Model Selection and Bias-Variance tradeoff

Prof. Alessandro Lucantonio

Aarhus University - Department of Mechanical and Production Engineering

?/?/2023

Motivations - Training vs test error

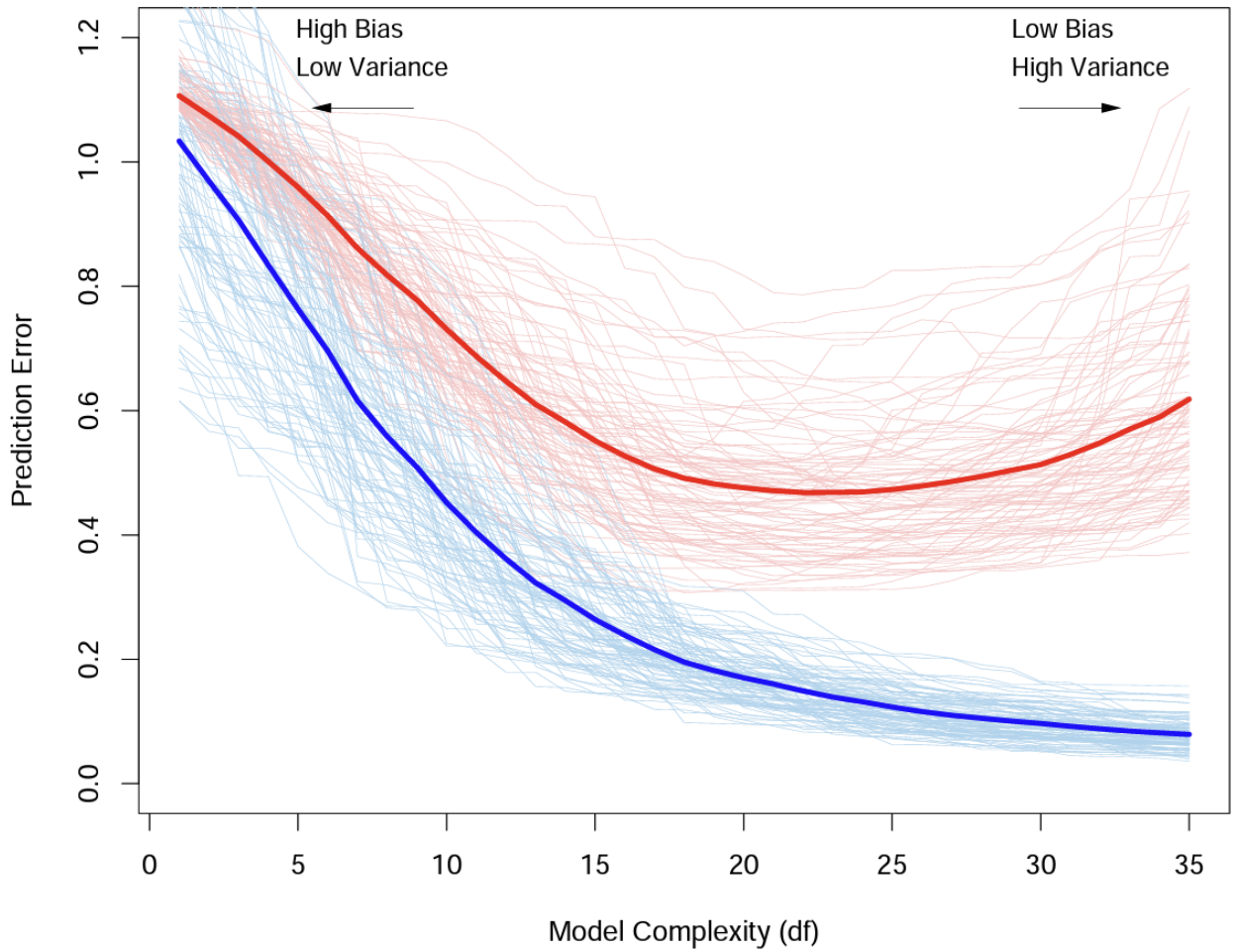


Figure: Training (blue) vs text (red) error as the model complexity varies.

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.

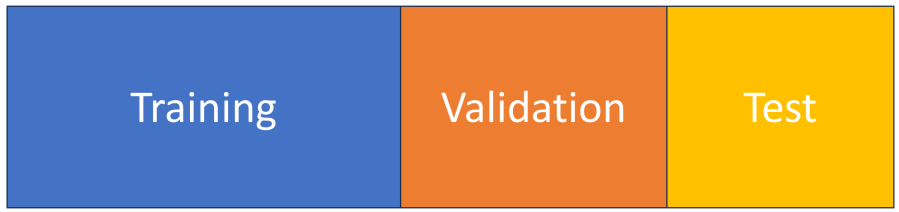
Model selection and model assessment

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)

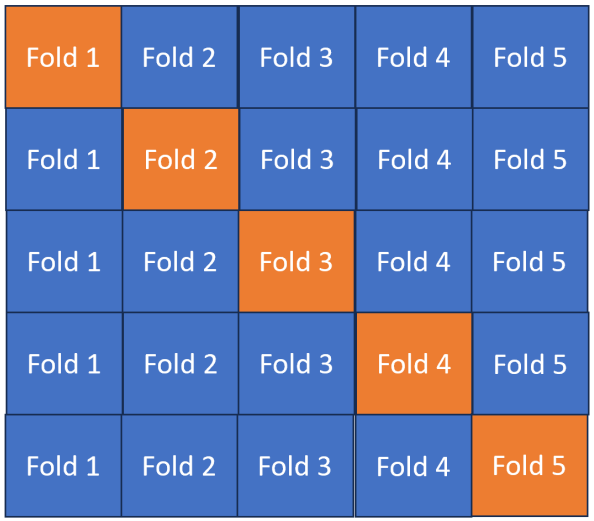


Figure: 5-fold CV

- 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).

CV for selection and hold out for assessment

1. Split dataset in two parts (hold out). One constitutes the development set D (model selection), the latter is the test set T (model assessment).
2. Use the cross validation method on D to select the best model.
3. Evaluate its performance on the **external** test set T.

Optional: after completing 2., hence after choosing the best hyperparameters and the best models, you can retrain (i.e. refit) the best model in the entire dataset D. In this way you exploit fully data you can for training.

Grid search (Selection)

How to search the best tuple of hyperparameters?

Example: two hyperparameters to choose (learning rate α and regularization parameter λ). Consider a set of values for both, e.g. $\alpha_{\text{vals}} = \{0.001, 0.01, 0.1\}, \lambda_{\text{vals}} = \{0.0001, 0.001, 0.01\}$.
Evaluate the model with $(\alpha, \lambda) = (i, j)$ for $i \in \alpha_{\text{vals}}, j \in \lambda_{\text{vals}}$.

α \ λ			
	(0.001,0.0001)	(0.001, 0.001)	(0.001, 0.01)
α	(0.01, 0.0001)	(0.01, 0.001)	(0.01, 0.01)
	(0.1, 0.0001)	(0.1, 0.001)	(0.1, 0.01)

Refinement: Suppose that the best is $(0.1, 0.001)$. Then we can "zoom" near the best couple, doing another grid search with e.g. $\alpha_{\text{vals}} = \{0.075, 0.1, 0.125\}, \lambda_{\text{vals}} = \{0.00075, 0.001, 0.00125\}$.

Bias and Variance definitions

Bias: discrepancy between true targets and the prediction of the current hypothesis.

Variance: variability of the predictions of the current hypothesis for different training data.

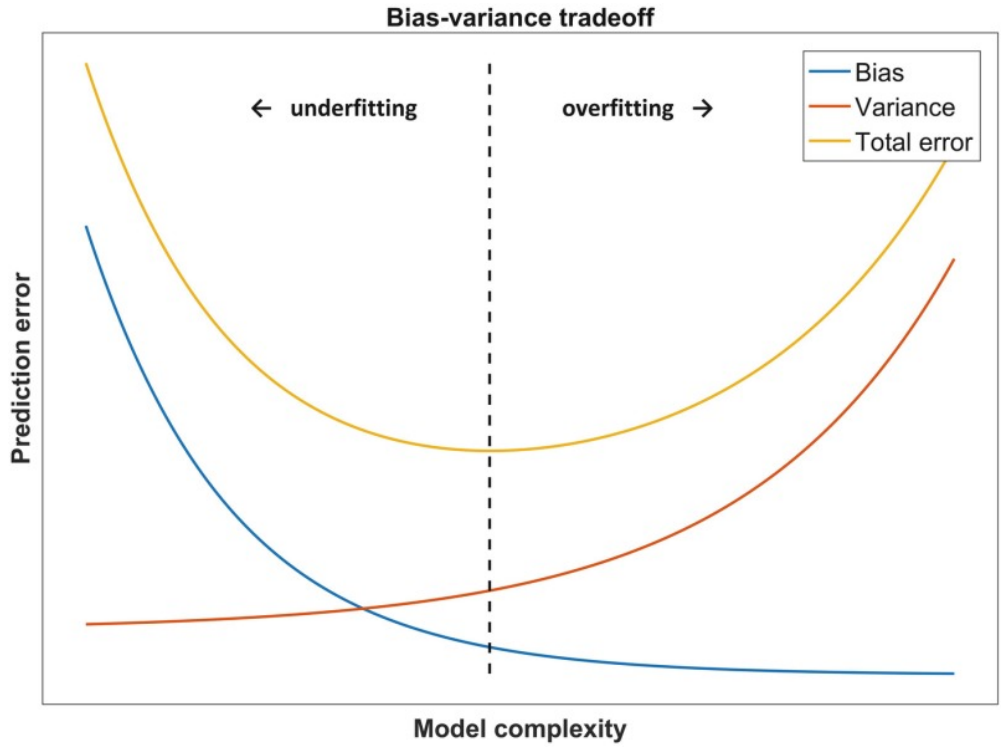


Figure: Yellow curve: validation error.

Bias-Variance: some intutitions

λ: regularization parameter.

- High λ: high bias (underfitting).
- Low λ: high variance (overfitting)
- Intermediate λ: optimal solution.

Complex model $\rightsquigarrow$ noise of the data captured $\rightsquigarrow$ high variance, low bias.

Simple model $\rightsquigarrow$ rigid hypothesis $\rightsquigarrow$ low variance, high bias.