Machine Learning

Lecture 16: Model Validation Techniques, Overfitting, Underfitting

COURSE CODE: CSE451

2023

Course Teacher

Dr. Mrinal Kanti Baowaly

Associate Professor

Department of Computer Science and Engineering, Bangabandhu Sheikh Mujibur Rahman Science and Technology University, Bangladesh.

Email: mkbaowaly@gmail.com

Model Validation Techniques

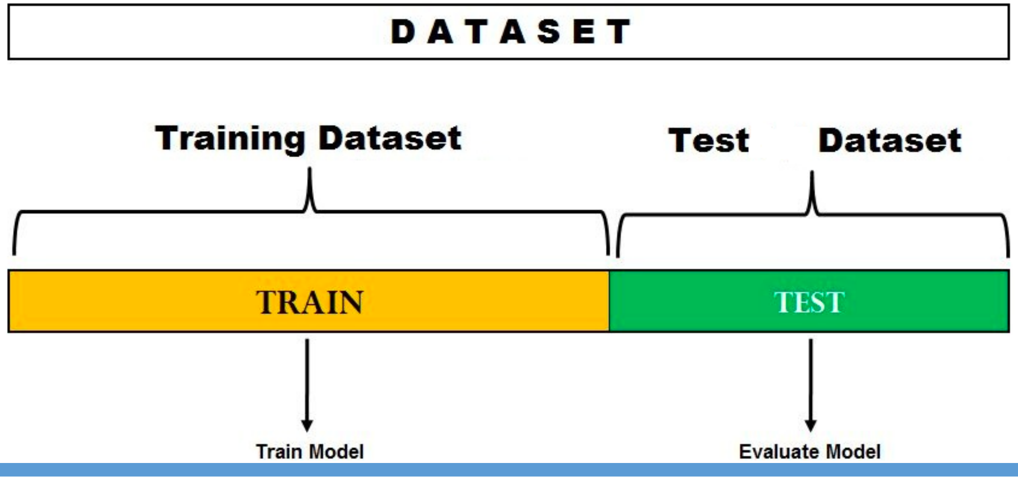
Purpose: To estimate performance of classifier on previously unseen data (test set)

- Holdout
- Random subsampling
- Cross validation (CV)
- Leave-one-out CV (LOOCV)

Detail: Introduction to Data Mining by Michael Steinbach, Pang-Ning Tan, and Vipin Kumar

Holdout Method

- The original data set is partitioned into two disjoint sets, called the Training set and Test set.
- Reserve k% for training and (100-k)% for testing
- A classification model is induced from the training set and its performance is evaluated on the test set.



Holdout Method’s Limitations

- Fewer data are available for training because a portion of the dataset is withheld for testing
- The model may be highly dependent on the composition of the training and test sets. The smaller training size, the larger the variance of the model and too large the training size, less reliable of the testing accuracy.
- Since it is a single train-and-test experiment, the holdout estimate of error rate will be misleading if we happen to get an “unfortunate” split

Random Subsampling Method

- The holdout method is repeated several times (k iterations) to improve classifier's performance.
- For each iteration, a fixed no. of observations is selected randomly without replacement and is kept aside as test set. The rest data is considered as training set.
- The overall accuracy is given by,

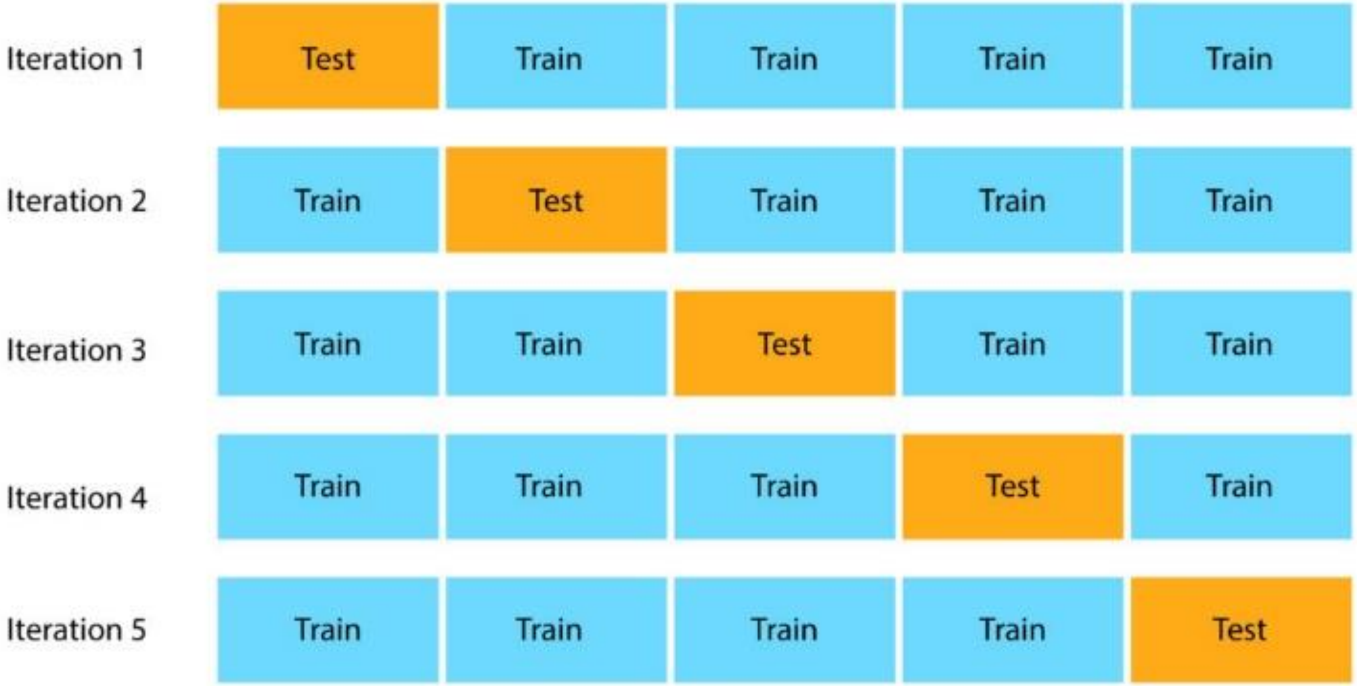
$$acc_{sub} = \frac{\sum_{i=1}^{k} acc_i}{k}$$

- Some observations may not be used for either training or testing.
- The number of iterations is not fixed.

Cross Validation Technique

1. Partition the dataset into k disjoint subsets/folds
2. For each fold in your dataset
   1. Take this fold as the test set
   2. Take the remaining k-1 folds as the training set
   3. Build your model on the training set and evaluate it on the test set
   4. Retain the evaluation score
3. Average your k recorded evaluation scores (also called the cross-validation scores) that will serve as your performance metric for the model.

5-folds Cross Validation (CV) - Example



Iteration 1	Test	Train	Train	Train	Train
Iteration 2	Train	Test	Train	Train	Train
Iteration 3	Train	Train	Test	Train	Train
Iteration 4	Train	Train	Train	Test	Train
Iteration 5	Train	Train	Train	Train	Test

Variations on Cross Validation

- Repeated cross-validation
  - Repeats cross-validation a number of times
  - Provides a way to improve the estimated performance of a machine learning model

  Implementation: see from here

- Stratified cross-validation
  - Guarantees the same percentage of class labels in training and test
  - Important when classes are imbalanced and the sample is small

  Implementation: see from here

Leave-one-out Cross Validation (LOOCV)

- A special case of the k-fold cross validation method when k=N, the size of the data set.
- Each test set contains only one record
- Utilizes as much data as possible for training
- It is computationally expensive
- But it results in a reliable and unbiased estimate of model performance

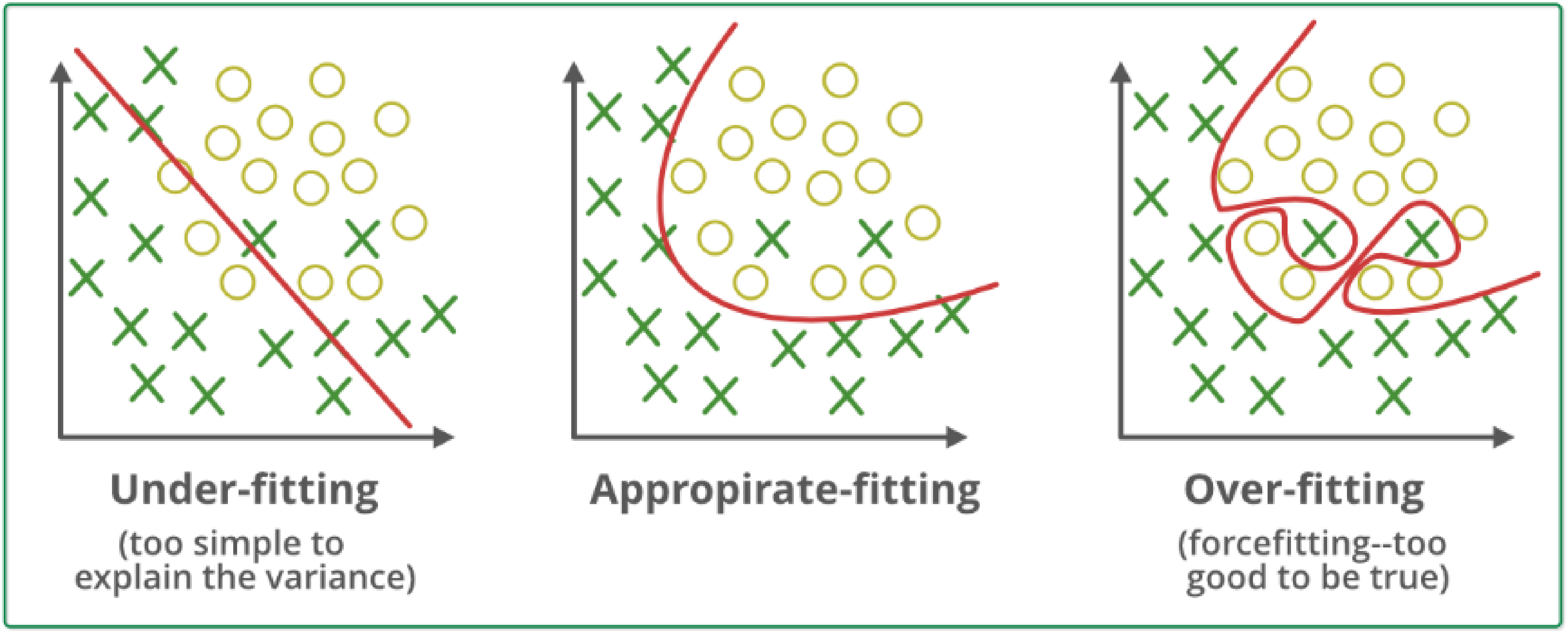
What is Generalization?

- In machine learning, generalization usually refers to the ability of a model to be effective across a range of unseen data.
- The goal of a good machine learning model is to generalize well from the training data to any data from the problem domain.
- There is a terminology used in machine learning when we talk about how well a machine learning model learns and generalizes to new data, namely overfitting and underfitting.
- Overfitting and underfitting are the two biggest causes for poor performance of machine learning algorithms.

Overfitting and Underfitting

A model is good if it neither Underfits or Overfits

Overfitting and Underfitting (Cont.)


Under-fitting
(too simple to explain the variance)
Appropirate-fitting
Over-fitting
(forcefitting--too good to be true)

Understanding Bias and Variance

- The prediction error for any machine learning model can be broken down into three parts as follows:

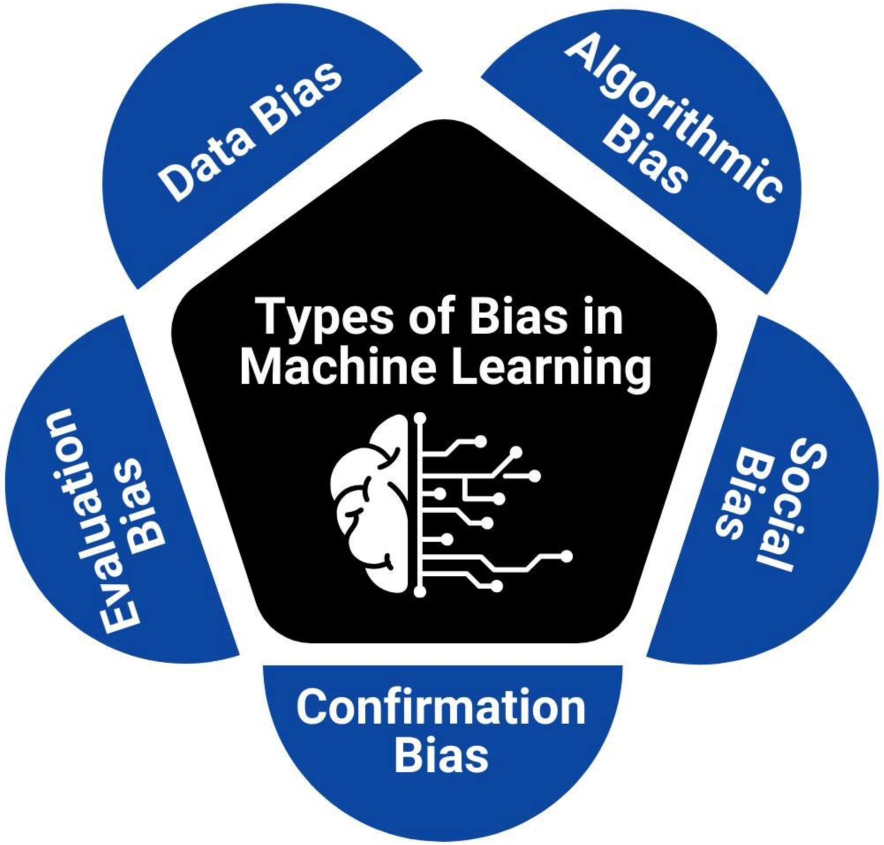
$$Error = \ Bias^2 + Variance + Irreducible\ Error$$

- Irreducible errors are errors which will always be present in a model, because of unknown variables, and can’t be reduced by creating good models or regardless of what algorithm is used.
- Reducible errors (Bias and Variance) are those errors whose values can be further reduced to improve a model.

Bias Error

- The bias error is an error from erroneous assumptions in the learning algorithm. **It is the difference between the actual and predicted values.**
- High bias can cause an algorithm to miss the relevant relations between features and target outputs (underfitting). It means it does not learn the training data very well.
- Examples of low-bias machine learning algorithms include: Decision Trees, k-Nearest Neighbors and Support Vector Machines.
- Examples of high-bias machine learning algorithms include: Linear Regression, Linear Discriminant Analysis and Logistic Regression.

HW: Types of Bias



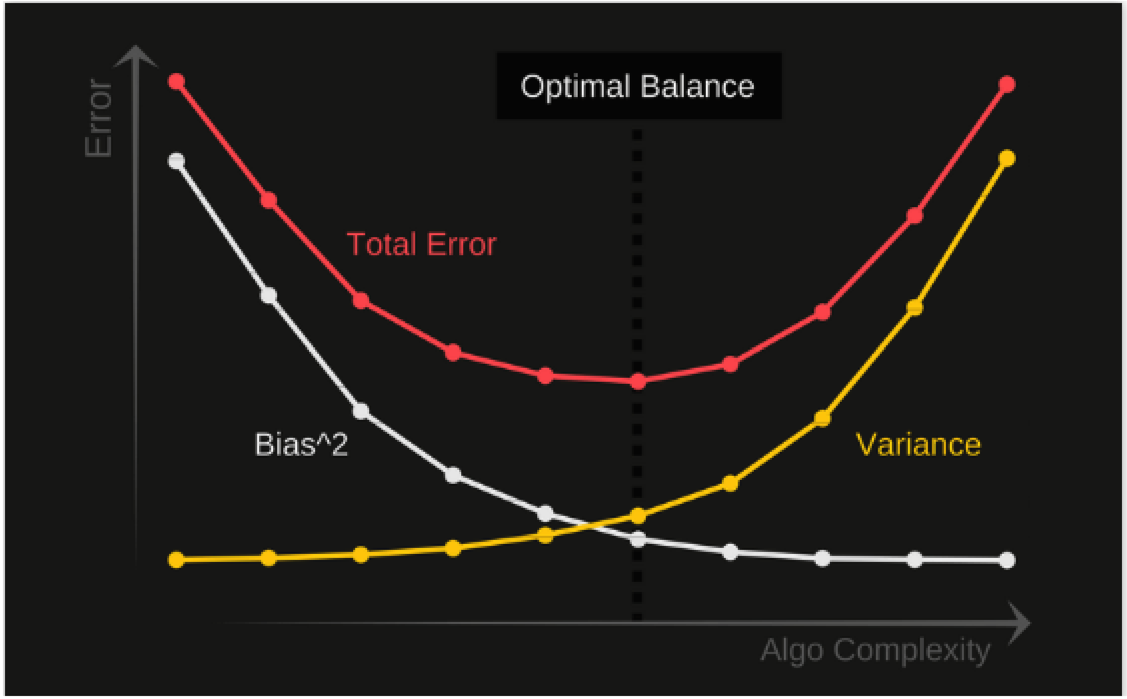
Variance Error

- The variance is an error from sensitivity to small fluctuations in the training set i.e. the noise as well.
- High variance can cause an algorithm to learn too much from the training data, so much so, that it is unable to predict new (testing) data accurately. It is called overfitting.
- Examples of low-variance machine learning algorithms include: Linear Regression, Linear Discriminant Analysis and Logistic Regression.
- Examples of high-variance machine learning algorithms include: Decision Trees, k-Nearest Neighbors and Support Vector Machines.

The Bias-Variance Tradeoff

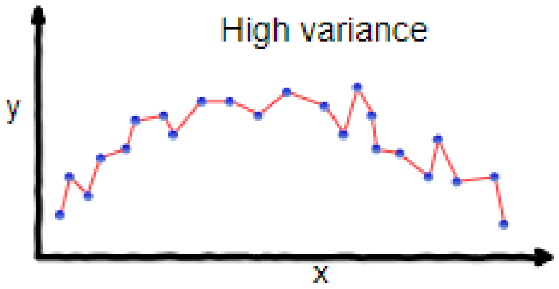
- The goal of any supervised machine learning algorithm is to achieve low bias and low variance. In turn, the algorithm should achieve good prediction performance.
- Bias–variance tradeoff is tension between the error introduced by the bias and the variance.
- Bias-variance trade-off refers to the property of a machine learning model such that model with a lower bias in parameter estimation have a higher variance of the parameter estimates across samples, and vice versa.
- To build a good model, we need to find a good balance between bias and variance such that it minimizes the total error.

The Bias-Variance Tradeoff (Cont.)

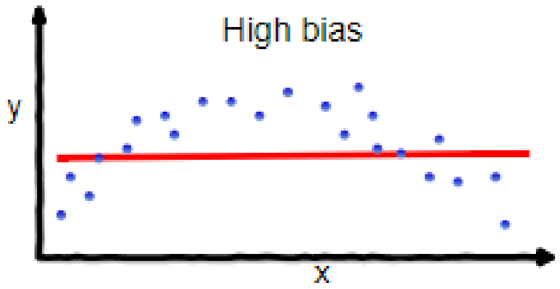


- An optimal balance of bias and variance would never overfit or underfit the model.

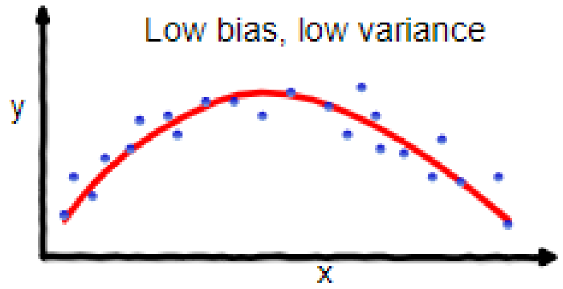
Overfitting and Underfitting (Cont.)



overfitting



underfitting



Good balance

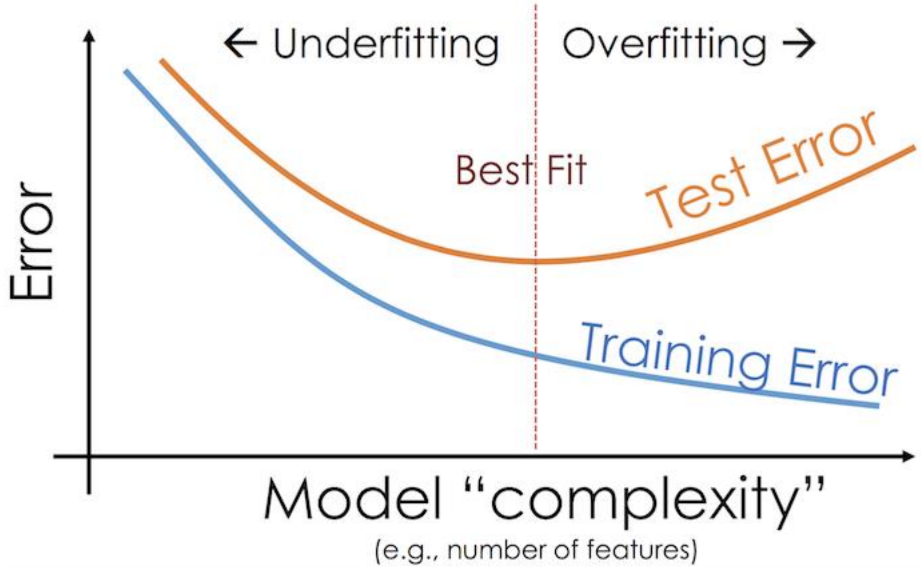
Underfitting

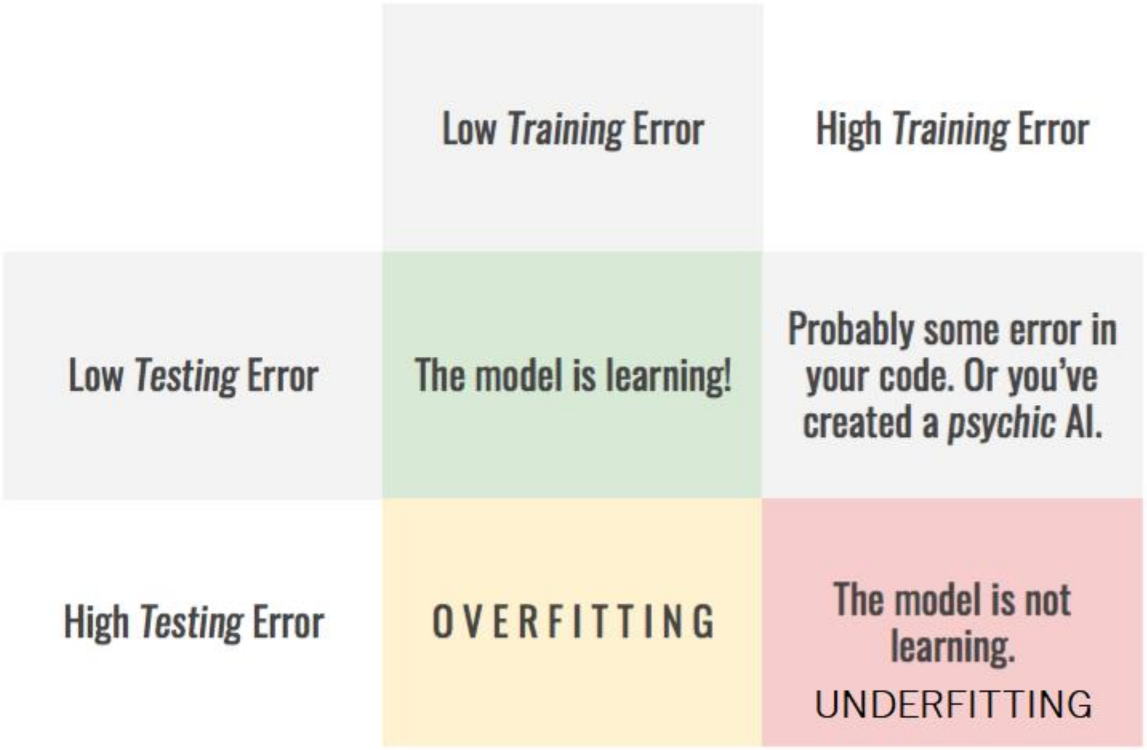
- Underfitting refers to a model that can neither fit the training data nor generalize to new data.
- Reasons for Underfitting:
  - when a model unable to capture the underlying pattern of the data.
  - when we have very less amount of data to build an accurate model.
  - When the model is too simple, e.g., when we try to build a linear model with a nonlinear data.
- These models usually have high bias and low variance.

Overfitting

- Overfitting refers to a model that fits the training data too well but generalizes poor to new data.
- Reasons for Overfitting:
  - when training data contains noise and the model captures the noise along with the underlying pattern in data.
  - when the model is too complex.
  - when the model trains for too long on a single sample set of data.
  - when you have less than enough training data, the model tries to memorize every single data point and fails to capture the general trends in the data.
- These models have low bias and high variance.

How to detect overfitting and underfitting?





	Low *Training* Error	High *Training* Error
Low *Testing* Error	The model is learning!	Probably some error in your code. Or you've created a *psychic* AI.
High *Testing* Error	OVERFITTING	The model is not learning. UNDERFITTING

Summary

- A model with a high bias error underfits data and makes very simplistic assumptions on it
- A model with a high variance error overfits the data and learns too much from it
- A good model is where both Bias and Variance errors are balanced

HW: How to avoid underfitting?

Find it by yourself

How to avoid underfitting?

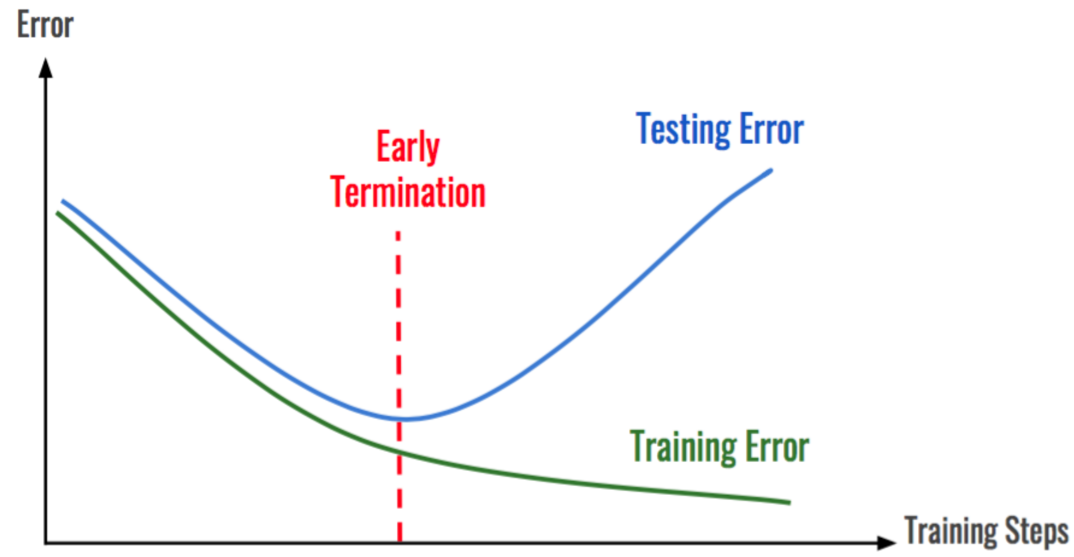
The main options to fix this problem:

- Collect more data
- Select a more powerful model, with more parameters
- Feature engineering (create new or modify features)
- Remove regularization
- Tune model’s hyper-parameters

How to avoid overfitting?

The commonly used solutions are:

- Train with more diverse data
- Remove irrelevant features or noise in the training data
- Reduce features (select important features)
- Cross-Validation
- Early Stopping
- Use ensemble methods
- Regularization (constraining a model to make it simpler)



Some Learning Materials

Bias and Variance in Machine Learning – A Fantastic Guide for Beginners!

Gentle Introduction to the Bias-Variance Trade-Off in Machine Learning

What Are Overfitting and Underfitting in Machine Learning?

Overfitting and Underfitting With Machine Learning Algorithms