Cross Validation

DS 301

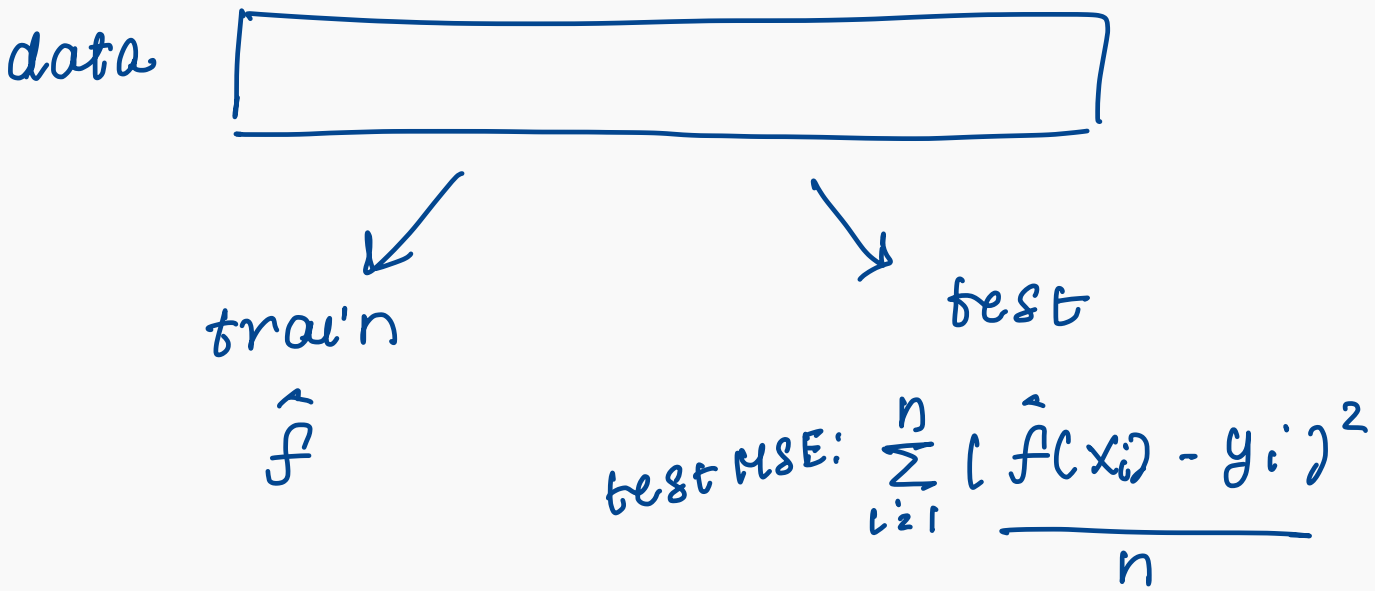
Iowa State University

Cross-validation

An alternative to the approaches we discussed is to directly estimate the test error using cross-validation.

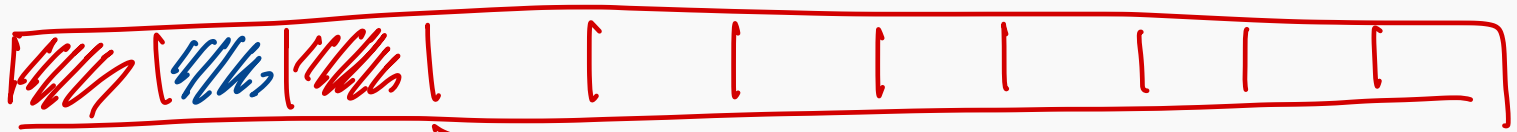
CMSE)

(1) validation set approach



Leave-one-out cross-validation (LOOCV)

- leave out a single observation
  ↳ test set.
- $(n-1)$ observations remaining ⟶ training set.



data
$n = 10$

test MSE_1 test MSE_2 test MSE_3 ..... test MSE_{10}.

⟹ average of these:

$$CV_{(n)} = \frac{\sum_{i=1}^{n} \text{test } MSE_i}{n}.$$

- utilize as much data as possible
- every observation gets the chance to be the test set.
- deterministic.

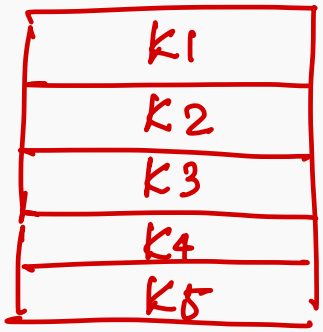
Drawbacks of LOOCV

- computationally expensive
  - ↳ need to fit model n times.

- theoretical problems w/ LOOCV.

k-fold cross-validation

data



$K = 5$

each fold gets the chance to be the test set.

- K2, K3, K4, K5 → training set
  K1 → test set.

  test MSE_{K1}

- K1, K3, K4, K5 → training set
  K2 → test set.

  test MSE_{K2}

  ⋮

  test MSE_{K5}

$$CV_{K=5} = \frac{1}{5} \sum_{i=1}^{K} \text{test } MSE_{Ki}$$

K = 5, K = 10, . . .

See R script: `cross_validation.R` and `cv_subset.R`