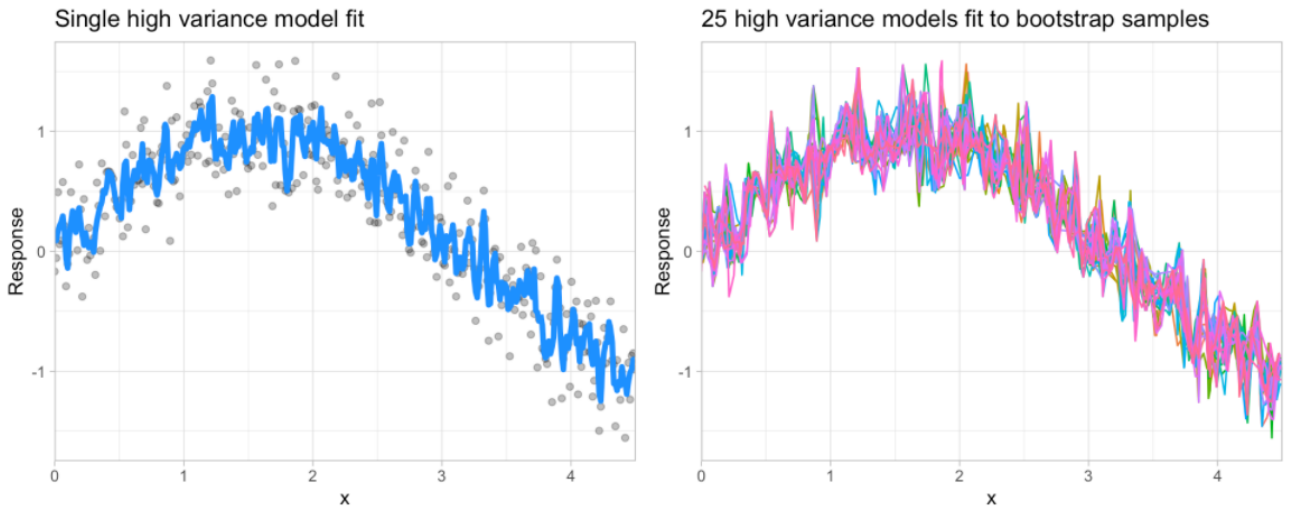


Figure 2.9: A high variance *k*-nearest neighbor model fit to a single data set captures the underlying non-linear, non-monotonic data structure well but also overfits to individual data points (left). Models fit to 25 bootstrapped replicates of the data are deterred by the noise and generate highly variable predictions (right).

https://bradleyboehmke.github.io/HOML

기계학습 모델 평가

k-fold 교차검증 (k-fold cross validation(CV))

- k-fold 교차 검증 (일명 k-fold CV)은 훈련 데이터를 동일한 크기의 k 그룹 (k-fold)으로 무작위로 나누는 리샘플링 방법
- k-fold CV 추정치는 k 테스트 오류를 평균화하여 계산.

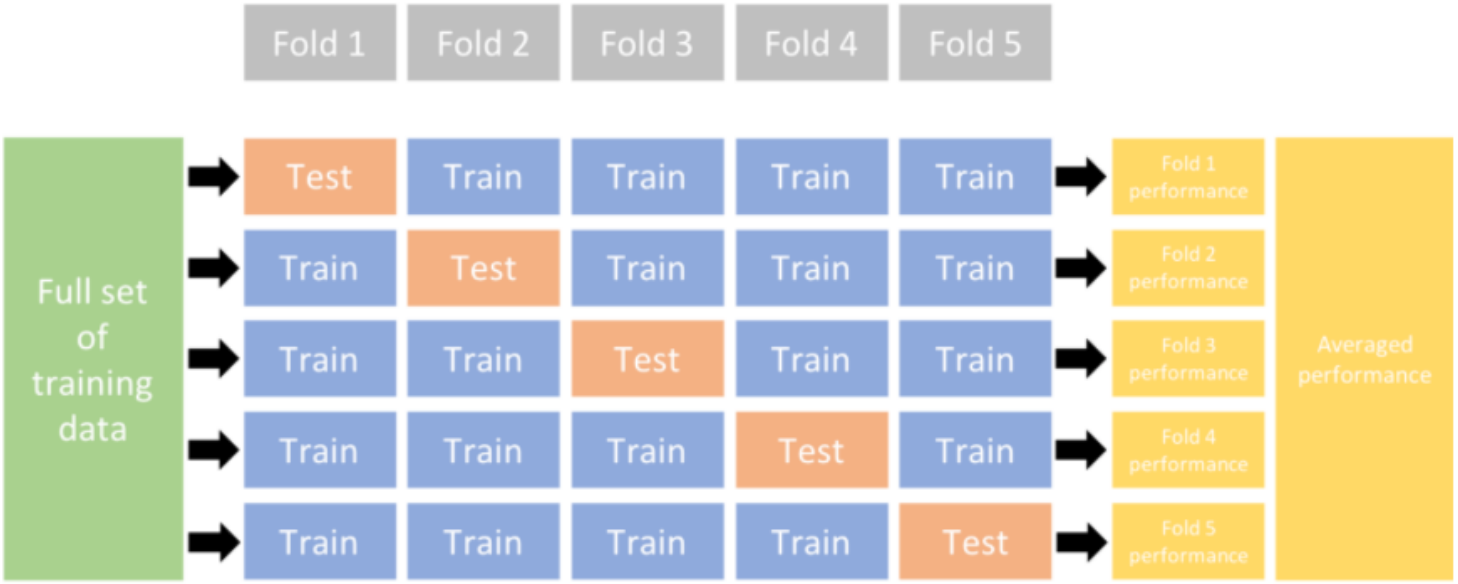


Figure 2.4: Illustration of the k-fold cross validation process.

https://bradleyboehmke.github.io/HOML

기계학습 모델 평가

Bootstrapping

- Bootstrapping 샘플은 복원추출을 이용한 데이터의 무작위 샘플.
- Bootstrapping은 선택한 샘플을 기반으로 모델을 구축하고 OOB (Out-of-Bag) 샘플을 이용하여 모델을 평가

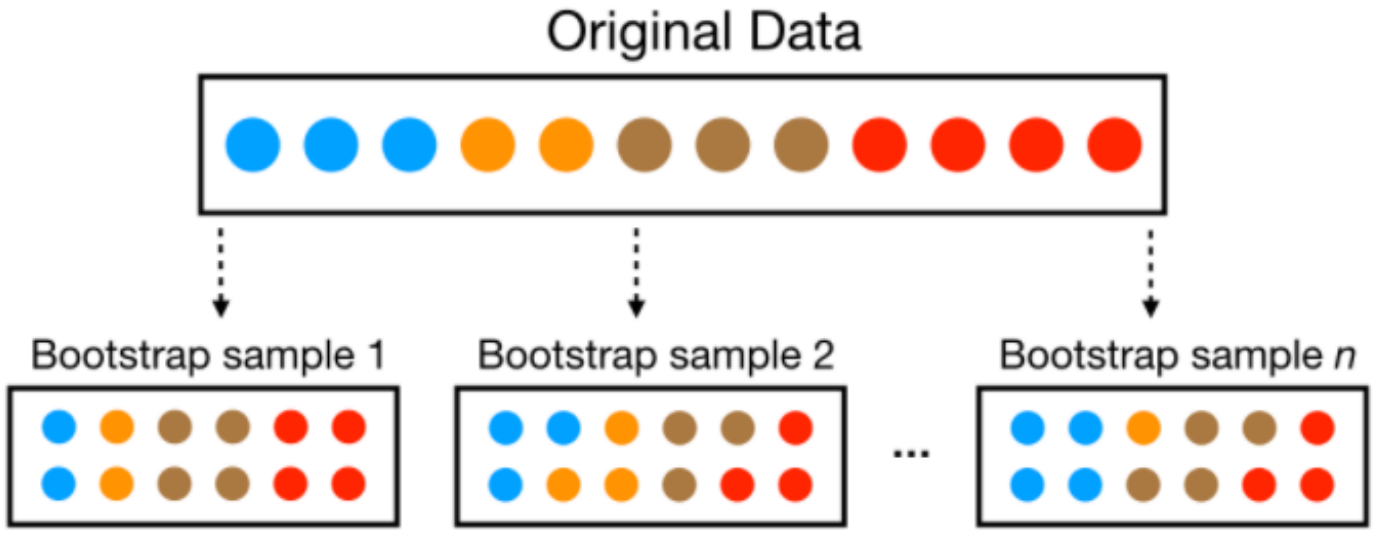


Figure 2.6: Illustration of the bootstrapping process.

https://bradleyboehmke.github.io/HOML

bootstrapping *vs* 10-fold CV (n = 32) 비교

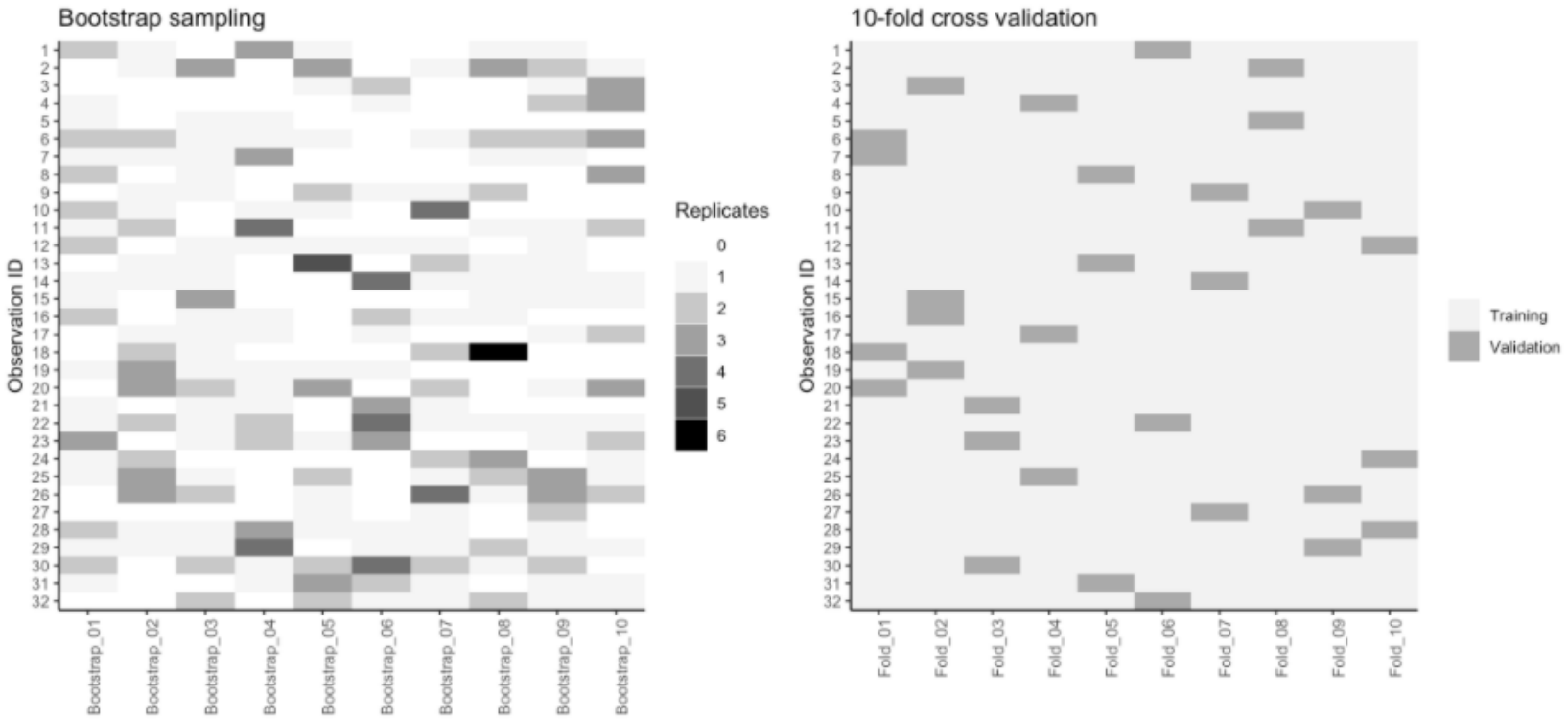


Figure 2.7: Bootstrap sampling (left) versus 10-fold cross validation (right) on 32 observations. For bootstrap sampling, the observations that have zero replications (white) are the out-of-bag observations used for validation.

https://bradleyboehmke.github.io/HOML