- Module 5 -
Model Assessment and Selection I

Outline

- Generalization
- Model Complexity
  - Underfitting and Overfitting
  - Bias-Variance Tradeoff
- Model Selection and Assessment

Recap

Linear Regression Models: Polynomial Regression

- **Polynomial regression** is a special case of the linear regression model in which the relationship between x and y is modelled as a p-th degree polynomial in x

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2 + \cdots + \hat{\beta}_p x_i^p, \quad \forall i = 1, 2, \ldots, N$$

- Considering all N samples in the data, we can rewrite in matrix notation as

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\beta}$$

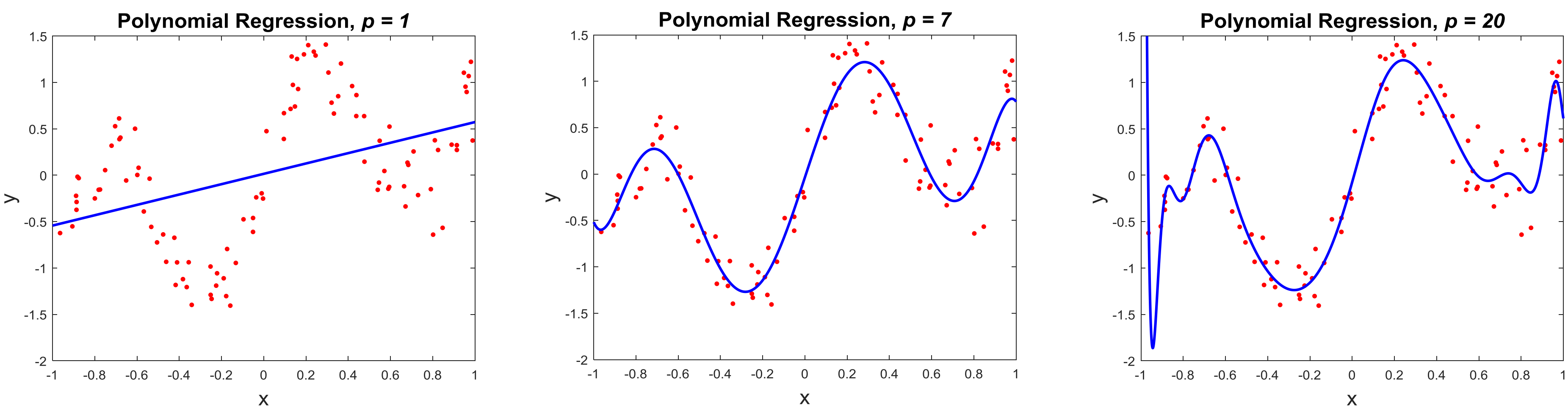
where $\hat{\boldsymbol{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_N \end{bmatrix}$, $\boldsymbol{X} = \begin{bmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^p \\ 1 & x_2 & x_2^2 & \ldots & x_2^p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_N & x_N^2 & \ldots & x_N^p \end{bmatrix}$ and $\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_p \end{bmatrix}$

- The least squares estimate for polynomial regression model remains as $\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$

Recap

Linear Regression Models: Model Complexity

- **Model complexity** refers to the number of parameters used in the model and represents its ability to capture the patterns in the data
  ➔ in polynomial regression models, model complexity is determined by p
- **Question:** What order of polynomial regression should be used to fit the following data?



➔ **Machine learning models** generally perform best when their complexity is appropriate for the true complexity of the task and the amount of training data provided

Generalization (1)

- 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
  - the ability to perform well on previously unobserved inputs is called **generalization**
  - the **generalization performance** of a trained model can be measured by its prediction capability on independent test data
    - ➔ guides the choice of learning method or model
    - ➔ provides a useful indication of the quality of the chosen model on new data

Generalization (2)

- In practice, when training a machine learning model for a given dataset, it is usually divided into two subsets
  - training set — subset to train the model (e.g., data fed to LinearRegression.fit() method)
  - test set — subset to test the model (e.g., data fed to LinearRegression.predict() method)

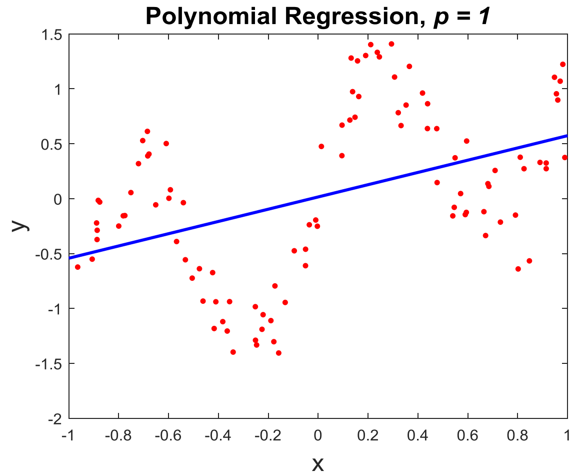
  (**CAUTION**: NEVER use test set to train your model)

- The **training error** can be computed on the training set
  ← desirable for the training error to be low
- The **test error** (or **generalization error)** is typically measured by its performance on a test set
  ← also desirable for the test error to be low

Model Complexity: Underfitting and Overfitting (1)

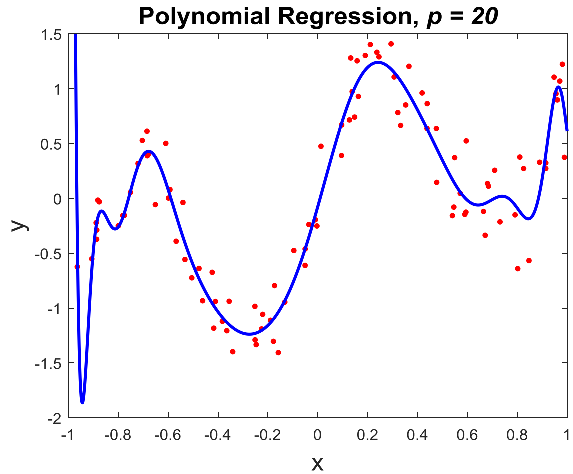
- There are two (2) factors that correspond to this central challenge in machine learning: **underfitting** and **overfitting**

- Underfitting occurs when the model is not complex enough and unable to obtain a sufficiently low error on the training set
  - ➔ the model is unable to capture the underlying trends in the observed data



Model Complexity: Underfitting and Overfitting (2)

- Overfitting occurs when the model is overly complex, and
  - the model is able to obtain a low training error on observed data by trying to fit to the particularities (usually noise)
  - the model fails to reflect the overall trend of the data and can vary greatly when given different sets of test data
    - ➔ the gap between the training error and test error is large



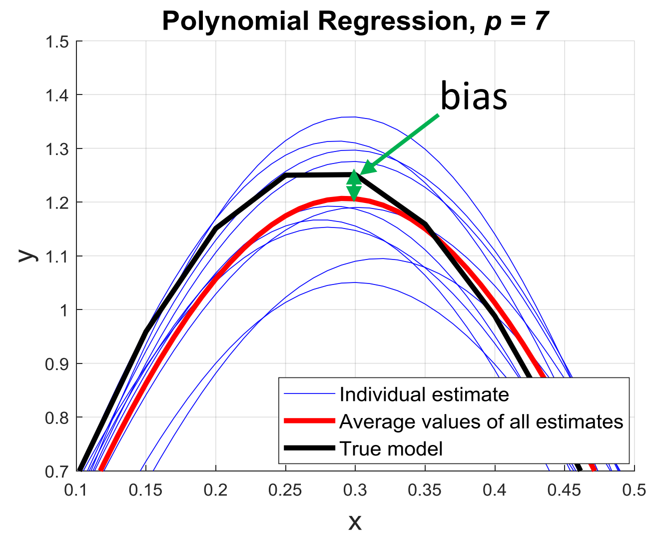
Model Complexity: Bias-Variance Tradeoff (1)

- The **bias-variance tradeoff** is a central problem in supervised learning

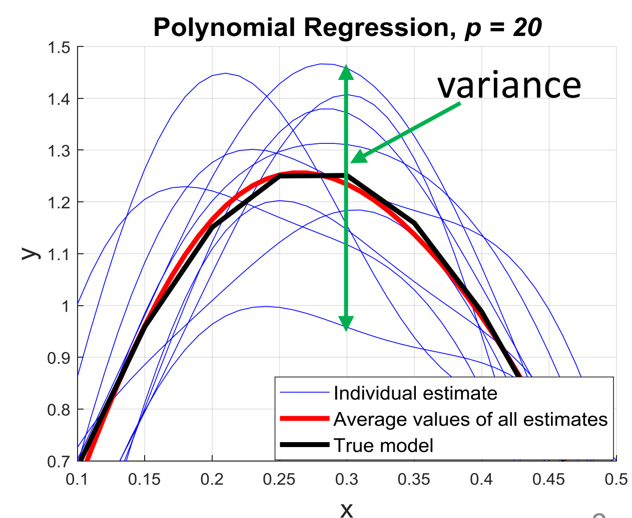
- Ideally, one wants to choose a model that both accurately captures the regularities in its training data, but also generalizes well to unseen data
  ➔ typically impossible to do both simultaneously

Model Complexity: Bias-Variance Tradeoff (2)

- The **bias** of a model measures the **difference** between the average value of a model that is fitted over all possible sets of data (**red line**) and the **true model** (black line) being estimated



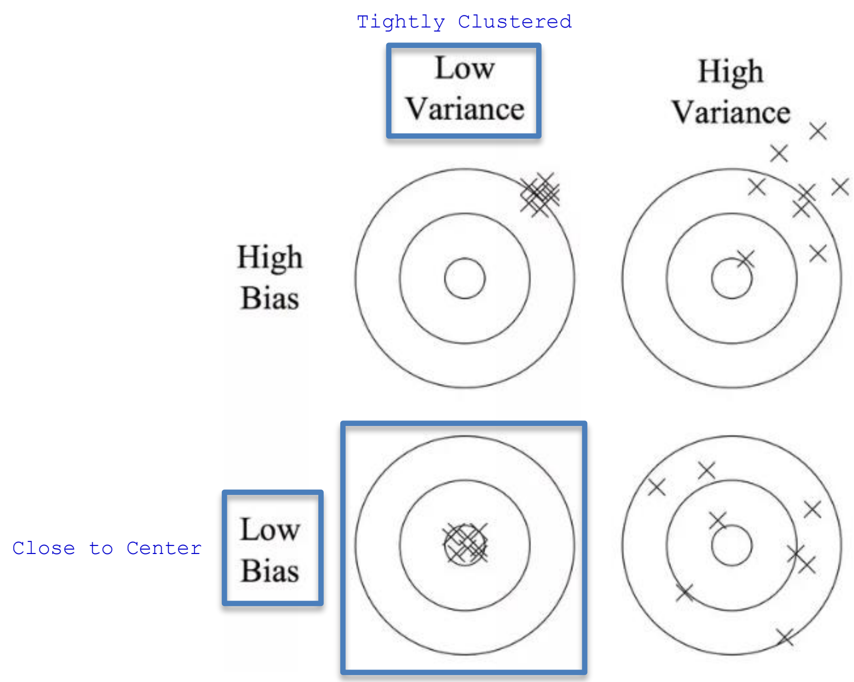
- The **variance** of a model describes how much the values of a model **spread** across all possible sets of data (**blue lines**)



Model Complexity: Bias-Variance Tradeoff (3)

- **Example:** Dart-throwing

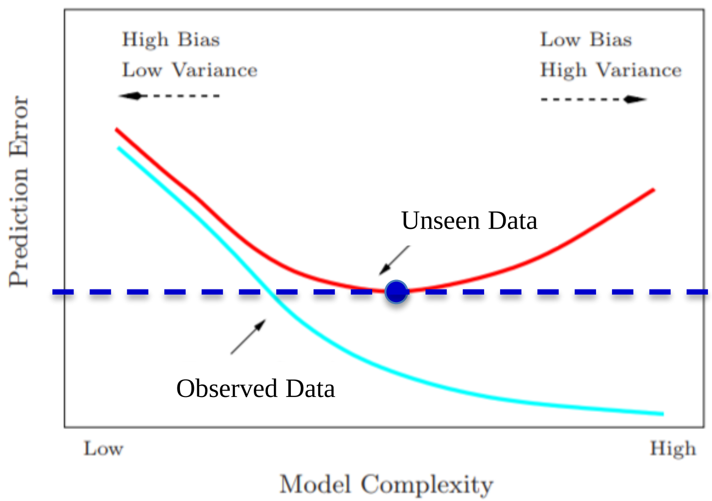
[Source: P. Domingos, "A few useful things to know about machine learning," Commun. ACM, vol. 55, no. 10, pp. 78-87, 2012]



Model Complexity: Bias-Variance Tradeoff (4)

- In general, as the model complexity increases, the bias tends to decrease and the variance tends to increase

- The figure shows the typical behavior of the training error (**cyan**) and test error (**red**), as model complexity is varied
  - The training error tends to decrease when the model complexity is increased (fitting the data harder)
  - However, with too much fitting, the model adapts itself too closely to the training data, and will not generalize well (i.e., have large test error)
  - In contrast, if the model is not complex enough, it will underfit and may have high bias, again resulting in poor generalization



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

- 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** (which includes model bias, model variance and observation noise) for unseen data

Model Complexity: Bias-Variance Tradeoff (5)

- As discussed, there is a bias-variance tradeoff as the model complexity (e.g., p-th degree polynomial regression model) varies
- Different ways to determine the best model complexity (or different models), include the following (simplest and most widely used)
  - **Cross-Validation**

    sub-divide the training data into

training set	used to train models with different complexities
validation set	used to choose the best one

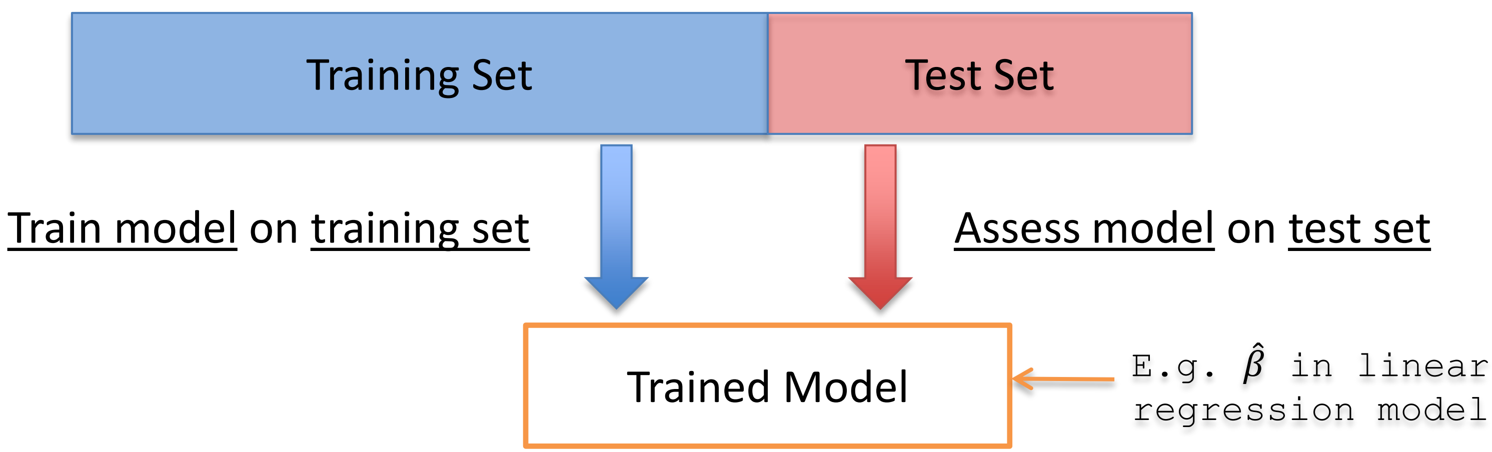
  - **Regularization**

    introduce **regularization term** (or **penalty term**) to penalize an overly complex model, thereby favoring simpler models with less room to overfit

Model Selection and Assessment (1)

- 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



Model Selection and Assessment (2)

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