- Module 6 -
Model Assessment and Selection II

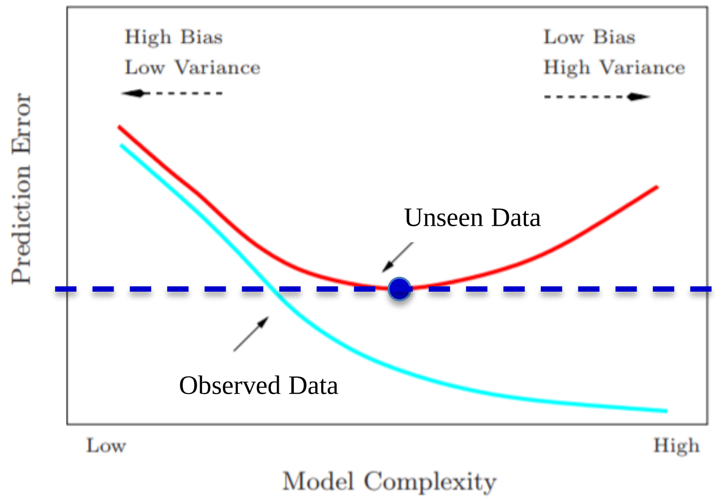
Outline

- Cross-Validation
  - K-fold Cross-Validation
  - Leave-one-out Cross-Validation
- Learning Curves

Recap

Generalization

- The central challenge in machine learning is that the algorithm must perform well on new, previously unseen inputs and not just the training data on which the model was trained



[Source: The Elements of Statistical Learning, ISBN: 978-0387848570]

> A hypothesis h with a low error rate on the training set (observed data) does not mean that it will generalize well to unseen data

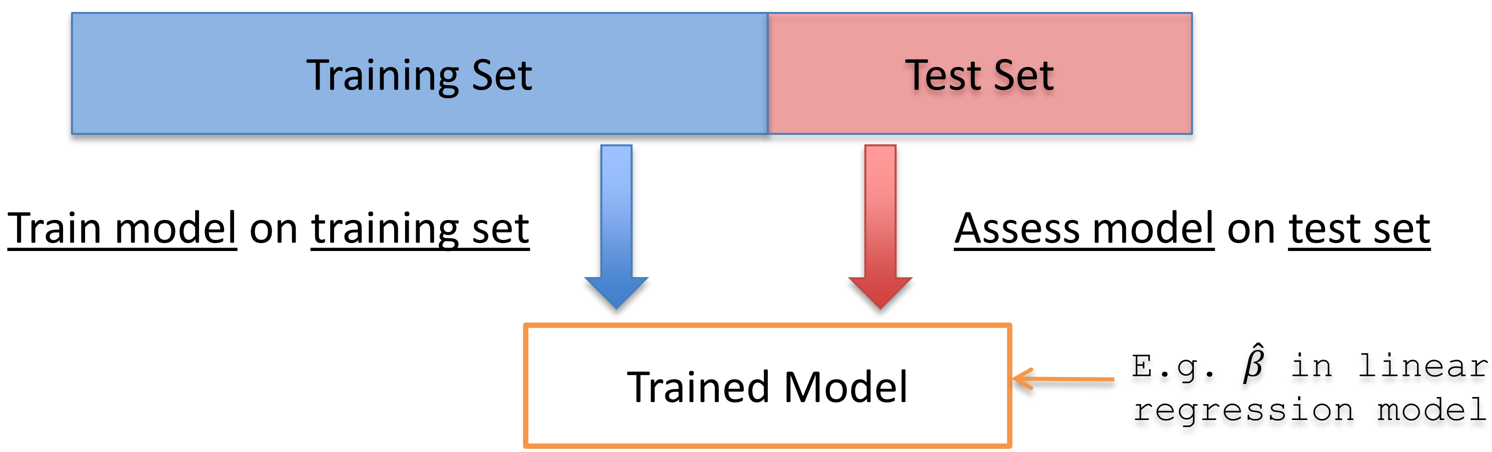
- In general, as the model complexity increases, the bias tends to decrease and the variance tends to increase
- The goal is to choose the model complexity to trade bias off with variance in such a way so as to **minimize** the **prediction error** for unseen data

Recap

Model Selection and Assessment (1)

- Recall in machine learning, there are two goals:
  - **Model selection:** estimating the performance of different models (or different model complexities) in order to choose the best one
  - **Model assessment:** having chosen a final model, estimating its prediction error on new data

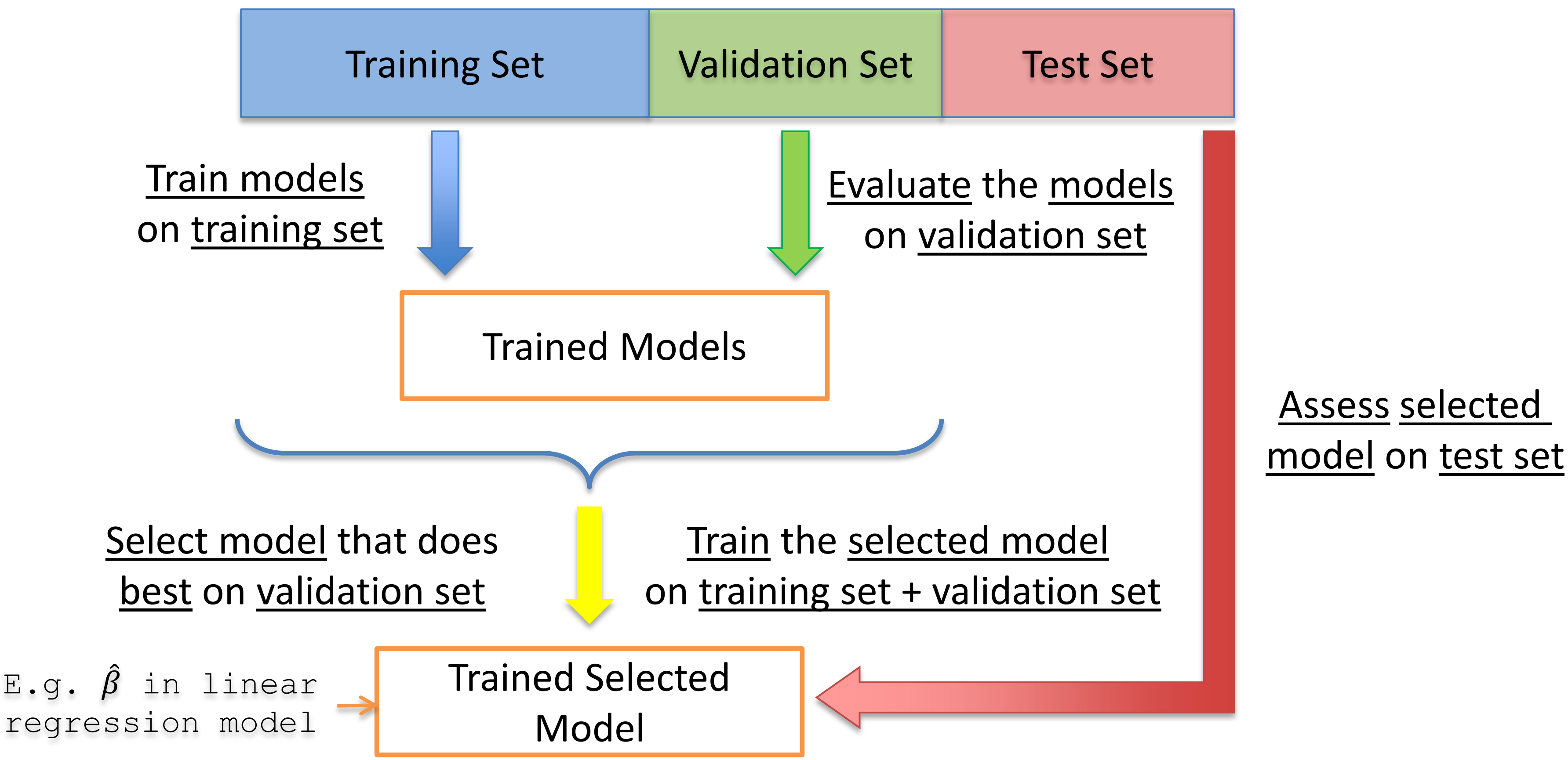
- If the model (and its complexity) to use is already known ➔ only model assessment is needed



Recap

Model Selection and Assessment (2)

- If the model (and its complexity) to use is not known
  ➔ need both model selection and model assessment



Evaluating and Choosing the Best Hypothesis (1)

- We want to learn a model (or hypothesis h), that fits the future data best

- Assumptions:
  - future data **stationarity**:
    - there is a **probability distribution** over samples (i.e., parameters such as mean and variance) that remains stationary (do not change) over time
  - best fit:
    - the **error rate** of a hypothesis is defined as the proportion of mistakes it makes (i.e., the proportion of times that $\hat{y} \neq y$ for a (x, y) sample)

Evaluating and Choosing the Best Hypothesis (2)

- To get an accurate evaluation of a hypothesis, we need to test it on a set of samples that it has not seen yet. The simplest approach, sometimes called **holdout cross-validation**, is one we have already seen
  - (randomly) split the available data into
    - a training set from which the learning algorithm produces h
    - a test set on which the accuracy of h is evaluated

      (**CAUTION**: NEVER use test set to train your model)
  - However, holdout cross-validation has the disadvantage that it fails to use all the available data
    - if we use half the data for the test set, then we are only training on half the data

      ➔ may get a poor hypothesis
    - if we reserve only 10% of the data for the test set

      ➔ may get a poor estimate of the actual accuracy

Cross-Validation (1)

- Dividing a dataset into a fixed training set and a fixed validation set can be problematic when the given dataset is small
  - ➔ results in the validation set being too small
  - ➔ a small validation set is unable to accurately estimate the prediction error

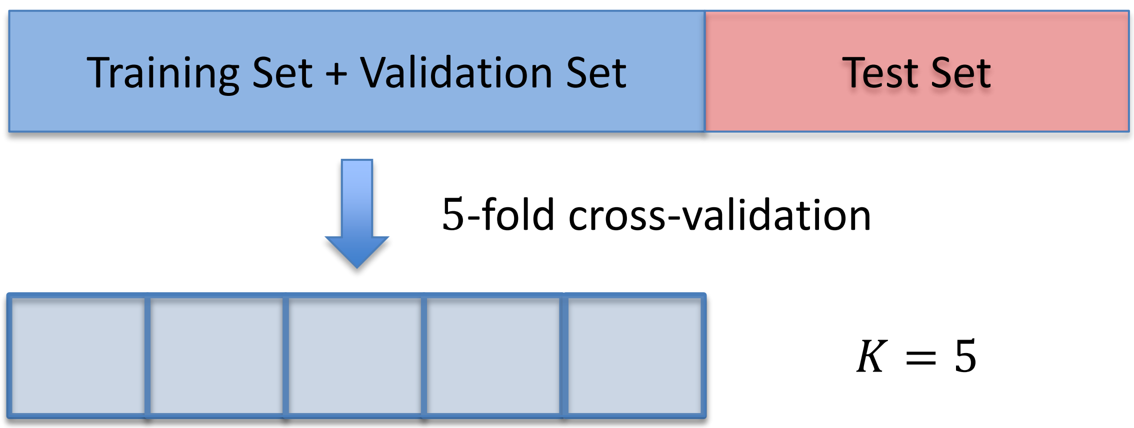
- A widely used method for estimating prediction error is **cross-validation**
  - a procedure that repeats training and validation on different (randomly chosen) subsets of the original training set
  - incurs increased computational cost

➔ **K-fold cross-validation**

➔ **leave-one-out cross-validation** (LOOCV)

Cross-Validation (2)

- K-fold cross-validation allows us to make more out of the data and still get an accurate estimate
  - the idea is that each sample serves double-duty: as training set and validation set



- Typical values for K are $K = 5$ and $K = 10$
  - enough to give an estimate that is statistically likely to be accurate
  - but at a cost of K times longer computation time
- The extreme is $K = N$, where N is the number of samples in the training set, known as leave-one-out cross-validation

Cross-Validation (3)

- K-fold cross-validation:

  - sub-divide the training set into K non-overlapping subsets of equal size (e.g., $K = 5$)
  - perform K trials of learning – on trial k,
    - all subsets except the k-th subset is used as the **training set** to train the model
    - the k-th subset is used as the **validation set** to evaluate the trained model
  - the **cross-validation estimate** of **prediction error** of the given model is the average of the cross-validation estimates across K trials

    ➔ it is expected that the average cross-validation estimate of the K trials should be a better estimate than that from a single trial

Cross-Validation (4)

- Specifically, for a p-th degree polynomial regression
  - divide the training set into K equal subsets
  - for each $k = 1, 2, \dots, K$
    - train the model with complexity p on the $K-1$ subsets (training set) with k-th subset removed, to obtain the least squares estimate, $\hat{\beta}^{-k}$
    - evaluate the trained model using $\hat{\beta}^{-k}$ on the k-th subset (validation set) to obtain **cross-validation estimate**, Err_{CV}^{k}
      (which can be mean absolute error (MAE), root mean squared error (RMS), etc.)
  - compute the average cross-validation estimate, Err_{CV}, across k

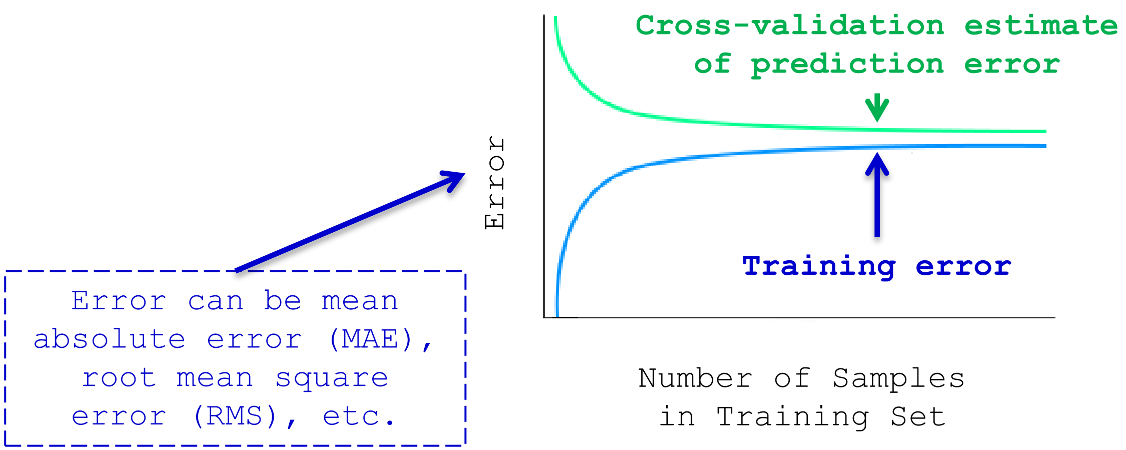
$$Err_{CV} = \frac{1}{K}\sum_{k=1}^{K} Err_{CV}^{k}$$

Cross-Validation (5)

- To select and assess the final model complexity p (for p-th degree polynomial regression model)
  - Final model selection:
    - repeat the K-fold cross-validation for different values of p
    - select the value of p that gives the smallest Err_{CV}
  - Final model assessment:
    - train final model with selected complexity p on all of the training set (training set + validation set)
    - evaluate trained model on the test set to obtain the final prediction error

Learning Curves (1)

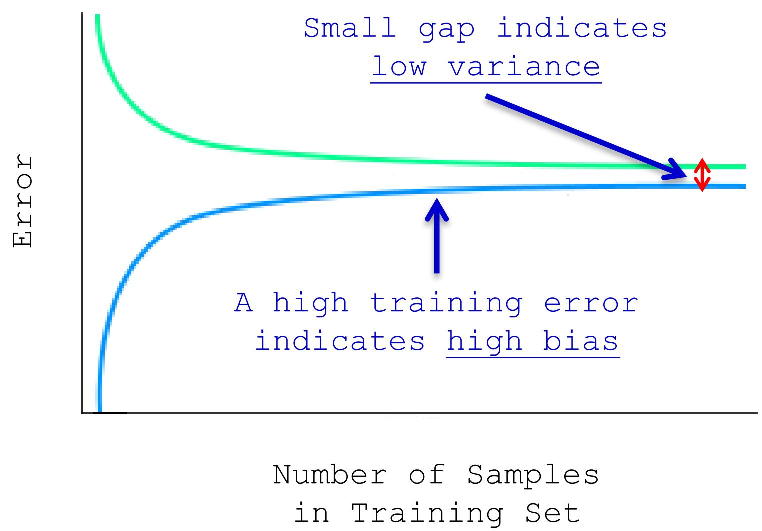
- The **learning curve** for a model is a plot showing the training error and the cross-validation estimate of prediction error as a function of the number of samples N in the training set



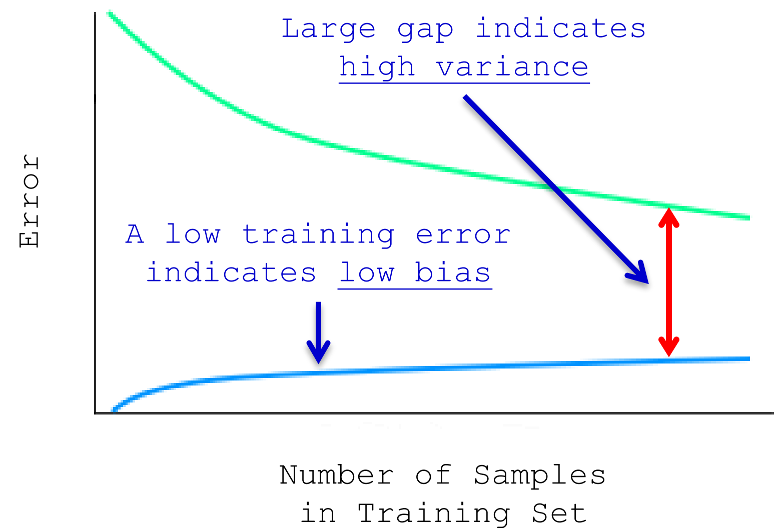
- The learning curve plot is a useful tool to diagnose bias and variance and can help determine whether a model is underfitting (low complexity) or overfitting (high complexity)
- More importantly, a learning curve can also provide insight as to whether more samples need to be collected

Learning Curves (2)

Low Complexity Model (underfit)



High Complexity Model (overfit)



$$gap = \; cross\_validation\ estimate\ \; - training\ error$$

- Ideally, a model with low bias and low variance is desired
  - for high complexity models (low bias), the cross-validation estimate could converge towards the training error (low variance) if more training samples were added (i.e., increase N)