Introduction to Machine Learning

William Brasic

The University of Arizona

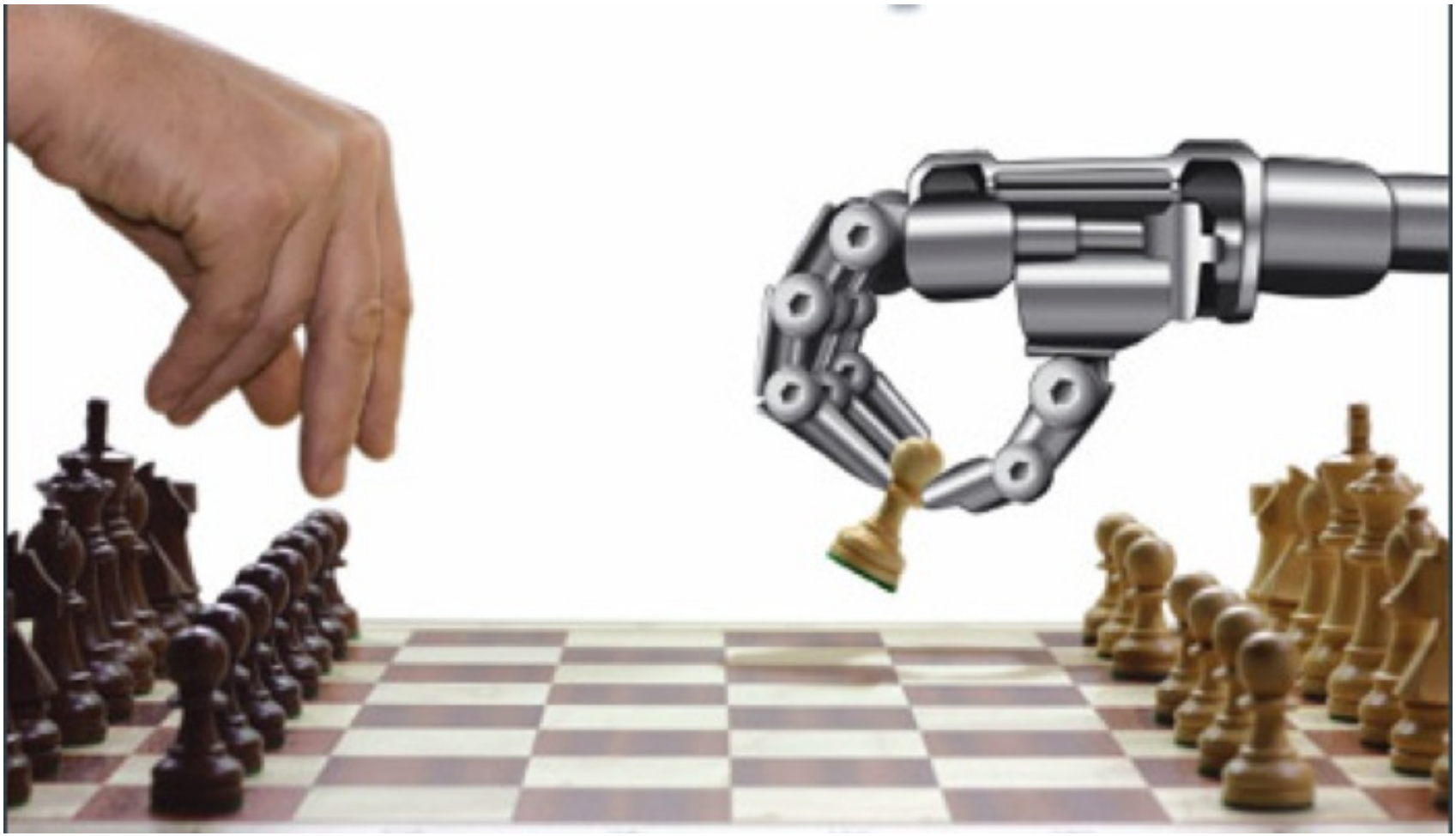
Artificial Intelligence (AI)

Definition 1: Artificial Intelligence (AI)

Artificial Intelligence (AI) is the simulation of human intelligence processes by machines, especially computer systems.

- These processes include learning (the acquisition of information and rules for using the information), reasoning (using rules to reach approximate or definite conclusions), and self-correction.

Artificial Intelligence (AI)



Artificial Intelligence (AI)

Algorithm

Definition 2: Algorithm

An algorithm is a set of instructions that maps inputs to outputs.

- One example is least squares.

Machine Learning

Definition 3: Machine Learning

Machine Learning (ML) is a field of artificial intelligence that focuses on developing algorithms and statistical models that enable computers to learn and make predictions or decisions without being explicitly programmed.

- ML involves training models on data to identify patterns, make inferences, and improve performance over time.
- ML only cares about prediction and doesn't concern itself with most issues in econometrics (e.g., endogeneity).

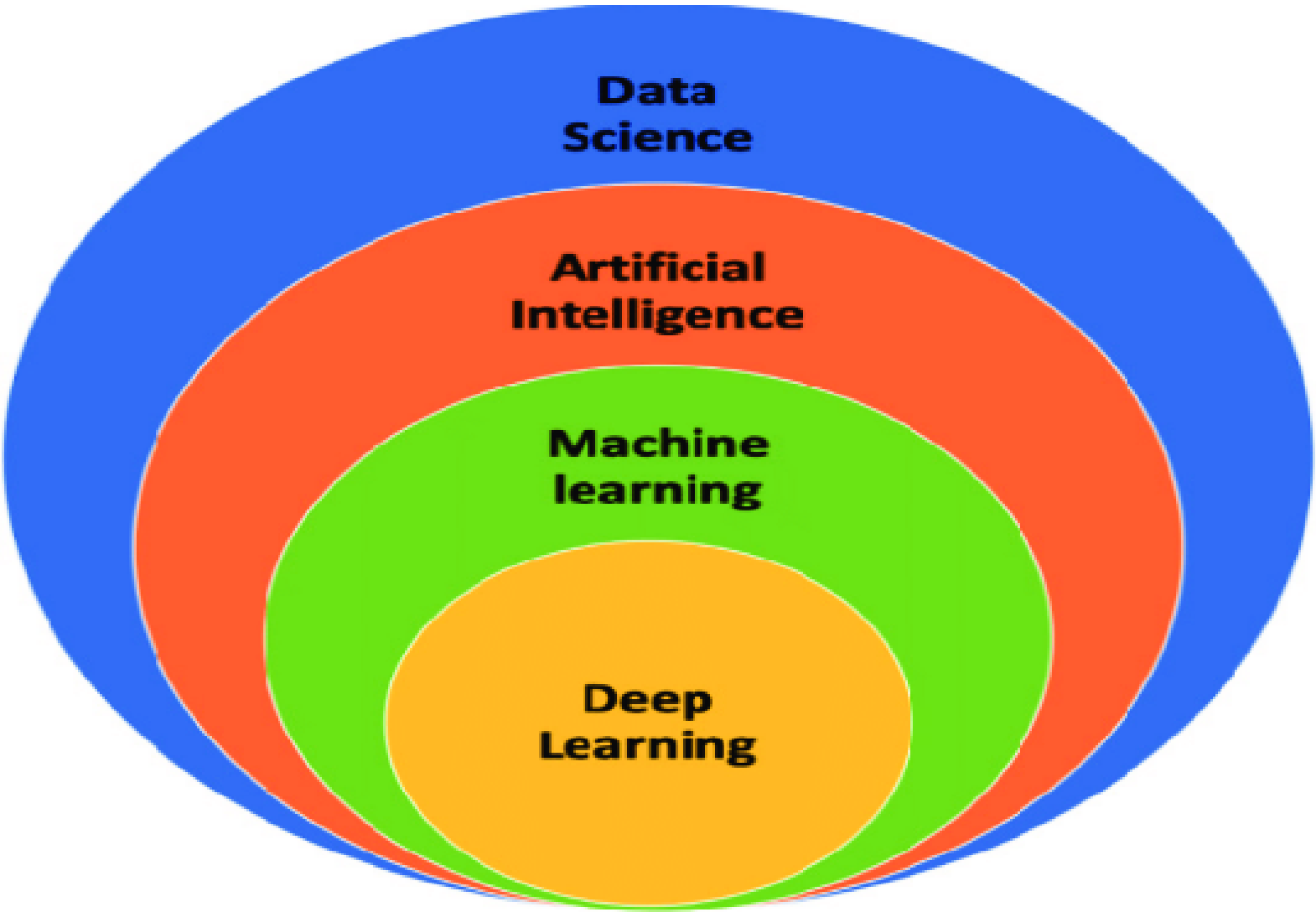
Machine Learning

Example 1: Machine Learning

Examples of machine learning include:

- Using linear regression to predict a person's income based on their education level.
- Using random forests to classify a patient has having cancer based on their individual characteristics.
- Using K-nearest neighbors (KNN) clustering to detect credit card fraud.
- Using reinforcement learning to construct a large language model (LLM) like ChatGPT.

Artificial Intelligence (AI)



Econometrics vs ML Terminologies

	Econometrics
X	Covariate Matrix
$\boldsymbol{y}$	Response/Outcome
$\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$	Sample
Function to Minimize	Objective Function
Data Analysis Goals	Consistency, Inference, Confidence Intervals
Practical vs Theoretical Concerns	Reluctant to use methods w/o theoretical justification
Computing	Important

Econometrics vs ML Terminologies

	ML
X	Input/Feature/Design Matrix
$\boldsymbol{y}$	Output
$\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$	Training Data
Function to Minimize	Loss Function
Data Analysis Goals	Prediction
Practical vs Theoretical Concerns	As long the method predicts well, we use it
Computing	Very Important

Supervised Learning

Definition 4: Supervised Learning

Supervised learning is a branch of machine learning where the algorithm is trained on labeled data.

- The model learns from input-output pairs, where the input data is known and the output is labeled with the correct answer.
- This is the most prevalent type of machine learning.
- Supervised learning consists of regression and classification.

Supervised Learning

Property 1: Supervised Learning

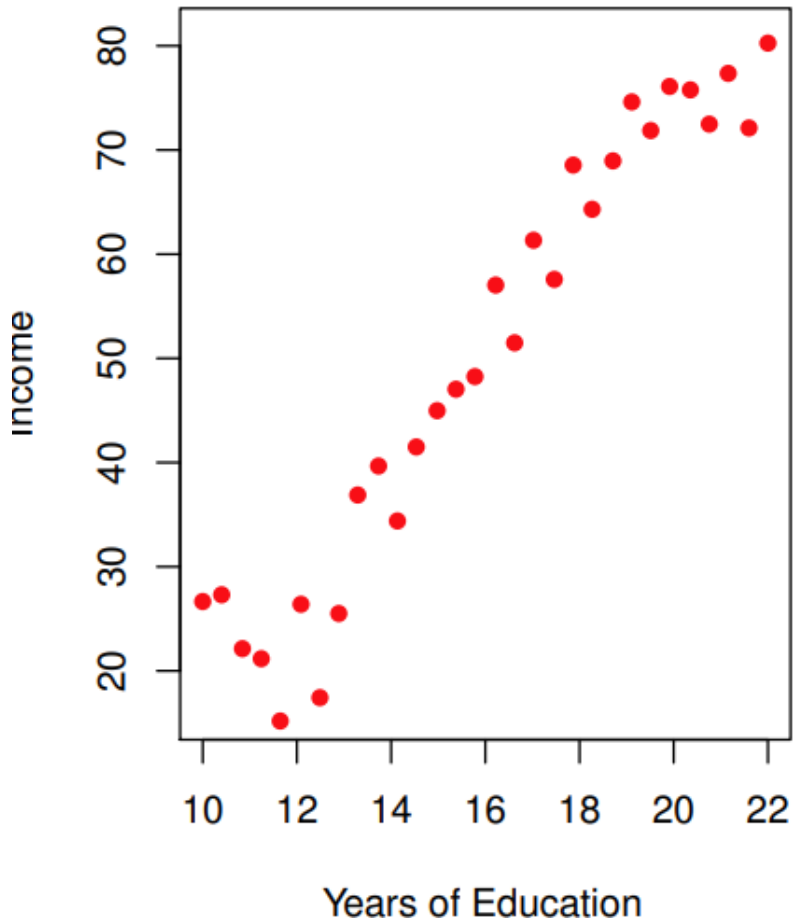
There are two types of supervised learning:

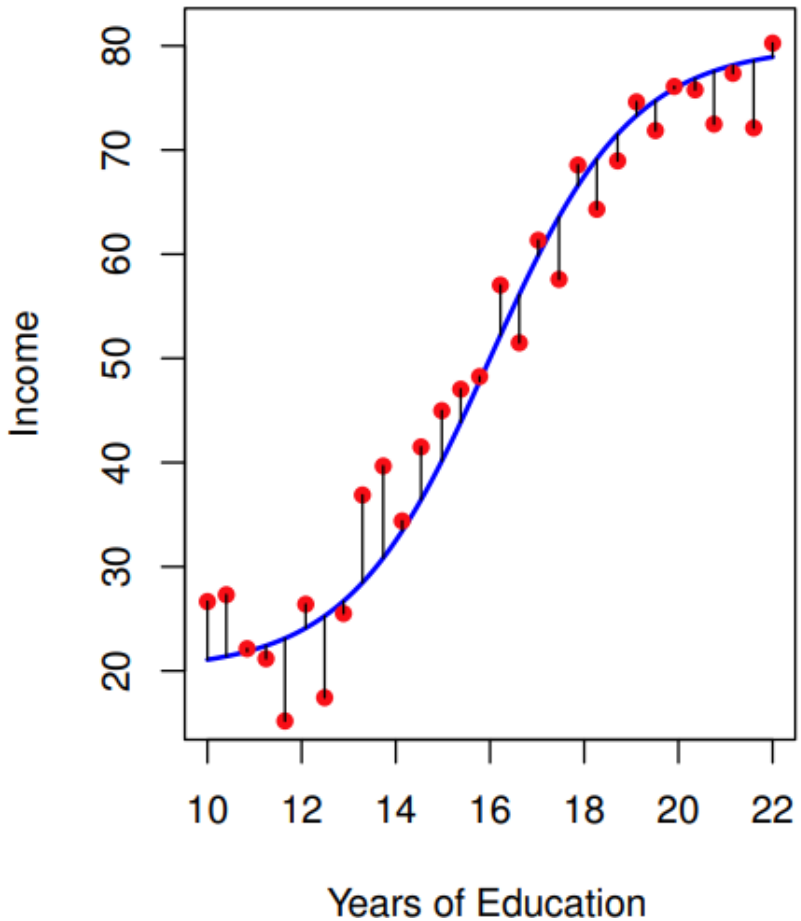
- Regression so $y_i = f(\boldsymbol{x}_i) + \epsilon_i$ is continuous
  - OLS
- Classification so $y_i = f(\boldsymbol{x}_i) + \epsilon_i$ is a discrete label
  - Probit/Logit

where ϵ_i is the zero mean error assumed to be uncorrelated with $\boldsymbol{x}_i$.

- In either case, the goal is to learn the best possible $\widehat{f}$.

Regression


income
20
30
40
50
60
70
80
10
12
14
16
18
20
22
Years of Education


Income
20
30
40
50
60
70
80
10
12
14
16
18
20
22
Years of Education

Classification

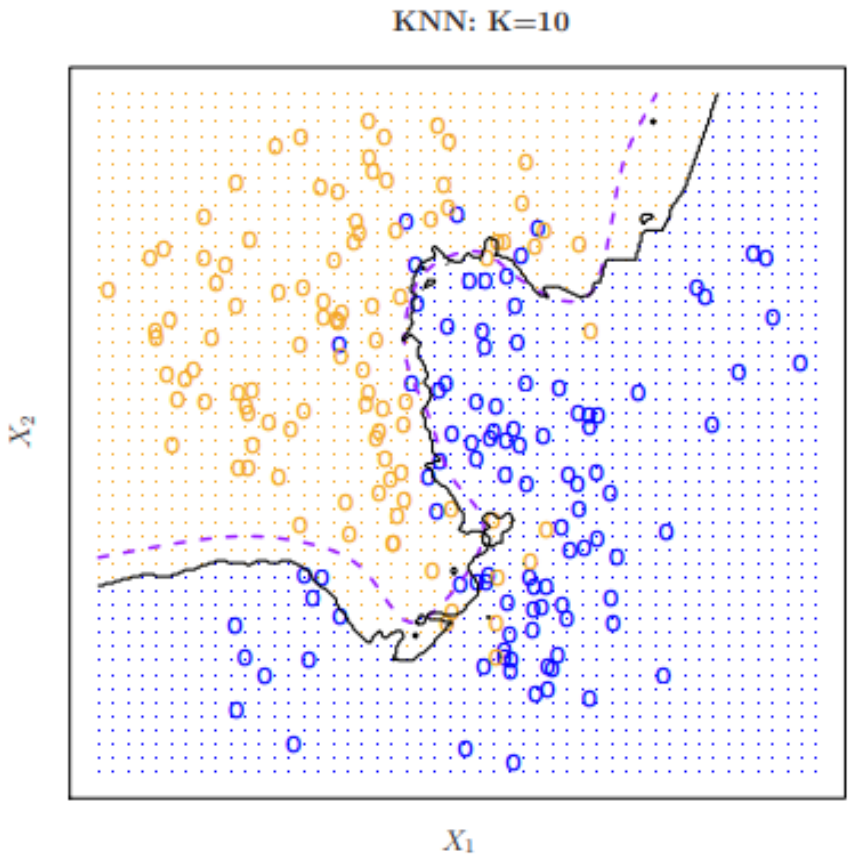


FIGURE 2.15. *The black curve indicates the KNN decision boundary on the data from Figure 2.13, using $K = 10$. The Bayes decision boundary is shown as a purple dashed line. The KNN and Bayes decision boundaries are very similar.*

How to Obtain Best $\widehat{f}$?

Question 1: How to Obtain Best $\widehat{f}$?

How do we obtain the best possible $\widehat{f}$?

How to Obtain Best $\widehat{f}$?

Question 1: How to Obtain Best $\widehat{f}$?

How do we obtain the best possible $\widehat{f}$?

Answer to Question 1

Let's think back to OLS. The goal was to minimize the empirical $SSR = \widehat{\boldsymbol{u}}'\widehat{\boldsymbol{u}}$. The estimators that minimize this are the same as the estimates that minimize $\frac{1}{n}SSR$ which is called the empirical mean squared error (MSE). Thus, a good place to start is by minimizing the empirical mean squared error in hopes that it will minimize the true MSE.

Expected Mean Squared Error (MSE)

Definition 5: Expected Mean Squared Error (MSE)

$$\mathbb{E}\left[(y_i - \widehat{y}_i)^2\right] = \mathbb{E}\left[\left(f(\boldsymbol{x}_i) - \widehat{f}(\boldsymbol{x}_i) + \epsilon_i\right)^2\right]$$

Expected Mean Squared Error (MSE)

Definition 5: Expected Mean Squared Error (MSE)

$$\begin{aligned}\mathbb{E}\left[(y_i - \widehat{y}_i)^2\right] &= \mathbb{E}\left[\left(f(\boldsymbol{x}_i) - \widehat{f}(\boldsymbol{x}_i) + \epsilon_i\right)^2\right] \\ &= \mathbb{E}\left[\left(f(\boldsymbol{x}_i) - \widehat{f}(\boldsymbol{x}_i)\right)^2\right] + 2\mathbb{E}\left[\left(f(\boldsymbol{x}_i) - \widehat{f}(\boldsymbol{x}_i)\right)\epsilon_i\right] \\ &\quad + \mathbb{E}\left[\epsilon_i^2\right]\end{aligned}$$

Expected Mean Squared Error (MSE)

Definition 5: Expected Mean Squared Error (MSE)

$$\begin{aligned}
\mathbb{E}\left[(y_i - \widehat{y}_i)^2\right] &= \mathbb{E}\left[\left(f(\boldsymbol{x}_i) - \widehat{f}(\boldsymbol{x}_i) + \epsilon_i\right)^2\right] \\
&= \mathbb{E}\left[\left(f(\boldsymbol{x}_i) - \widehat{f}(\boldsymbol{x}_i)\right)^2\right] + 2\mathbb{E}\left[\left(f(\boldsymbol{x}_i) - \widehat{f}(\boldsymbol{x}_i)\right)\epsilon_i\right] \\
&\quad + \mathbb{E}\left[\epsilon_i^2\right] \\
&= \mathbb{E}\left[\left(f(\boldsymbol{x}_i) - \widehat{f}(\boldsymbol{x}_i)\right)^2\right] + \mathbb{V}\left[\epsilon_i\right].
\end{aligned}$$

- We can't do much about $\mathbb{V}\left[\epsilon_i\right]$.
- So, must reduce $\mathbb{E}\left[\left(f(\boldsymbol{x}_i) - \widehat{f}(\boldsymbol{x}_i)\right)^2\right]$.

Empirical Mean Squared Error (MSE)

Definition 6: Empirical Mean Squared Error (MSE)

The empirical MSE is given by

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{f}(\boldsymbol{x}_i)\right)^2.$$

- By minimizing this, we hope to minimize the expected MSE. If so, our model *generalizes* well.
- What does this represent when y_i is binary?

Loss Function

Definition 7: Loss Function

The loss function for a given ML algorithms is the function the algorithm attempts to minimize in order to obtain $\widehat{f}$.

- Examples include:
  1. SSR
  2. MSE
  3. Classification Error Rate
- Before we called this the objective function

How to estimate f?

Question 2: How to Estimate f?

How can we estimate f?

Answer to Question 2

- Parametric Methods
- Non-parametric Methods

Parametric Methods

Definition 8: Parametric Methods

Parametric methods make distinct and often strong assumptions on the functional form of f.

- These assumptions could result in $\widehat{f}$ being very different from f.
- Least squares where we make the linearity assumption.

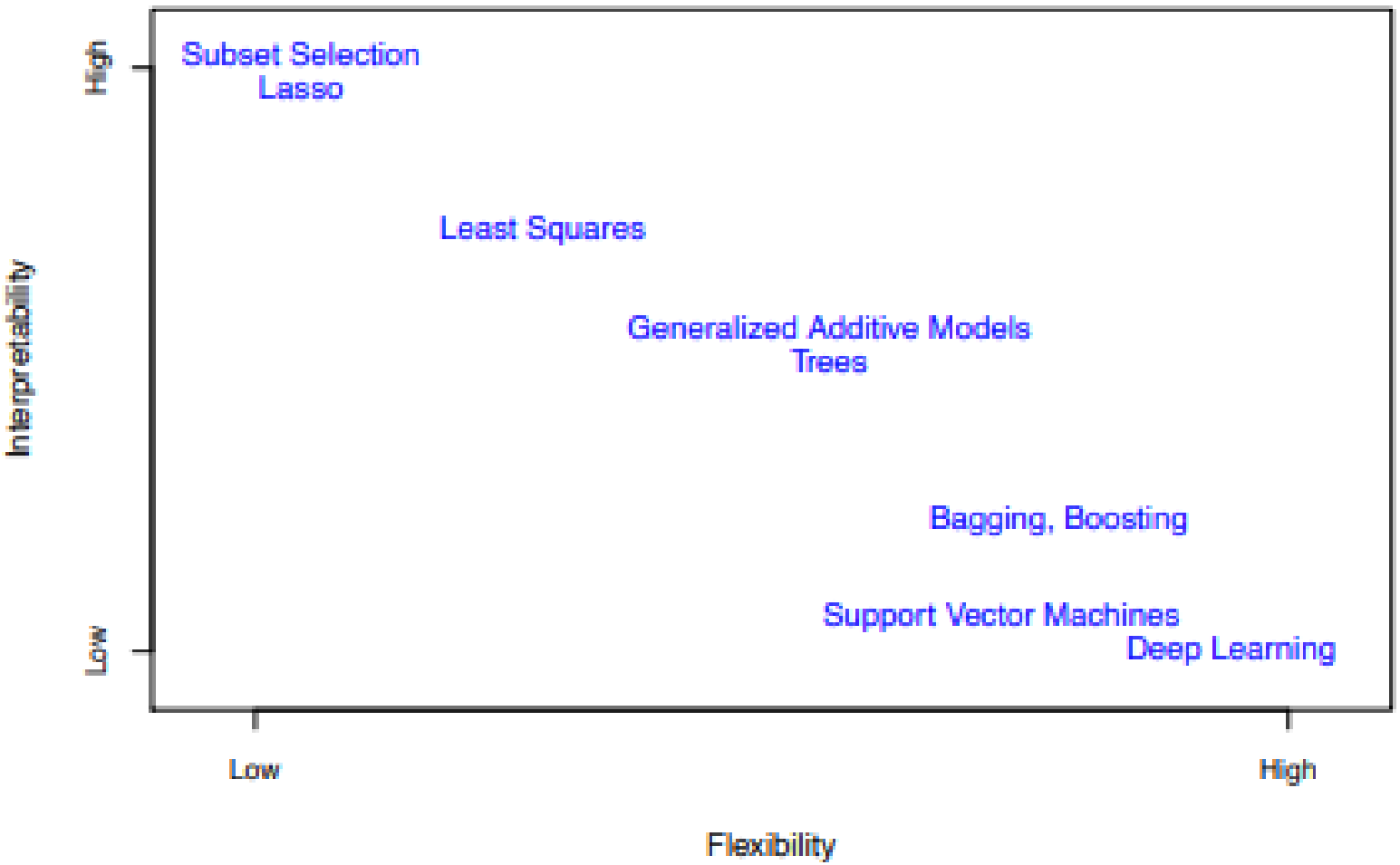
Non-Parametric Methods

Definition 9: Non-Parametric Methods

Non-parametric methods avoid making assumptions on the functional form of f.

- Could suffer from overfitting.
- Example is random forests where no parameters are estimated and the predictor $\widehat{f}$ is entirely determined by the data.

Accuracy vs. Interpretability



Training Set

Definition 10: Training Set

A training set is a subset of the original dataset used to train a machine learning model.

- The model learns to understand the patterns and relationships in the data.
- We'll often use 60% - 80% of the original data for training.

Validation Set

Definition 11: Test Set

A validation set is a subset of the dataset used to evaluate the performance of a machine learning model during training.

- It helps in tuning the models' hyperparameters and preventing overfitting by providing an independent set of data for testing.

Test Set

Definition 14: Test Set

A test set is a subset of the original dataset used to evaluate the final performance of a trained machine learning model.

- It is independent of the training and validation sets and provides an unbiased assessment of how well the model generalizes to new, unseen data.
- **We only use the testing set after selecting a final model. We cannot change our model or adjust its hyperparmaters using the testing set.**

ML Procedure

Definition 15: ML Procedure

1. Split data into 70% training, 15% validation, and 15% testing.

ML Procedure

Definition 16: ML Procedure

1. Split data into 70% training, 15% validation, and 15% testing.
2. Evaluate multiple models using the training set.

ML Procedure

Definition 17: ML Procedure

1. Split data into 70% training, 15% validation, and 15% testing.
2. Evaluate multiple models using the training set.
3. Tweak the estimated models' hyperparameters using the validation set.

ML Procedure

Definition 18: ML Procedure

1. Split data into 70% training, 15% validation, and 15% testing.
2. Evaluate multiple models using the training set.
3. Tweak the estimated models' hyperparameters using the validation set.
4. Select the best model based on its validation set performance.

ML Procedure

Definition 19: ML Procedure

1. Split data into 70% training, 15% validation, and 15% testing.
2. Evaluate multiple models using the training set.
3. Tweak the estimated models' hyperparameters using the validation set.
4. Select the best model based on its validation set performance.
5. Evaluate the selected model on the testing set and future new data.

ML Procedure

Definition 20: ML Procedure

1. Split data into 70% training, 15% validation, and 15% testing.
2. Evaluate multiple models using the training set.
3. Tweak the estimated models' hyperparameters using the validation set.
4. Select the best model based on its validation set performance.
5. Evaluate the selected model on the testing set and future new data.

- The idea is that the model that performs best on the validation data should *generalize* best to new and unseen data.

Generalization

Definition 21: Generalization

Generalization is the ability of a machine learning model to perform well on new and unseen data.

- It refers to the model's capability to apply what it has learned from the training data to make accurate predictions on different datasets.
- An unbiased estimate of model's generalization ability is its performance on the testing set.

Overfitting

Definition 22: Overfitting

Overfitting is a modeling error that occurs when a machine learning model captures noise or random fluctuations in the training data instead of the underlying pattern.

- This leads to a model that performs well on training data, but poorly on new and unseen data.
- The model does not *generalize* well to unseen data.
- This leads to high *variance*.

Underfitting

Definition 23: Underfitting

Underfitting is a modeling error that occurs when a machine learning model is too simple to capture the underlying pattern in the data.

- This leads to poor performance on both training and new data.
- This leads to high *bias*.

Bias-Variance Tradeoff

Definition 24: Bias-Variance Tradeoff

The bias-variance tradeoff is the balance between two sources of error in a machine learning model: bias, which is the error due to overly simplistic assumptions in the learning algorithm, and variance, which is the error due to too much complexity in the model.

- High bias (underfitting) and high variance (overfitting) leads to poor *generalization*
- A goal in ML is to find a good balance between these two ideas.

Bias-Variance Decomposition

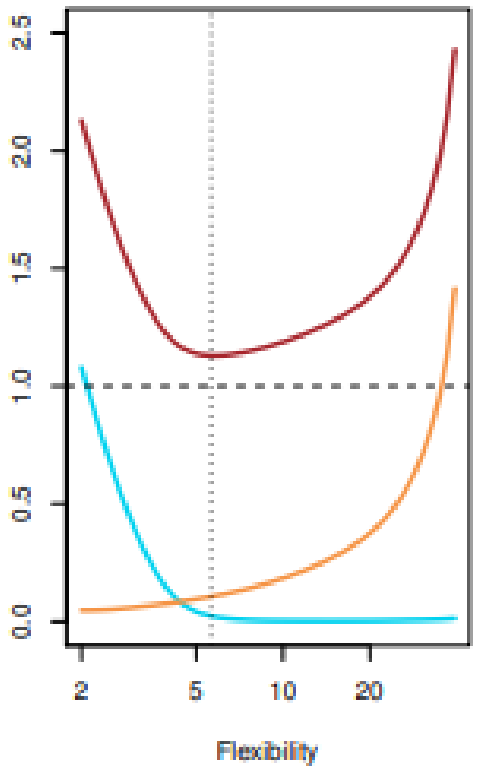
Definition 25: Bias-Variance Decomposition

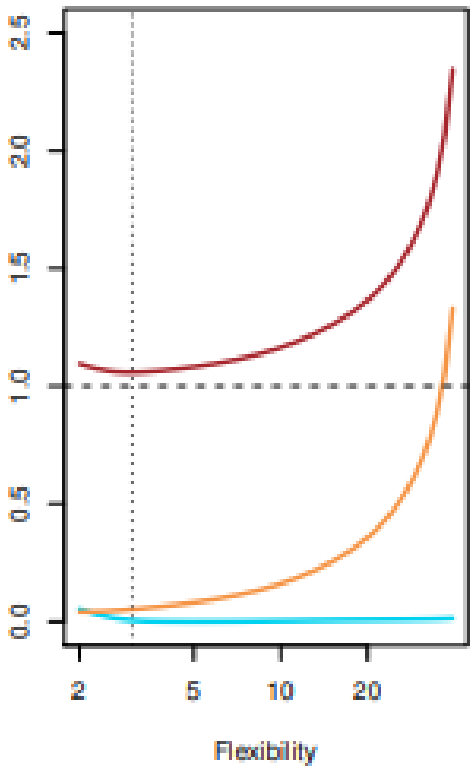
The bias-variance decomposition expresses the expected MSE as

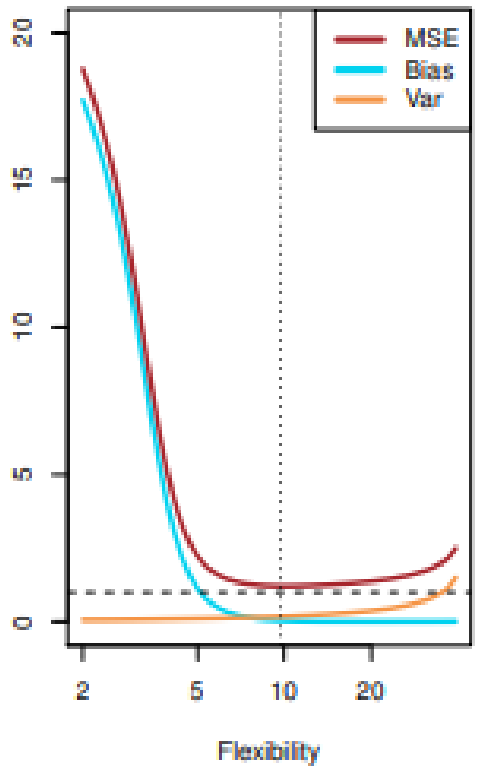
$$\begin{aligned}
&\mathbb{E}\left[\left(f(\boldsymbol{x}_i) - \widehat{f}(\boldsymbol{x}_i)\right)^2\right] + \mathbb{V}\left[\epsilon_i\right] \\
&= \mathbb{V}\left[\widehat{f}(\boldsymbol{x}_i)\right] + \left[\text{Bias}\left[\widehat{f}(\boldsymbol{x}_i)\right]\right]^2 + \mathbb{V}\left[\epsilon_i\right].
\end{aligned}$$

- Thus, to get the best $\widehat{f}$ possible, we must reduce bias and variance of our estimator, but we can't do both *simultaneously*

Bias-Variance Tradeoff







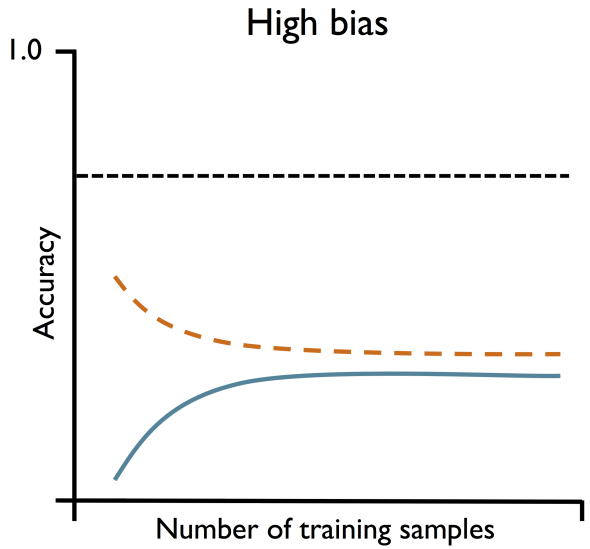
Learning Curve

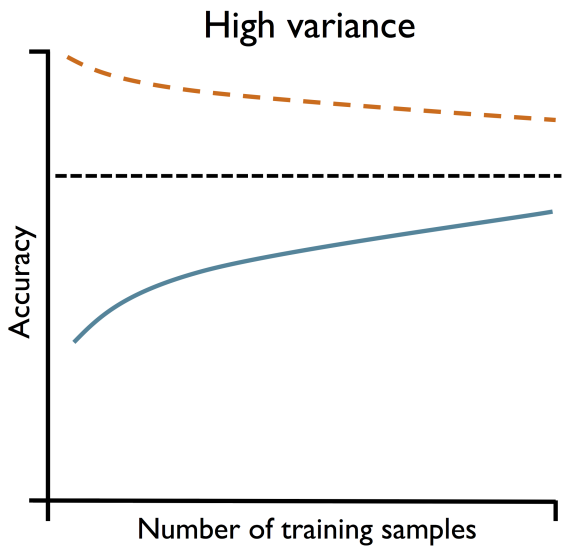
Definition 26: Learning Curve

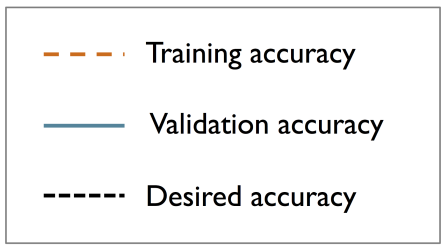
A learning curve depicts the training and validation accuracy as a function of the number of training samples seen.

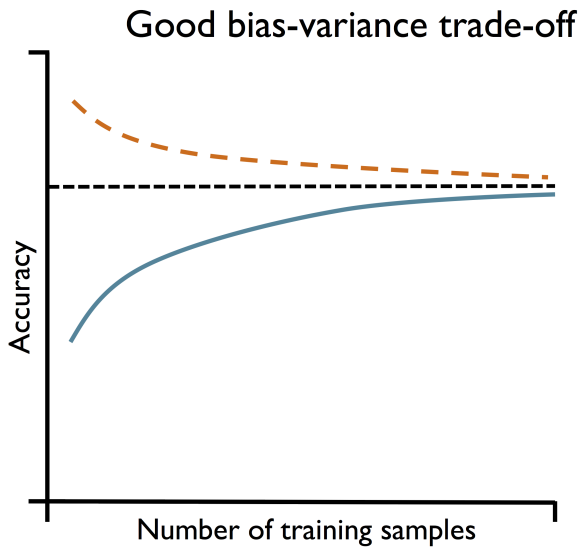
- Helps to visually analyze the bias-variance tradeoff.
- We want these curves to converge to each other at a high accuracy level as the number of seen data samples.
- Extremely prevalent in deep learning.

Learning Curve









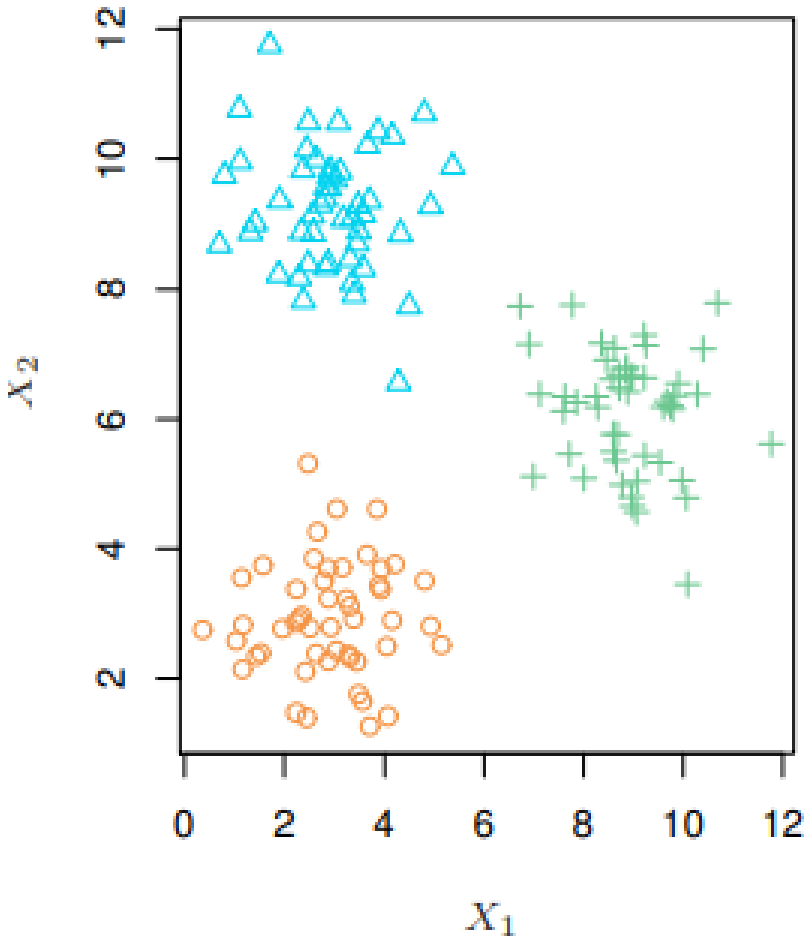
Unsupervised Learning

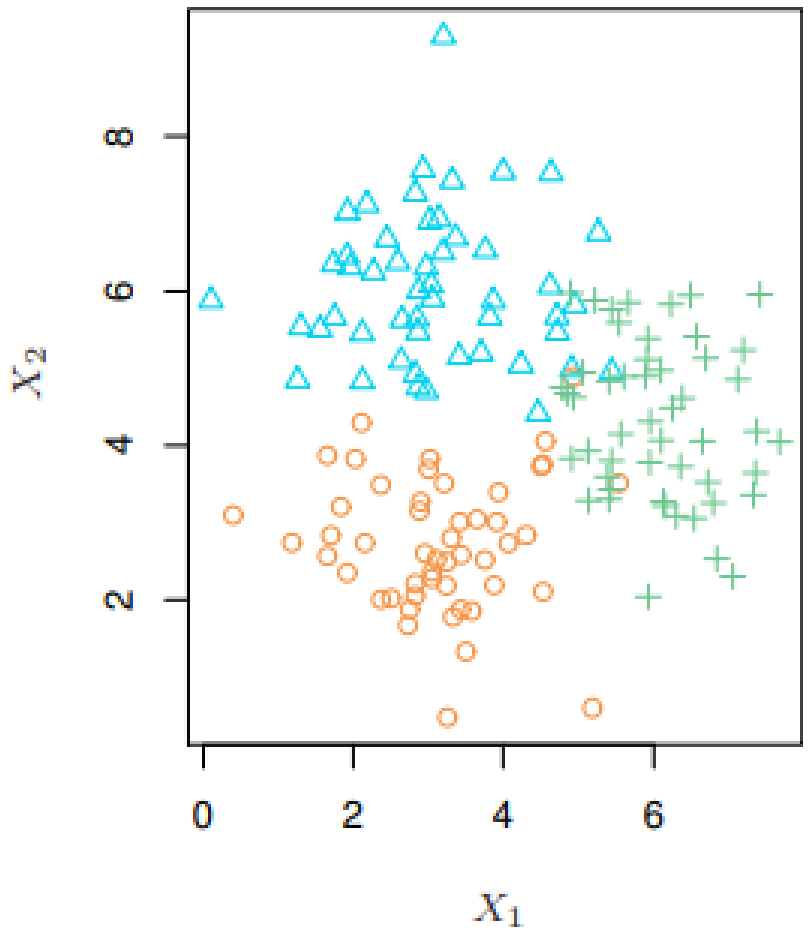
Definition 27: Unsupervised Learning

Unsupervised learning is a branch of machine learning where the algorithm is trained on unlabeled data.

- The model tries to identify patterns and relationships within the data without any prior knowledge of the correct output.

Unsupervised Learning





- Example is fraud detection

Econometrics vs ML Terminologies

	Supervised Learning	**Unsupervised Learning**
y	Observed	Unobserved
Goal	Predict y given X	Find patterns in X
Examples	Regression, Classification	Clustering, Dimensionality Reduction

Reinforcement Learning

Definition 28: Reinforcement Learning

Reinforcement learning is a branch of machine learning where an agent learns to make decisions by performing certain actions and receiving rewards or penalties.

- The goal is to maximize the cumulative reward over time by learning the best strategies.
- ChatGPT is built on reinforcement learning with human feedback.

Evaluation Metrics

Definition 29: Evaluation Metrics

An evaluation metric is method to evaluate a model's performance.

- We use different evaluation metrics depending on whether we are doing regression, classification, clustering, etc.

Regression Evaluation Metrics

Definition 30: Empirical Root Mean Squared Error (RMSE)

The empirical root mean squared error (RMSE) is given by

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[y_i - \widehat{f}(\boldsymbol{x}_i)\right]}.$$

- This is the main evaluation metric used for regression.
- Taking the square root just makes for an easier interpretation relative to the standard empirical MSE.

Regression Evaluation Metrics

Definition 31: Empirical Mean Absolute Error (MAE)

The empirical mean absolute error (MAE) is given by

$$\frac{1}{n}\sum_{i=1}^{n}\left|y_i - \widehat{f}(\boldsymbol{x}_i)\right|.$$

Regression Evaluation Metrics

Definition 32: Adjust R-Squared

The adjusted R-Squared is given by

$$\widetilde{R}^2 = 1 - \frac{SSR/(n-k)}{SST/(n-1)}.$$

Classification Evaluation Metrics

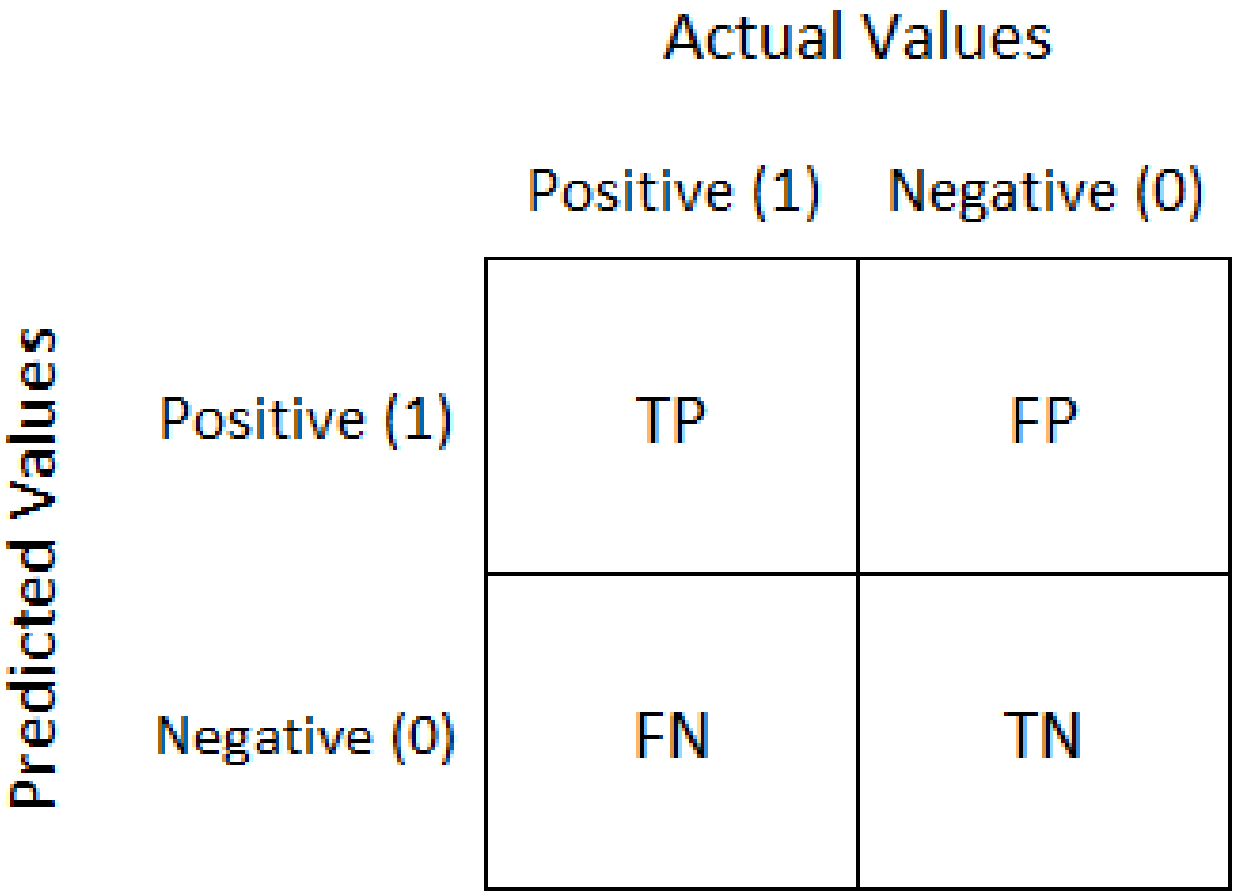
Definition 33: Success Rate

The success rate is given by

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\left(y_i = \widehat{f}(\boldsymbol{x}_i)\right).$$

- Represents the fraction of correctly classified observations.
- This is main evaluation metric used for classification.

Classification Evaluation Metrics



Feature Engineering

Definition 34: Feature Engineering

Feature engineering is the process of creating new features or modifying existing ones to improve the performance of a machine learning model.

- Involves transforming raw data into meaningful features that better represent the underlying problem to the model

Feature Engineering

Example 2: Feature Engineering

- Creating interaction, squared, cubed, etc. terms
- Normalization and scaling (critical for almost every ML algo outside of trees and forests)
- Encoding categorical variables (one-hot encoding)
- Handling missing values

Feature Selection

Definition 35: Feature Selection

Feature selection is the process of selecting a subset of relevant features for use in model construction.

- Aims to improve model performance by reducing overfitting, enhancing generalization, and decreasing computational cost.

Feature Selection

Example 3: Feature Selection

- Filter methods
  - Correlation coefficient
- Wrapper methods
  - Recursive feature elimination
- Embedded methods
  - Lasso regularization

Hyperparameters

Definition 36: Hyperparameters

Hyperparameters are parameters that are set before the training process begins.

- They control the learning process and the model structure.
- Examples include:
  - Polynomial degree in linear regression
  - Trees in a forest
  - Layers in a neural network
  - Regularization term in lasso/ridge regression

Hyperparameter Tuning

Definition 37: Hyperparameter Tuning

Hyperparameter tuning is the process by which we tweak the hyperparmeters during the validation stage to enhance model performance.

- Proper tuning can significantly improve model performance.
- Incorrect hyperparameters can lead to overfitting or underfitting.

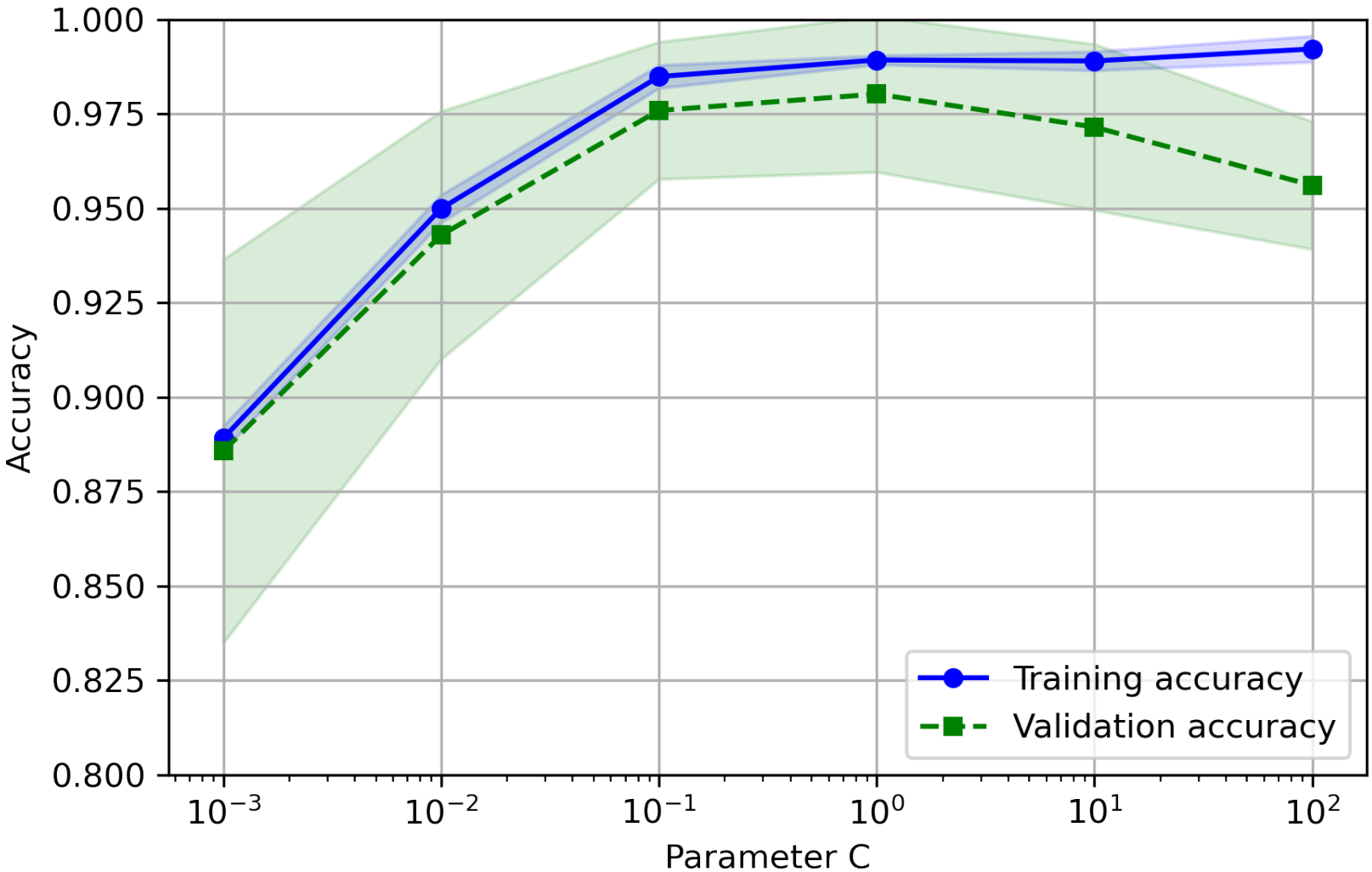
Validation Curve

Definition 38: Validation Curve

A validation curve is used to pick the optimal hyperparameter from a set of contending hyperparameters.

- Useful in addressing over- and underfitting.

Validation Curve



Grid Search

Definition 39: Grid Search

Grid search exhaustively searches through a specified parameter grid.

- Simple, but often computationally expensive.

Grid Search

Example 4: Grid Search

Suppose we are tuning the learning rate and number of hidden layers in a neural network. We define a possible *grid* of hyperparameters as

- Learning Rate: $[0.001,\ 0.01,\ 0.1]$
- Hidden Layers: $[1,\ 2,\ 3]$.

Using the validation set, grid search will evaluate the model on each of the nine possible combinations and we will select the combination with the best validation performance.

Random Search

Definition 40: Random Search

Random search randomly samples hyperparameters from a distribution.

- Simple and often less computationally expensive relative to grid search.

Random Search

Example 5: Random Search

Suppose we are tuning the learning rate and number of hidden layers in a neural network. We define a possible *grid* of hyperparameters as

- Learning Rate: $[0.001,\ 0.01,\ 0.1]$
- Hidden Layers: $[1,\ 2,\ 3]$.

Using the validation set, random search will evaluate the model on a certain number of *random* hyperparameter combinations and we will select the combination with the best validation performance.

Thank You!