Resampling Methods

Jesús Calderón

Learning Objectives

By the end of this week, students will be able to:

- Explain two resampling methods: cross-validation and the bootstrap.
- Explain how to assess model performance systematically.
- Explain the model selection process.
- Implement models and model workflows using the tidymodels library.

Resampling Methods

Resampling Methods

- Resampling methods are indispensable in modern statistics.
- Involve repeated sampling from the training set and refitting a model on each sample to obtain additional information about the fit.
- Can be computationally expensive, but modern computing enables these methods.

Methods

- **Cross-Validation**: used to estimate the test error associated with a given statistical learning method. Two applications:
  - *Model assessment*: the process of evaluating a model's performance.
  - *Model selection*: the process of selecting the right level of flexibility for a model.
- **Bootstrap**: many applications, but commonly used to measure the accuracy of a parameter estimate or of a given ML method.

Validation Set and Leave-One-Out Cross-Validation

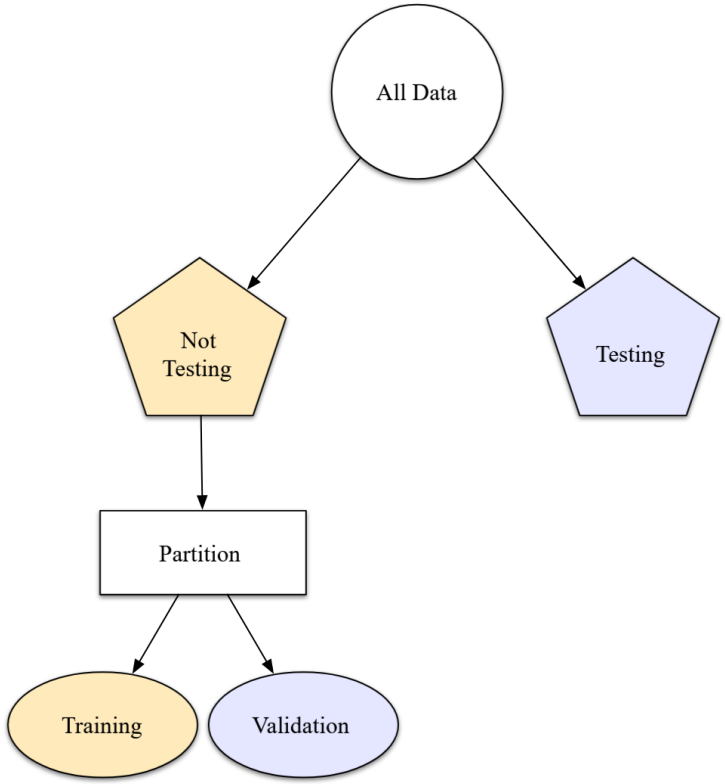
Introduction

- Test error can be easily calculated if plenty of test cases are available. However, we don’t always have access to an abundance of test data.
- In cases where we only have limited data, we will need to reserve some of it for testing. With the remaining data, we can apply four strategies:
  - Validation set.
  - Leave-One-Out Cross-Validation (LOOCV)
  - K-Fold Cross-Validation (CV)

Validation Set

Randomly divide the available set of observations into:

- Training Set
- Validation or Hold-Out Set



A predictive model case study (tidymodels.org)

Validation Set: Drawbacks

There are two potential drawbacks to the Validation Set approach:

- The test error can be highly variable, depending on the observations that are included in the training set and which observations in the validation set.
- Only a subset of the observations are used to fit the model. Since learning methods tend to perform worse when trained on fewer observations, this may imply that the validation set error rate may tend to *overestimate* the test error rate for the model fit on the entire data set.

Leave-One-Out Cross-Validation (LOOCV)

- Attempts to address the drawbacks of the Validation Set approach.
- Uses each observation as the validation set and trains on the remaining n-1 cases, calculate validation performance, and repeat n-1 times.

- For each observation *i*:
  - Use *i* as a validation set.
  - Train model on remaining *n-1* observations.
  - Calculate validation performance and store.
- Average performance over all observations.

LOOCV: Pros and Cons

Advantages

- *Less bias*: repeatedly fit learning method on n-1 samples vs a single validation set of X%. As a consequence, LOOCV tends not to overestimate the test error as much as the validation set approach.
- *Produces More Stable Results*: performance measures tend to vary less than in the validation set approach.

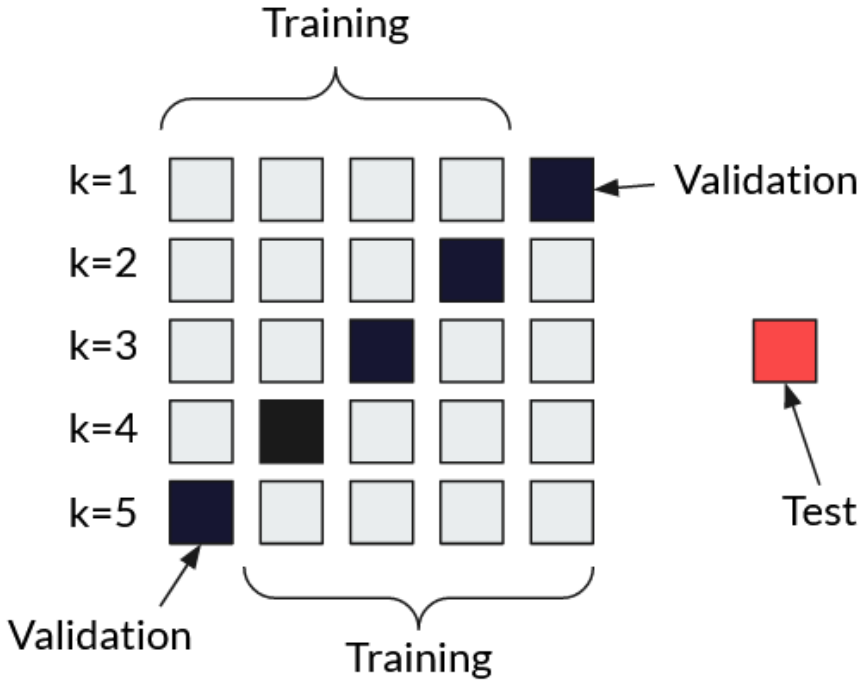
Disadvantage

- *Expensive to implement*: the model is fit n times. Can be costly and time-consuming if n is large or when model is slow to fit.

K-Fold Cross-Validation

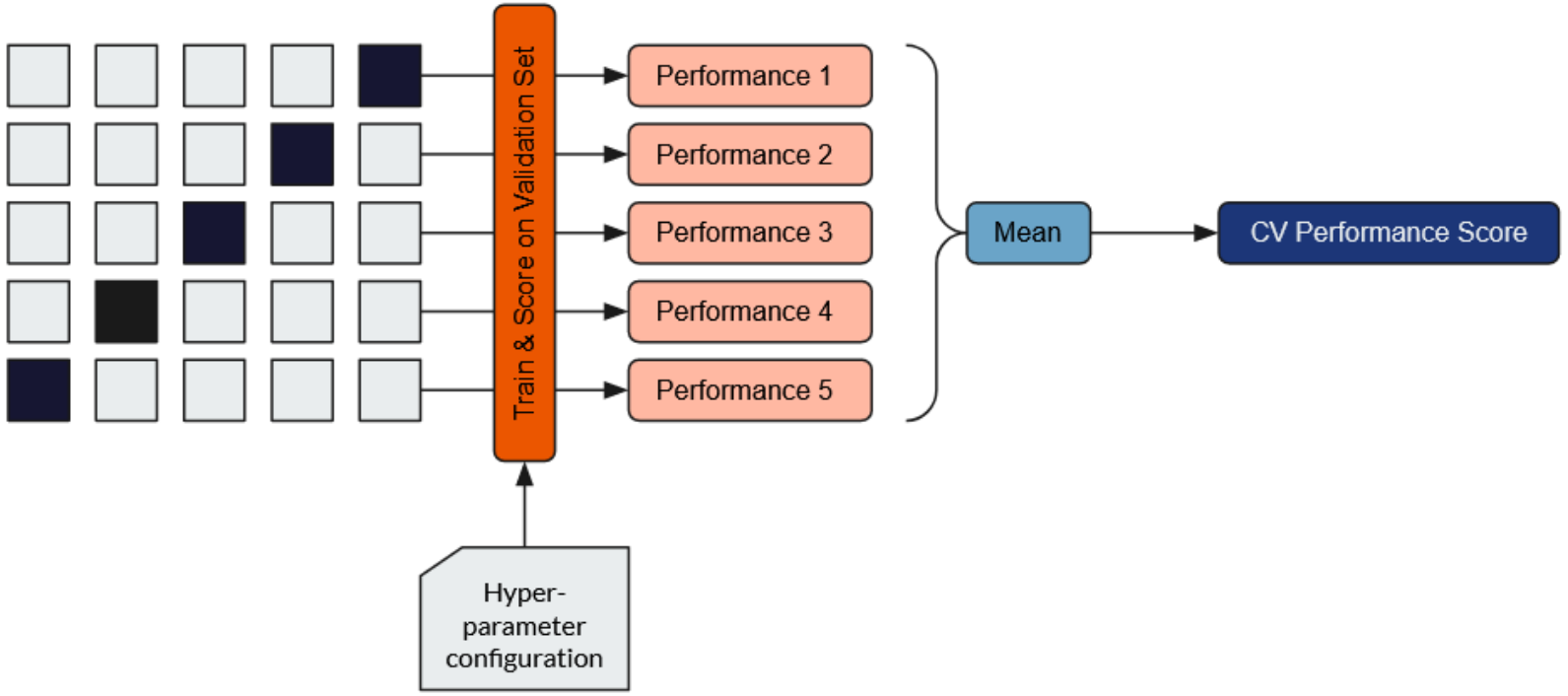
K-Fold Cross-Validation

- Choose K: in practice between 5 and 10.
- Randomly partition data in K subsamples, also called *folds*.
- For each fold *j*: use *j* as the validation set, train on K-1 remaining folds, calculate performance measures and store.
- Average performance over all repetitions.



K-Fold CV initial Set Up

Model Assessment using K-Fold Cross-Validation

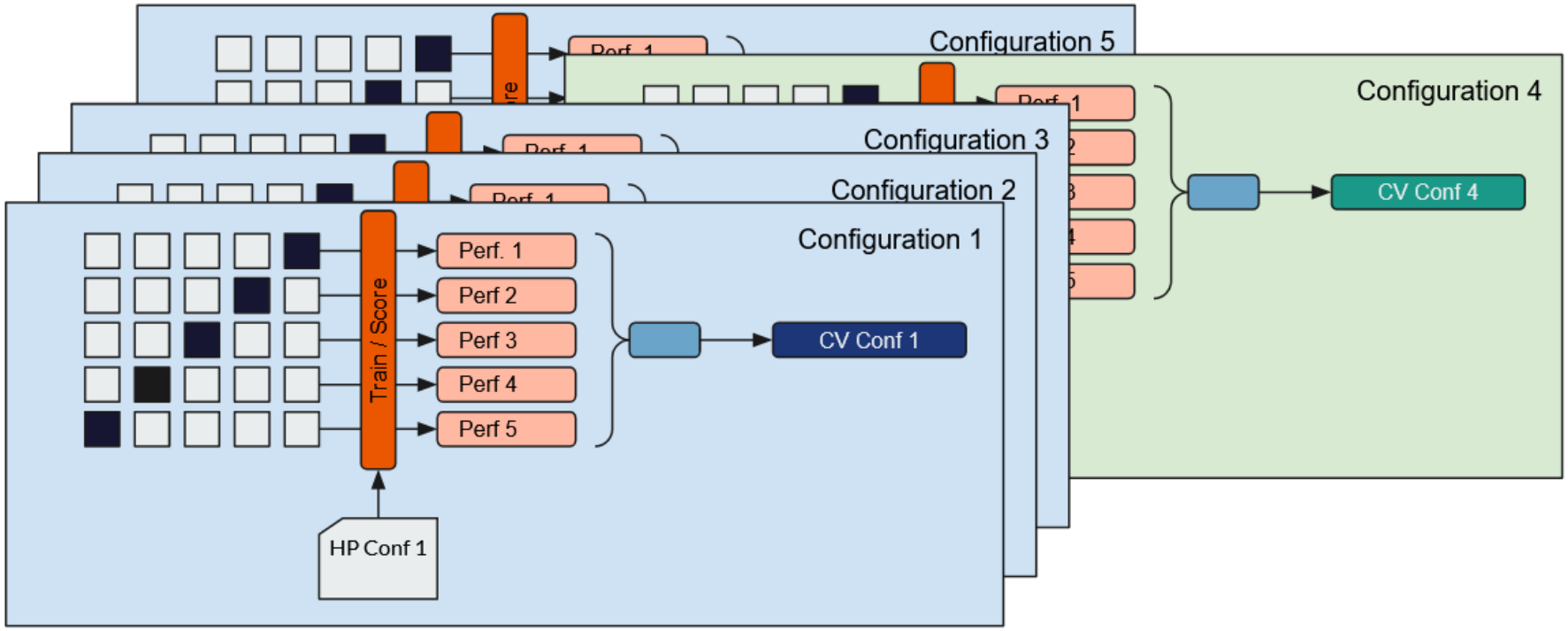


Model Assessment

K-Fold Cross-Validation: Pros and Cons

- Lower computational cost than LOOCV.
- Less bias than the Validation Set approach. K-Fold CV uses a larger number of samples, and this reduces bias.
- Higher bias than LOOCV since K-Fold CV does not use as many samples.
- Lower variance than LOOCV. When we apply LOOCV we are training models on almost identical training sets, and this induces variance.
- Notice that the choice of K is a trade-off between variance and bias of test error rate estimates.
- As well, we can use K-Fold CV to assess the performance of different models, hyperparameter configurations, and features over the *training* set (remember, we no longer have the “validation” set).

Model Selection



Model Selection

The Bootstrap

The Bootstrap

- Statistical tool used to quantify the uncertainty associated with a given estimator or statistical learning method.
- Learning methods do not need to be analytical: some methods do not naturally produce an estimator standard error (as in the case of linear/logistic regression).
- The method can be thought to estimate the assumption that “we have infinitely many samples.”
- Sample with replacement several times from training set, estimate model parameters, measure their variability (standard error).
- Sample with replacement: the same observation can appear more than once.