PS5841

Data Science in Finance & Insurance

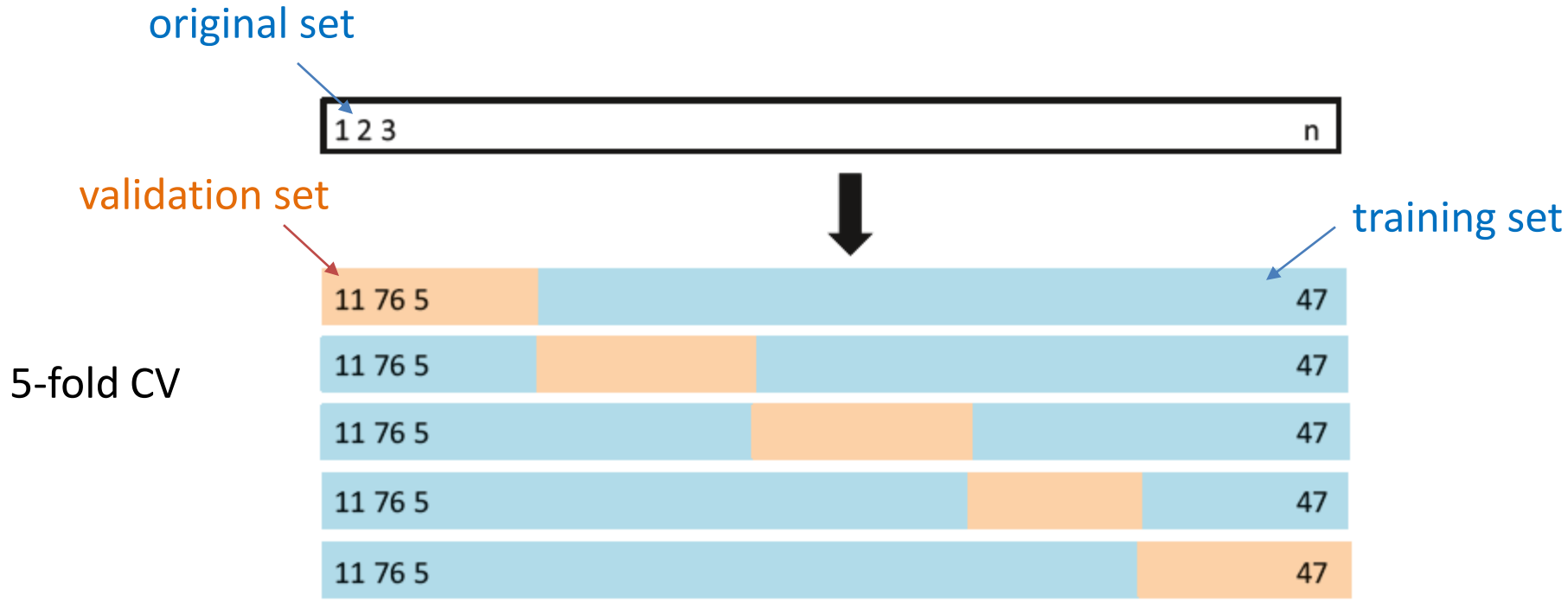
Cross Validation

Yubo Wang

Autumn 2021

Columbia University

k-Fold Cross-Validation



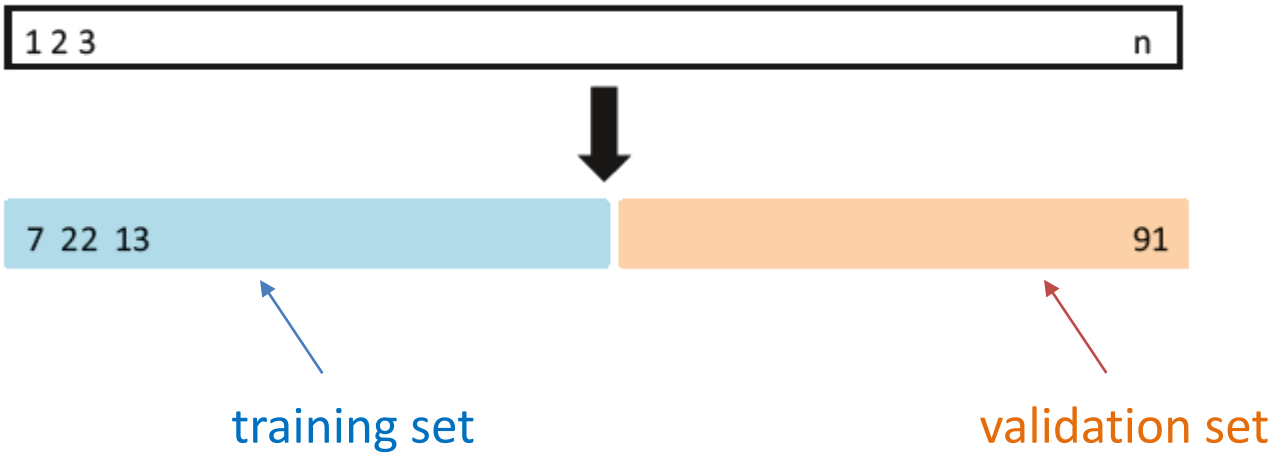
k-fold CV: For each validation set

- Fit model on $k-1$ folds (training set)
- Compute "$Error_i$" on the hold-out fold (validation set)
- Compute the CV estimate of the "test error"

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

COLUMBIA UNIVERSITY

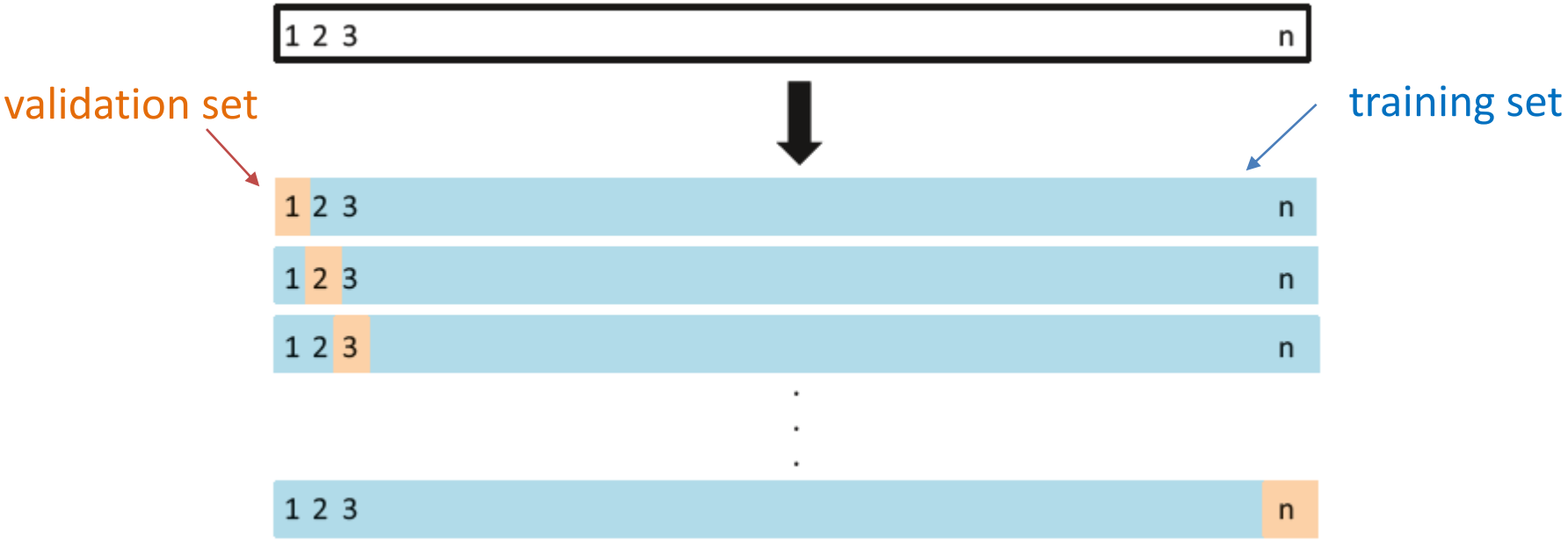
The Validation Set Approach



- Fit model on the training set
- Compute MSE on the validation set

COLUMBIA UNIVERSITY

Leave-One-Out Cross Validation



Same as n-fold CV

COLUMBIA UNIVERSITY

Magic Numbers

- There is a bias-variance trade-off associated with the choice of k in k-fold CV.
- Empirically, k=5, or k=10 are the best
  - Neither excessively high bias nor high variance

COLUMBIA UNIVERSITY

That was

COLUMBIA UNIVERSITY