SOMAIYA VIDYAVIHAR UNIVERSITY
K J Somaiya College of Engineering

Introduction To Need Of Estimation And Validation For Added Value Due To Data Science

Vaibhav P. Vasani

Assistant Professor

Department of Computer Engineering

K. J. Somaiya College of Engineering

Somaiya Vidyavihar University

Introduction To Need Of Estimation And Validation For Added Value Due To Data Science

Model selection at different scales

- Hyperparameters are the parameters of the learning method itself which have to specify a priori, i.e., before model fitting.
- In contrast, ***model parameters*** are parameters which arise as a result of the fit.

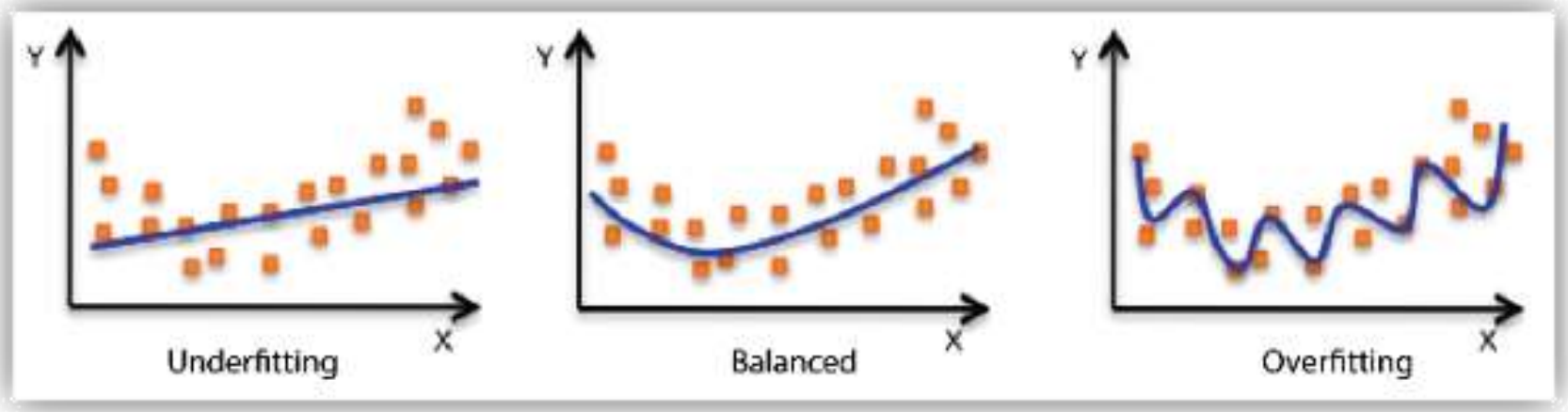
- Example
  - In a logistic regression model, for example, the regularization strength (as well as the regularization type, if any) is a hyperparameter which has to be specified prior to the fitting, while the coefficients of the fitted model are model parameters. Finding the right hyperparameters for a model can be crucial for the model performance on given data.

- the best ***learning method*** (and their corresponding "optimal" hyperparameters) from a set of eligible machine learning methods.
  - ***algorithm selection***.

Model evaluation

- Model evaluation aims at estimating the generalization error of the selected model, i.e., how well the selected model performs on unseen data.

Fitting of curve slide

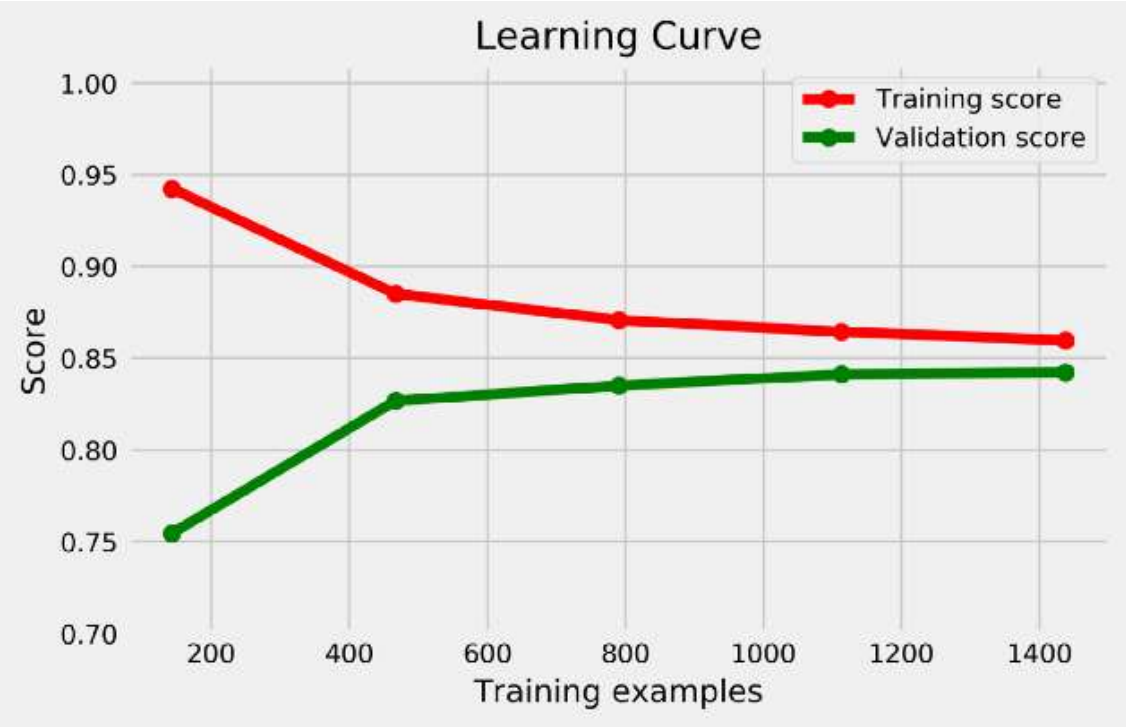


If data is not an issue

- The recommended strategy for model selection depends on the amount of data available. If ***plenty of data*** is available, It can split the data into several parts, each serving a special purpose.
- For instance, for ***hyperparameter tuning*** we may split the data into three sets: ***train / validation / test***.
- There is no general rule as to how the data should be split. A typical split is e.g. 50%/25%/25%.

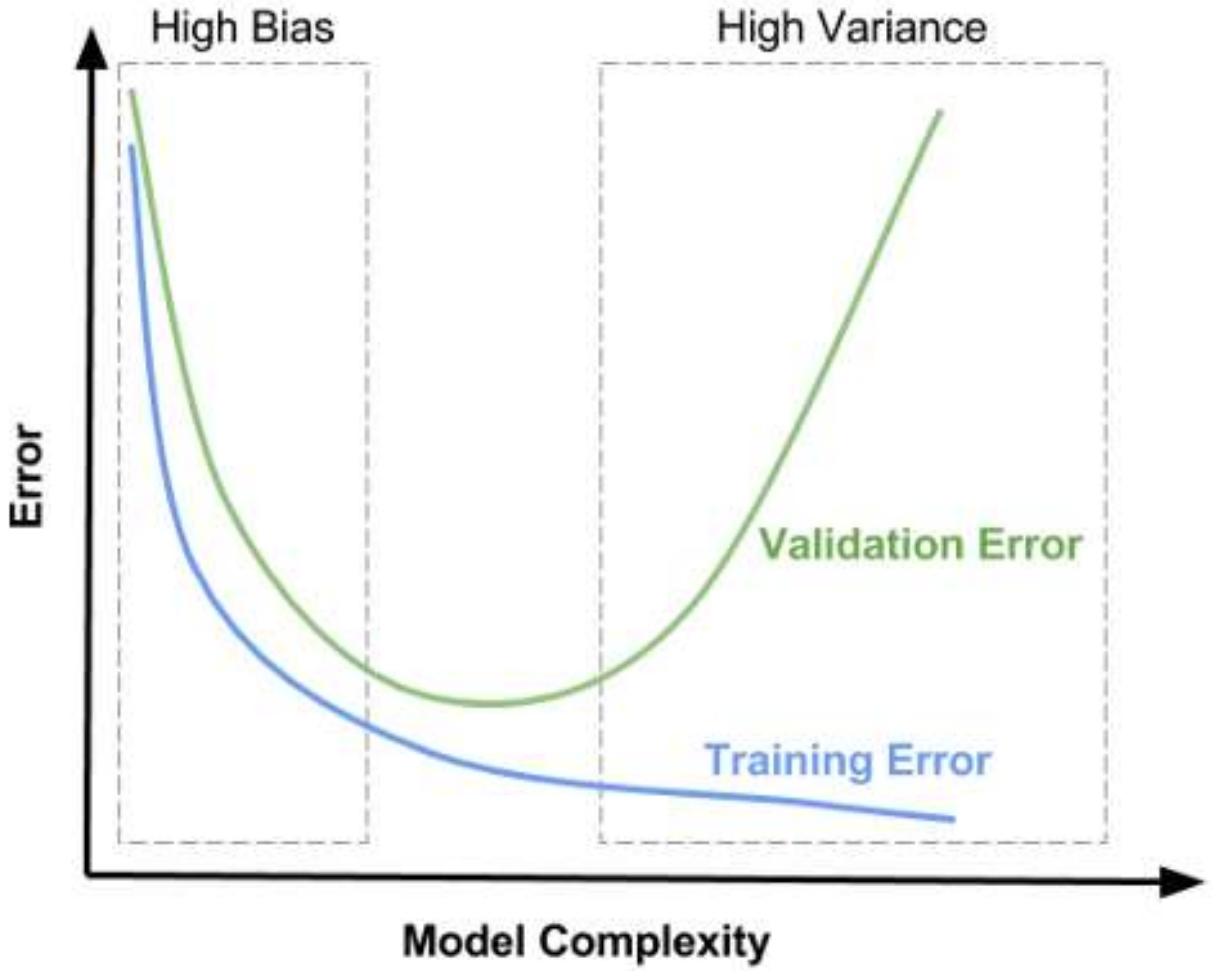
Learning curves, and why they are useful

- In a learning curve, the performance of a model both on the training and validation set is plotted as a function of the training set size.

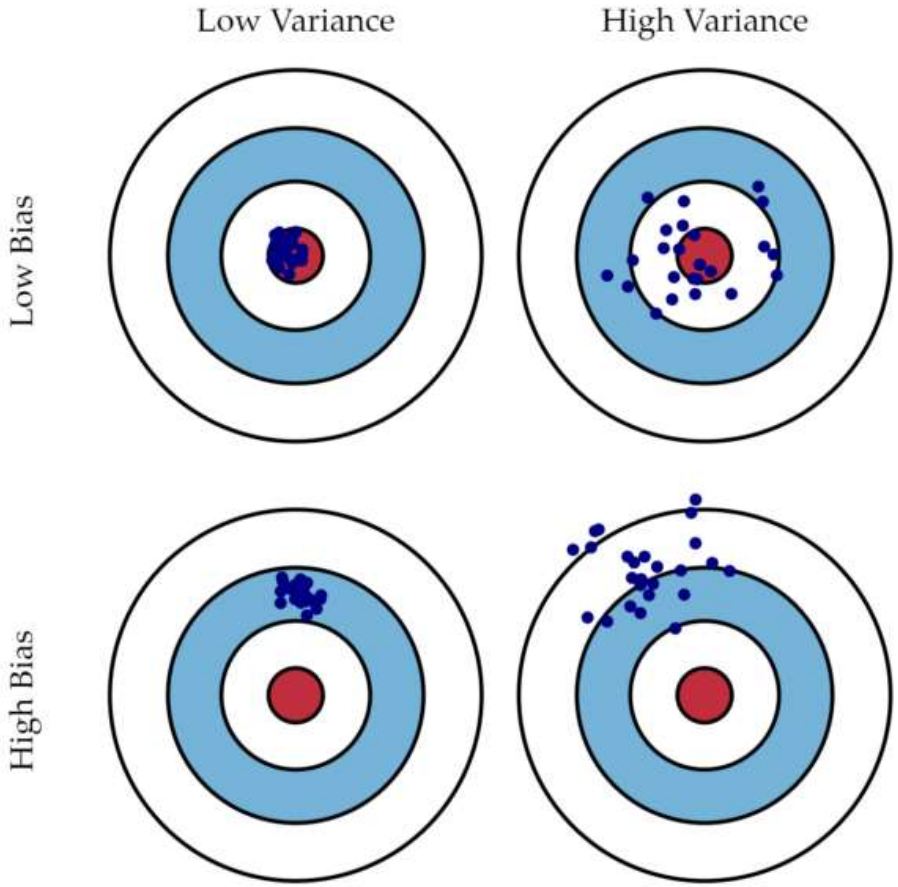


- The training score (performance on the training set) decreases with increasing training set size while the validation score increases at the same time.
- High training score and low validation score at the same time indicates that the model has overfit the data, i.e., has adapted too well to the specific training set samples.
- As the training set increases, overfitting decreases, and the validation score increases.
- Especially for data-hungry machine learning models, the learning curve might not yet have reached a plateau at the given training set size, which means the generalization error might still decrease when providing more data to the model.
- Hence, it seems reasonable to increase the training set (by adding the validation set) before estimating the generalization error on the test set, and to further take advantage of the test set data for model fitting before shipping the model. Whether or not this strategy is needed depends strongly on the slope of the learning curve at the initial training set size.

- Learning curves further allow to easily illustrate the concept of (statistical) ***bias*** and ***variance***. Bias in this context refers to erroneous (e.g. simplifying) model assumptions, which can cause the model to underfit the data. A high-bias model does not adequately capture the structure present in the data. Variance on the other hand quantifies how much the model varies as we change the training data. A high-variance model is very sensitive to small fluctuations in the training data, which can cause the model to overfit. The amount of bias and variance can be estimated using learning curves: A model exhibits high variance, but low bias if the training score plateaus at a high level while the validation score at a low level, i.e., if there is a large gap between training and validation score. A model with low variance but high bias, in contrast, is a model where both training and validation score are low, but similar. Very simple models are high-bias, low-variance while with increasing model complexity they become low-bias, high-variance.


High Bias
High Variance
Error
Validation Error
Training Error
Model Complexity

- Bias
- Bias is the error resulting from the difference between the expected value(s) of a model and the actual (or “correct”) value(s) for which we want to predict over multiple iterations. In the scientific concepts of accuracy and precision, bias is very similar to accuracy.
- Variance
- Variance is defined as the error resulting from the variability between different data predictions in a model. In variance, the correct value(s) don’t matter as much as the range of differences in value between the predictions. Variance also comes into play more when we run multiple model creation trials.


Low Variance
High Variance
Low Bias
High Bias

Divide and conquer — but do it carefully

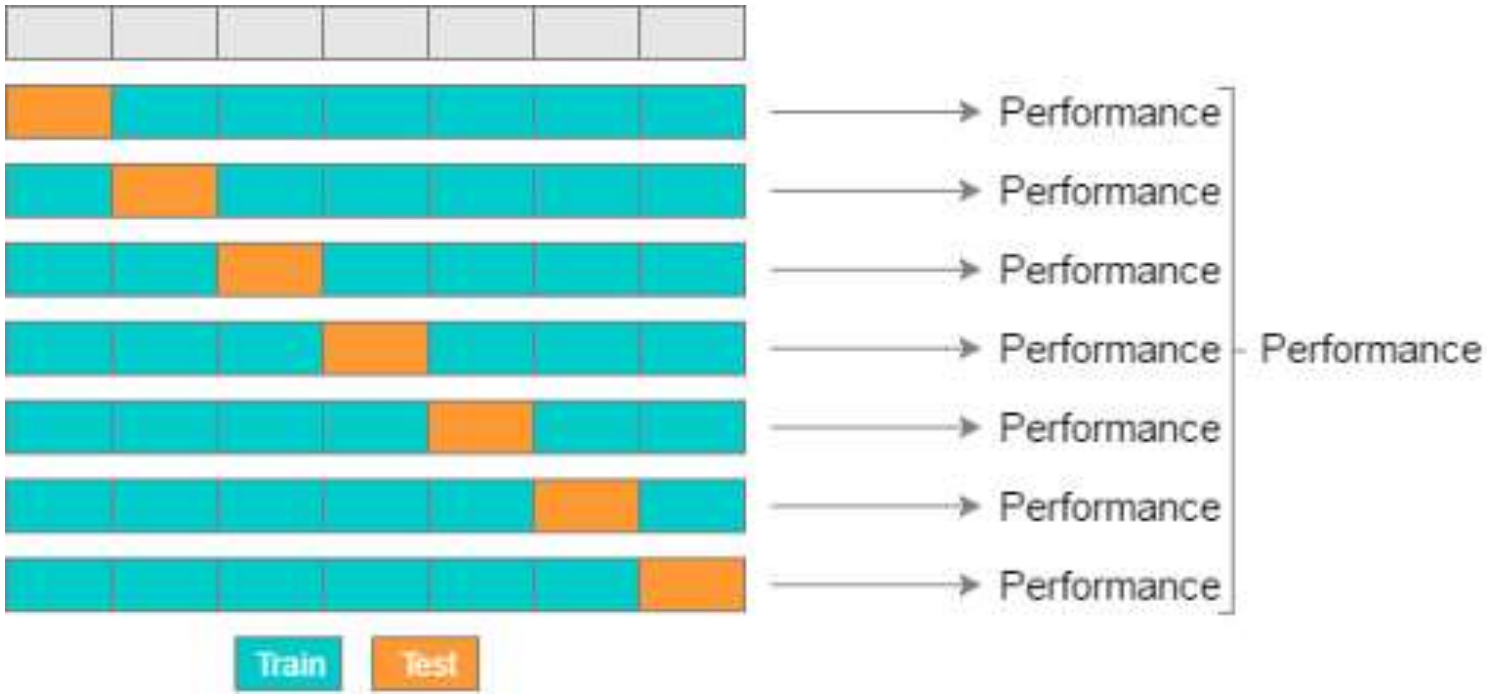
- **Training, validation, and test set** are ***sampled from the same distribution***.
- Ensure before model building that the distribution of the data is not affected by partitioning your data.
  - Example
- use ***stratified sampling***

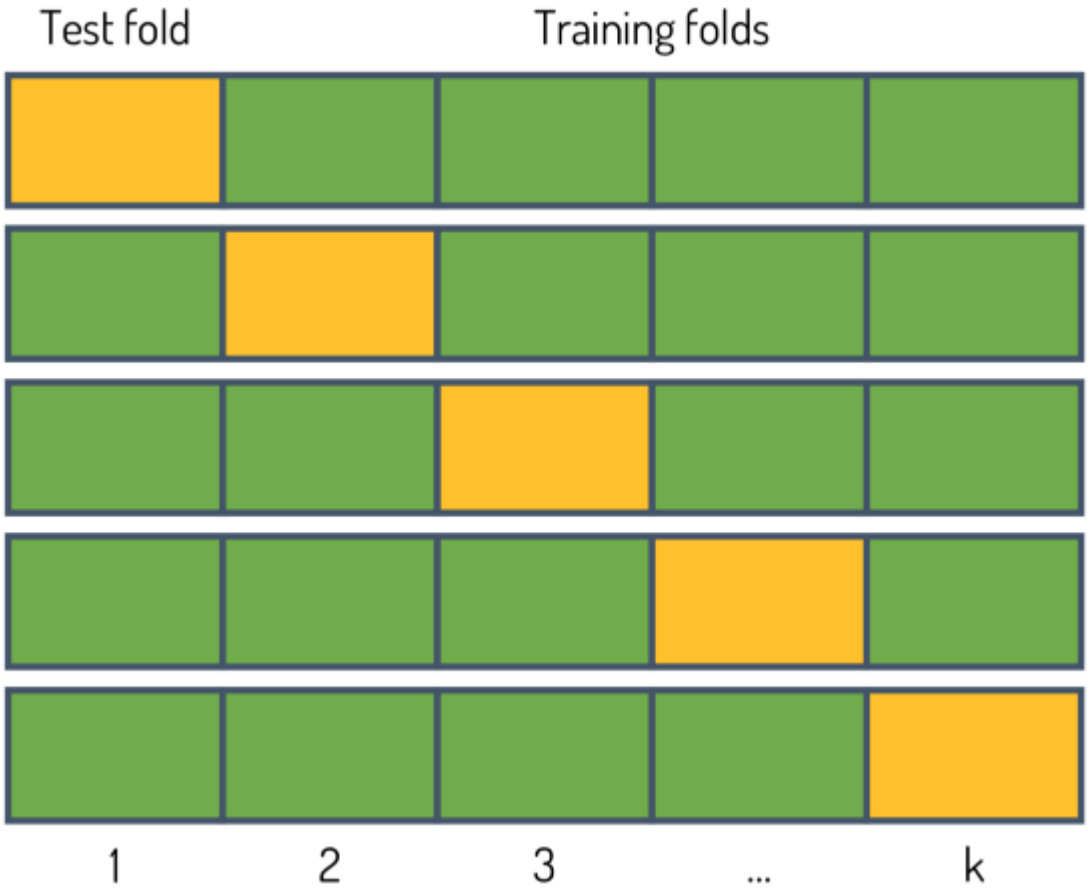
If all you have is small data

- split the data into two sets, a training and a test set.
- Augmentation of Data

K-Fold Cross Validation

- shuffles the data and splits it into *k* number of folds (groups).




Test fold
Training folds
1
2
3
...
k

Question ?