Robotics & Internet-of-Things

جامعة الأمير سلطان PRINCE SULTAN UNIVERSITY

CS316: INTRODUCTION TO AI AND DATA SCIENCE

CHAPTER 7
STATISTICAL LEARNING

LECTURE

Dataset Distribution | Bias vs Variance
Overfitting vs. Underfitting

Prof. Anis Koubaa

Nov 2024

www.riotu-lab.org

CHAPTER 7
STATISTICAL LEARNING

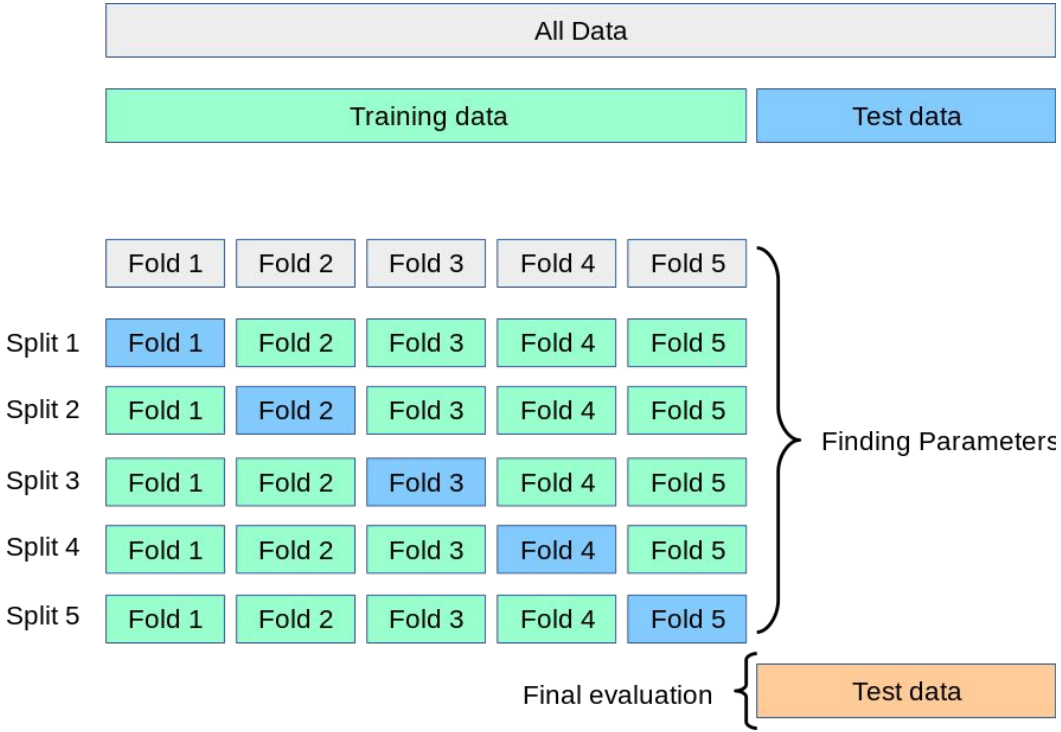
LECTURE 2
FOUNDATIONS OF PREDICTIVE MODELING IN DATA SCIENCE

Dataset Distribution for Training

CS316: INTRODUCTION TO AI AND DATA SCIENCE

Prof. Anis Koubaa

Application: Cross Validation



https://scikit-learn.org/stable/modules/cross_validation.html

- τ - **Specific Training Dataset:**
  - Represents a specific instance of a training dataset.
  - Used to discuss outcomes related to this particular dataset.
  - Focuses on the concrete realization of data points.
- T - **Random Training Set:**
  - Denotes the concept of a random training set, highlighting variability in dataset selection.
  - Generalizes findings across different instances of training data.
  - Used in theoretical analyses for expected performance over all possible training sets.

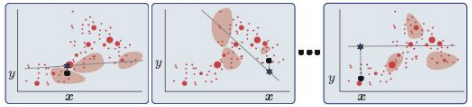


Figure 2.2: The expected generalization risk is the weighted-average loss over all possible pairs (x, y) and over all training sets.

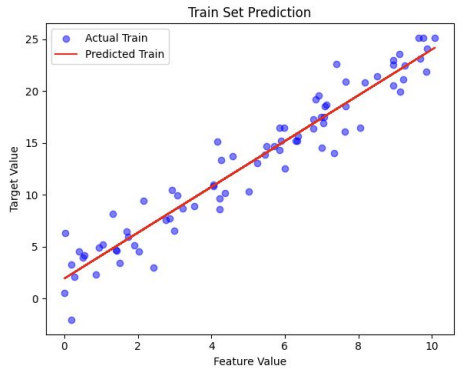
Unbiased Estimation of Generalization Risk

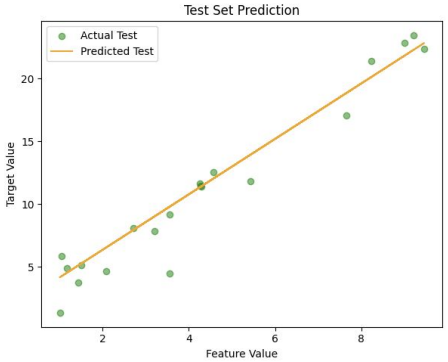
- **Objective**: Estimate the unbiased generalization risk of a predictive model.
- **Notation**:
  - **Training Data**, T: The dataset used to train the model.
  - **Test Data**, T': A separate dataset used to evaluate the model.
  - **Generalization Function**, $\hat{g}_\tau^G$: The predictive model trained on T and evaluated on T'.
- **Key Concept**:
  - The **Test Loss** is used to estimate the generalization risk of $\hat{g}_\tau^G$ without bias.
- **Formula**:

$$\frac{1}{n'}\sum_{i=1}^{n'} \text{Loss}(Y_i', \hat{g}_\tau^G(X_i')) =: l_{T'}(\hat{g}_\tau^G)$$

  where n' is the size of the test sample T', and $\text{Loss}(Y_i', \hat{g}_\tau^G(X_i'))$ calculates the discrepancy between the predicted and actual outcomes.
- **Assumption**: Independence between training set T and test set T' is crucial for unbiased estimation.
- **Application**: This methodology allows for the assessment of how well the model generalizes to new, unseen data, a fundamental aspect of supervised learning.

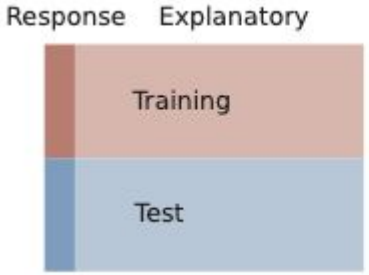
```
Train Loss (MSE): 3.301
Test Loss (MSE): 2.901
Train Accuracy (R^2 Score): 0.931
Test Accuracy (R^2 Score): 0.937
```



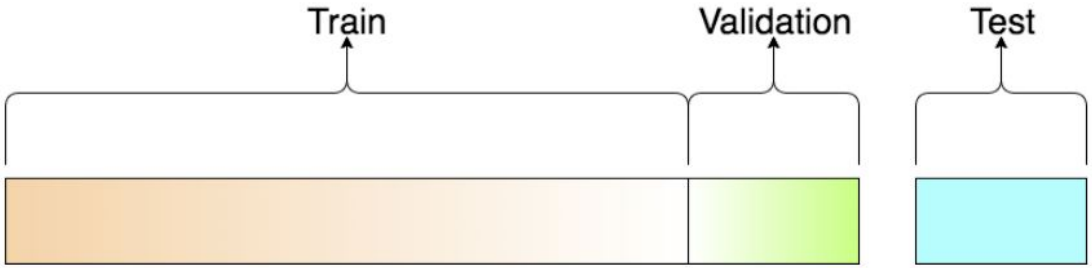


Comparing Predictive Performance

- **Objective**: Assess and compare the predictive performance of various learners within the function class G.
- **Notation**:
  - **Training Set**, τ: Used to train different learners ($g^{G_1}, g^{G_2}, \ldots$).
  - **Test Set**, τ': Initially used as a validation set to select the best learner based on the smallest test loss.
  - **Validation Set**: A subset of the dataset used to tune models' parameters and select the best performing model.
  - **Final Test Set**: An additional, separate dataset used to evaluate the predictive performance of the selected best learner.
- **Process**:
  - **Data Splitting**: The overall dataset is randomly divided into training and test/validation sets.
  - **Model Training**: Use the training data to construct and train various learners.
  - **Model Selection**: Use the validation set (initial test set) to select the learner with the best performance.
  - **Performance Evaluation**: Utilize a third set, a final test set, to assess the predictive performance of the chosen learner.



Train/Dev/Test Split



Typical 100-1000-10000	60%	20%	60%
Huge data size 1000000	98%	1%	1%

Mismatched Data Distribution

Train Dataset

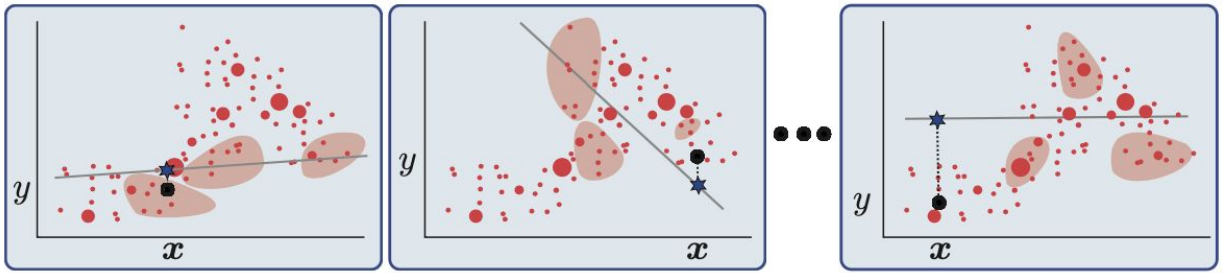
High Resolution

Different Distribution

Dev/Test Dataset

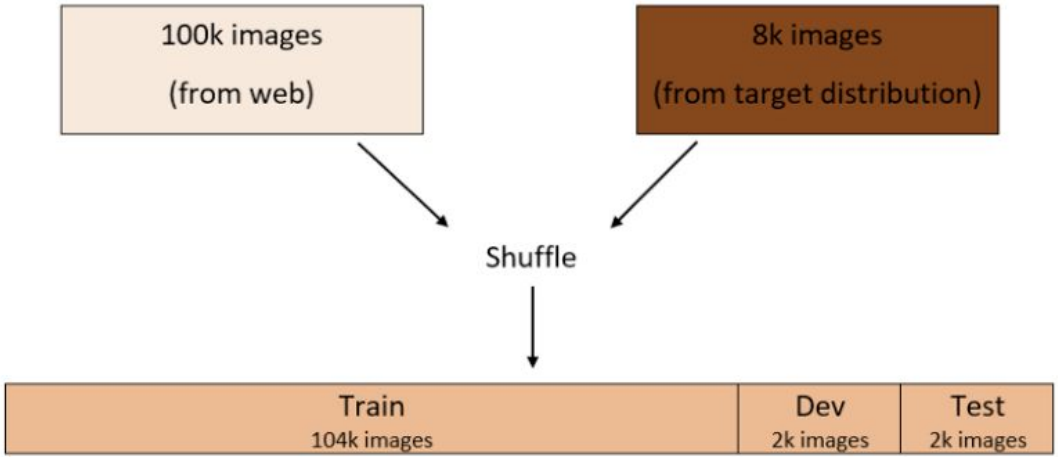
Low Resolution

Make sure that distributions of Train dataset and Dev/Test dataset are similar

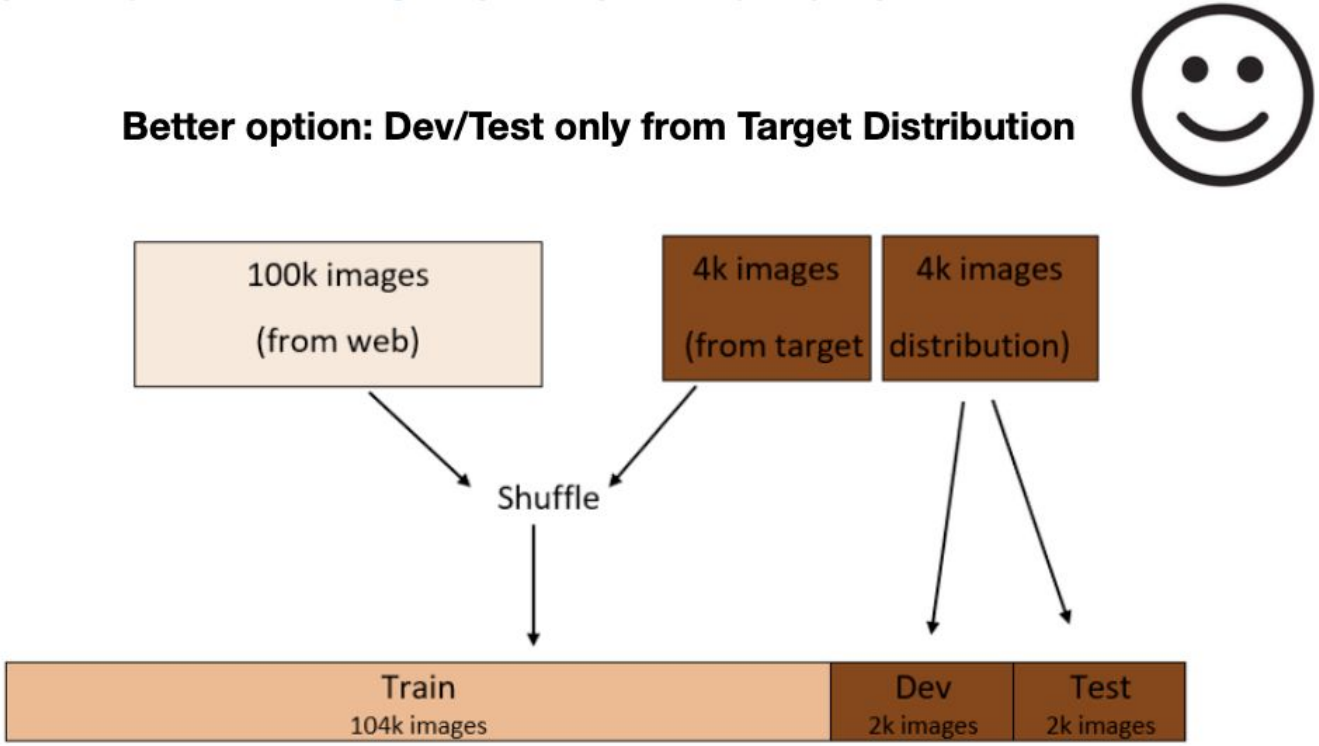


How to Split Mismatched Distribution Data

Option 1: **Shuffle images from different distributions**



How to Split Mismatched Distribution Data



CHAPTER 7
STATISTICAL LEARNING

LECTURE 2
FOUNDATIONS OF PREDICTIVE MODELING IN DATA SCIENCE

Bias vs. Variance

CS316: INTRODUCTION TO AI AND DATA SCIENCE

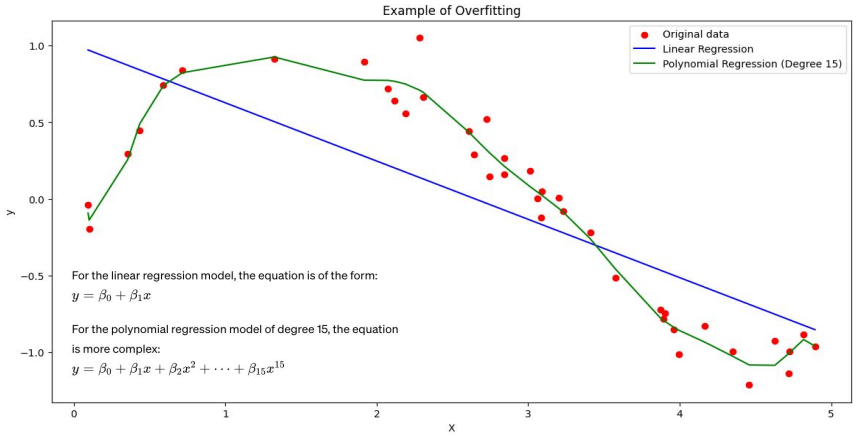
Prof. Anis Koubaa

Understanding Bias vs. Variance

الانحياز والتباين

- **Objective**: Dissect the trade-offs between bias and variance in statistical learning models.
- **Definitions**:
  - **Bias**: The error due to overly simplistic assumptions in the learning algorithm. High bias can cause the model to miss relevant relations between features and target outputs (underfitting).
  - **Variance**: The error from sensitivity to small fluctuations in the training set. High variance can cause model to capture random noise in the training data (overfitting).
- **Key Concept**:
  - **Bias-Variance Tradeoff**: Balancing the complexity of the model to minimize both errors and achieve good generalization.
- **Illustration**:
  - High Bias: Assumes a linear relationship in a non-linear problem.
  - High Variance: Fits noise in the data as if it were a real pattern.
- **Mathematical Perspective**:
  - Expected prediction error for a given point x can be decomposed as:

$$\text{Error}(x) = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error}$$



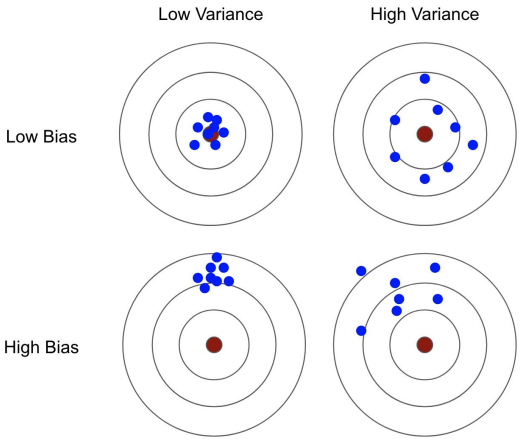
Simplifying the Bias-Variance Tradeoff

$$\text{Error}(x) = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error}$$

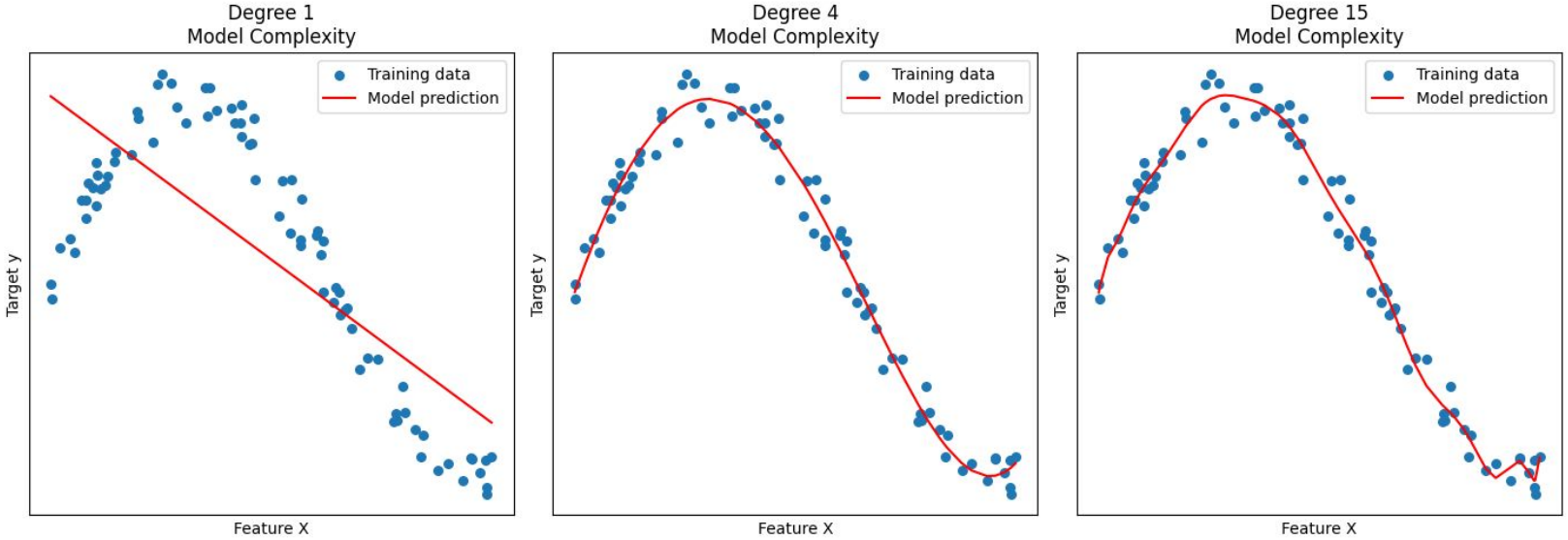
- **Understanding Prediction Error**:
  - Imagine you're trying to hit a target with arrows. Your goal is to be as close to the bullseye (the true value) as possible with each shot (prediction).
- **Components of Prediction Error**:
  - **Bias**: Suppose you consistently hit the target to the left of the bullseye. The average distance of your shots from the bullseye represents **Bias**. High bias means you're not even aiming correctly; your model is missing the mark because it's too simplistic.
  - **Variance**: Now, imagine your shots are spread out, sometimes to the left, sometimes to the right, above, and below the bullseye. The spread of your shots represents **Variance**. High variance means your model is sensitive to the training data and shoots all over the place.
  - **Irreducible Error**: There's always a bit of wind or some unseen bump in the ground that affects where your arrow lands, no matter how good your bow and your aim are. This is the **Irreducible Error**, which comes from randomness or unknown factors in the problem itself that we cannot predict or reduce.

- **Bias-Variance Tradeoff**: Balancing the complexity of the model to minimize both errors and achieve good generalization.
- **Illustration**:
  - High Bias: Assumes a linear relationship in a non-linear problem.
  - High Variance: Fits noise in the data as if it were a real pattern.
- **Mathematical Perspective**:
  - Expected prediction error for a given point x can be decomposed as:

$$\text{Error}(x) = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error}$$



Understanding Bias vs. Variance

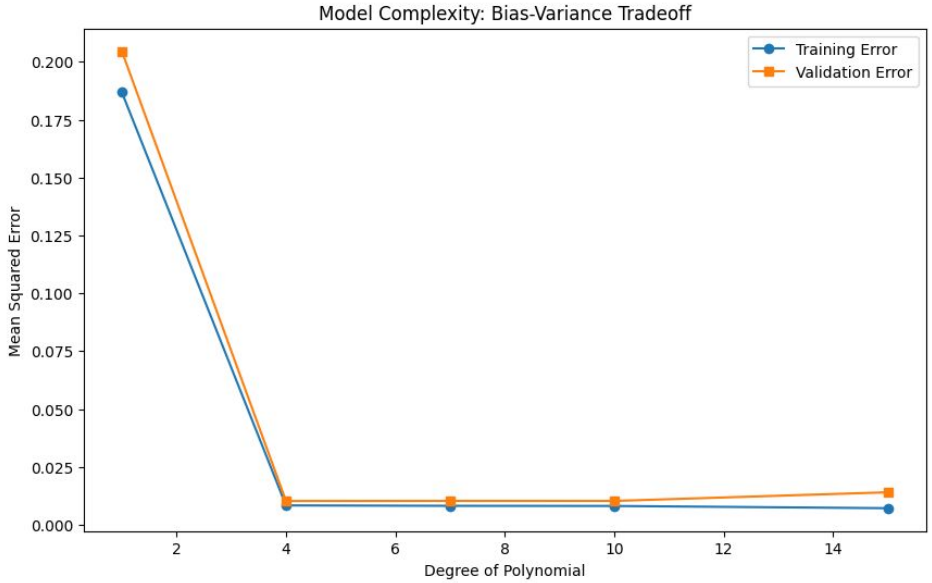


Understanding Bias vs. Variance

- **Training Error** decreases as the model complexity increases, which shows that the model is getting better at fitting the training data. However, this does not necessarily

mean it will perform well on unseen data.

- **Validation Error** initially decreases as the model becomes more capable of capturing the underlying pattern in the data but starts to increase at higher degrees of polynomial, indicating overfitting. This increase in validation error is due to the model's high variance, capturing noise in the training data as meaningful patterns, which do not generalize to new data.

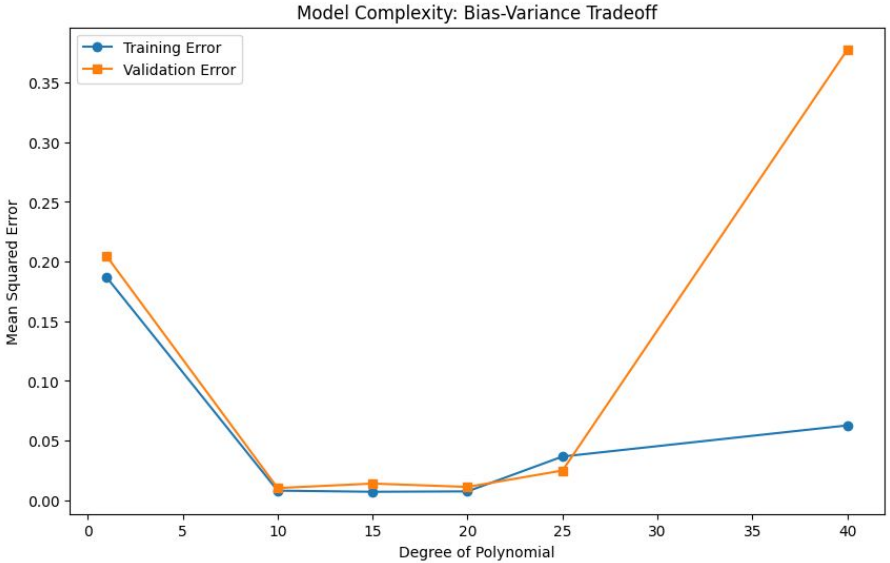


Understanding Bias vs. Variance

- **Training Error** decreases as the model complexity increases, which shows that the model is getting better at fitting the training data. However, this does not necessarily

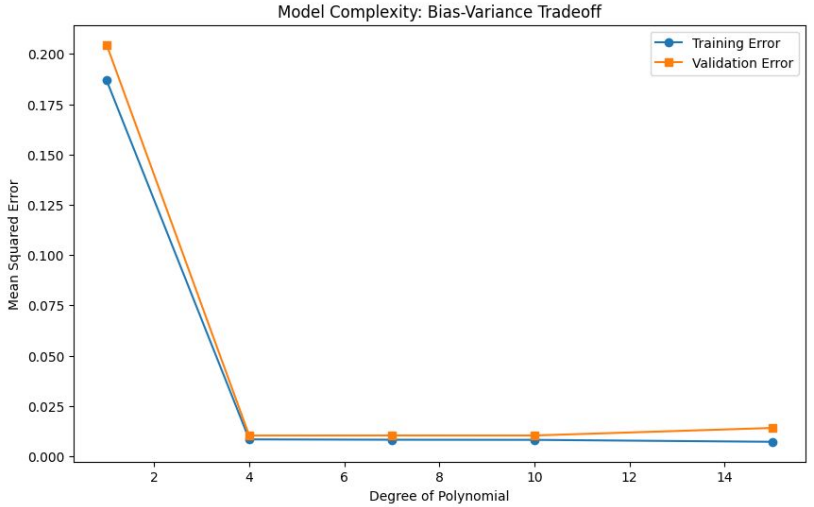
mean it will perform well on unseen data.

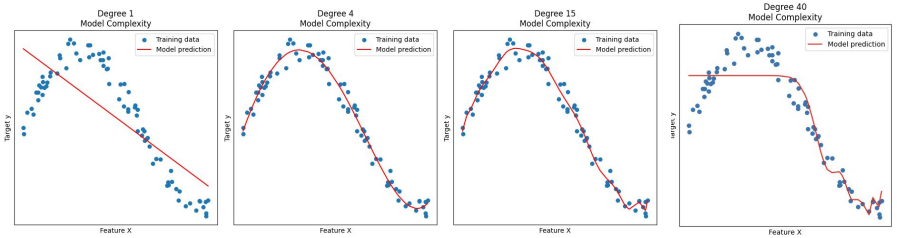
- **Validation Error** initially decreases as the model becomes more capable of capturing the underlying pattern in the data but starts to increase at higher degrees of polynomial, indicating overfitting. This increase in validation error is due to the model's high variance, capturing noise in the training data as meaningful patterns, which do not generalize to new data.



Understanding Bias vs. Variance

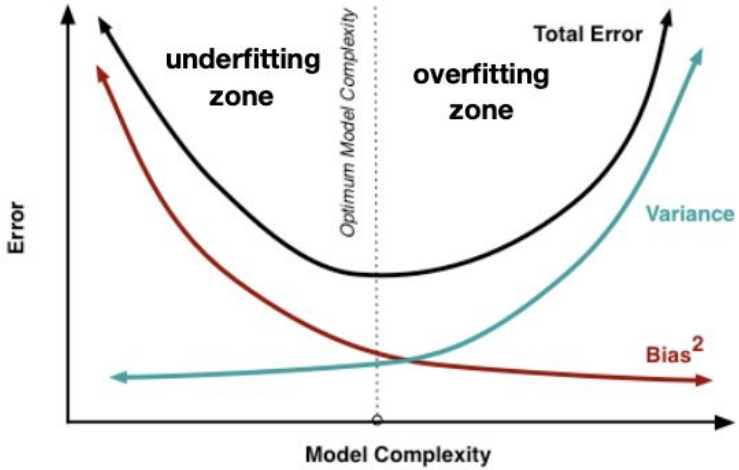
- **Bias**: At lower degrees of polynomial (e.g., 1), both training and validation errors are high because the model is too simple to capture the underlying pattern in the data (underfitting). This is indicative of high bias, where the model's assumptions are too far from the true relationship.
- **Variance**: At higher degrees of polynomial (e.g., 15), the training error is very low because the model fits the training data too closely, including noise. However, the validation error is high because the model fails to generalize to new data. This is indicative of high variance, where the model is too sensitive to small fluctuations in the training data.
- **Bias-Variance Tradeoff**: The optimal model complexity (in this case, around degree 4 to 7) achieves a balance between bias and variance. It is complex enough to capture the underlying pattern in the data (low bias) but not so complex that it fits the noise in the training data (controlled variance). This balance results in the lowest validation error, indicating the best generalization to unseen data.





How to Split Mismatched Distribution Data

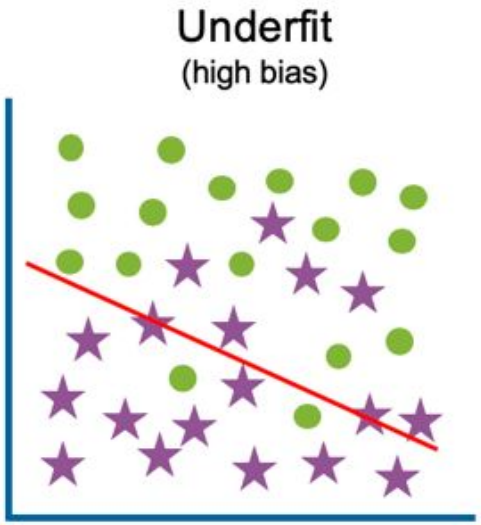
To build a good model, we need to find a good balance between bias and variance such that it minimizes the total error.



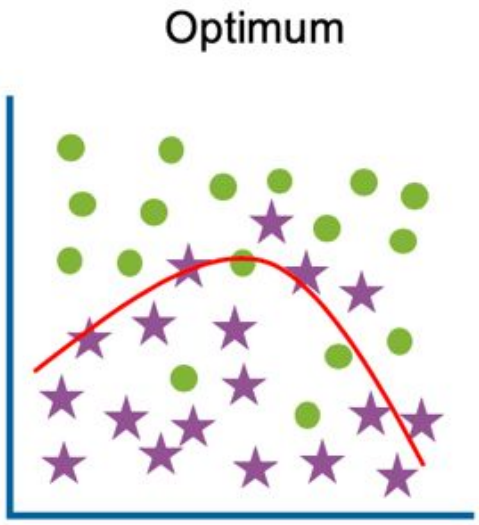
$$Err(x) = \left(E[\hat{f}\,(x)] - f(x)\right)^2 + E\left[\left(\hat{f}\,(x) - E[\hat{f}\,(x)]\right)^2\right] + \sigma_e^2$$

$$Err(x) = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error}$$

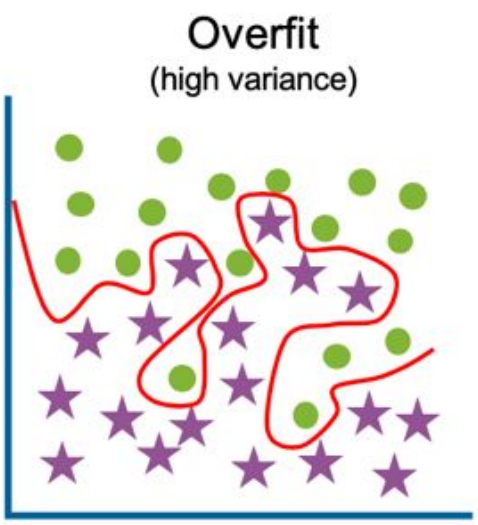
Overfitting | Optimal | Underfit



High training error
High test error



Low training error
Low test error



Low training error
High test error

Bias and Variance Cases

Train Error	1%	15%	15%	0.5%
Test Error	11%	16%	30%	1%
Result	High Variance	High Bias	High Bias High Variance	Low Bias Low Variance
Fitting Type	Overfitting	Underfitting	Overfitting and Underfitting	Good Fitting