Model Selection and Bias-Variance tradeoff

Prof. Alessandro Lucantonio

Aarhus University - Department of Mechanical and Production Engineering

?/?/2023

Motivations - Training vs test error

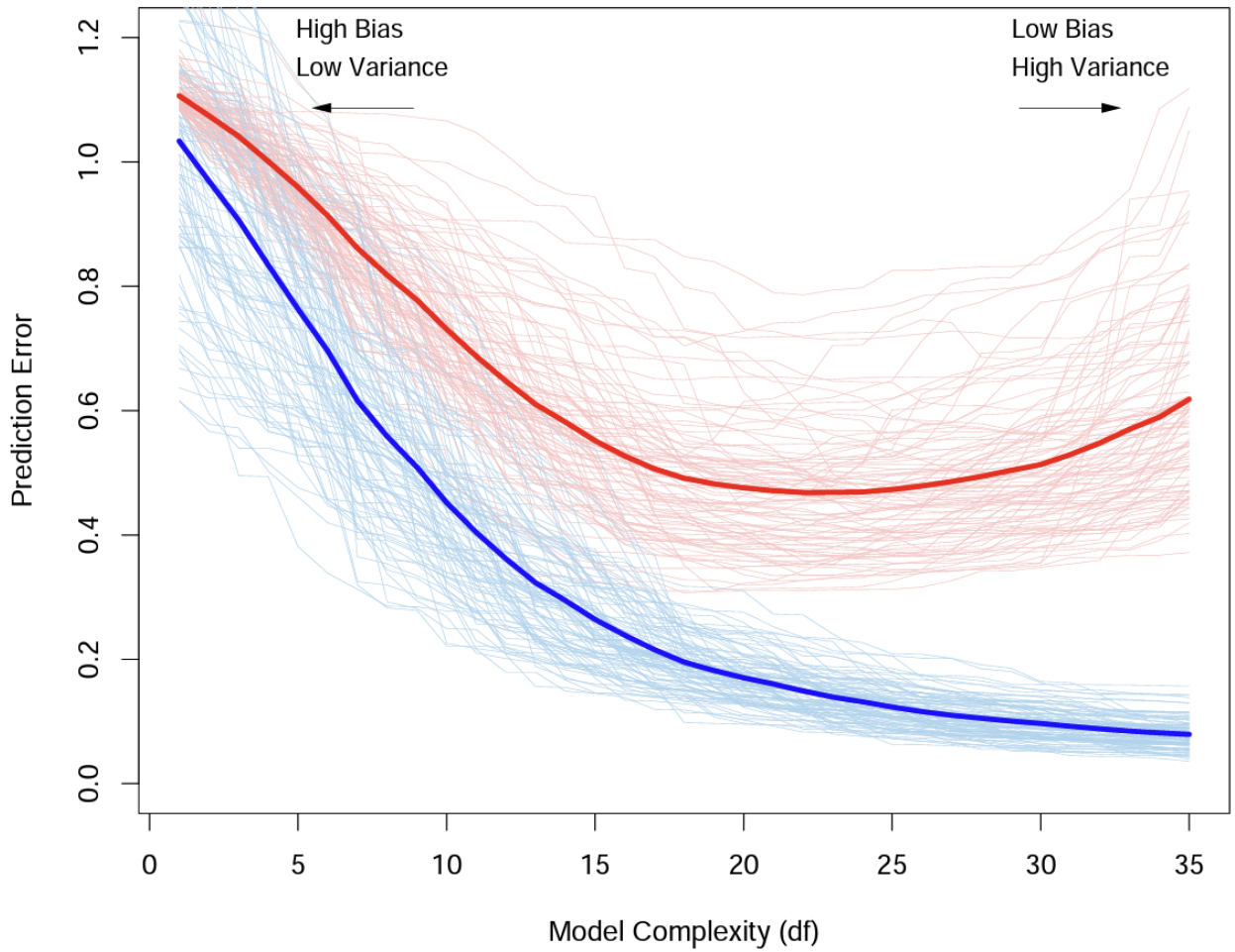


Figure: Training (blue) vs text (red) error as the model complexity varies.

Motivations - general idea

ML in one word: **generalization**!

Recall that we have to find a balance between fitting, on training data, and model complexity. Even though we limit the complexity of our model, training set does not provide a good estimate of test error.

In other words, generalization is compromised if we choose hyperparameters according to (only) training error.

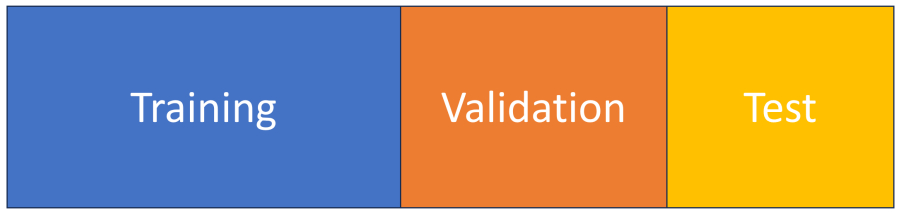
Model selection and model assessment

Model selection: estimate the performance of different models trained with different hyperparameters.

Model assessment: after choosing a final model we evaluate its perfomance on *completely new* test data.

N.B. Model selection and model assessment must be kept separated. Once we have chosen a final model, we are done with the model selection phase and model assessment is needed only to test the model on new data.

Double Hold out (Selection+Assessment)



- We split the entire dataset in three parts: training set, validation set and test set. Usually, training+validation constitutes the majority of the data available.
- Training set is used to fit. When the model is fitted, we evaluate its performance computing the error on the validation set. **Validation set is not a good estimation for the test error**.
- When a model is chosen, we evaluate its generalization capability computing the error on the test set. **Test set must not be used for hyperparameters selection**.

Cross validation (Selection or Assessment)

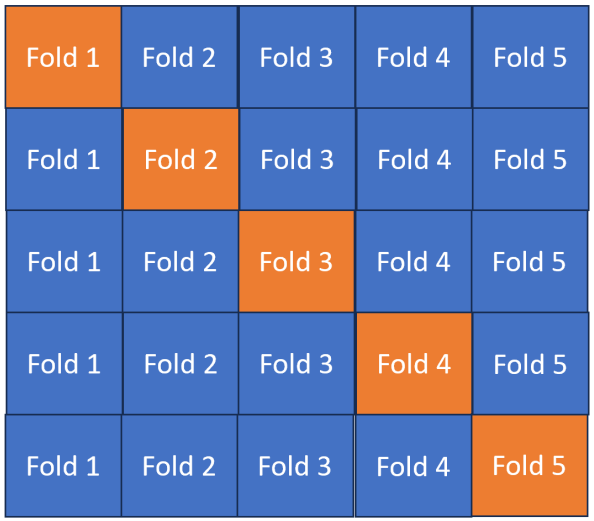


Figure: 5-fold CV

- Split the data in k disjoint folds.
- Use $k-1$ folds as the training set and the other fold as the validation set. Repeat it k time (look figure).
- The performance will be the mean $\pm$ standard deviation computed across the k runs.

Pro: Not sensible to a particular partition of the data. Mean filter the error.
Contra: Computationally expensive (but parallellizable).