Introduction to Machine Learning

By Van Dinh Tran

January 12, 2025

Introduction to ML

- What is learning? Machine learning (ML)? and When ML?
- Traditional program vs ML
- ML vs Data Science (DS) vs Artificial intelligence (AI)
- Types of ML
- ML model evaluation
  - Data splitting
  - Evaluation metrics
- Bias and variance
- Underfitting vs overfiting

What is learning?

- Learning is the process of acquiring new understanding, knowledge, behaviors, skills, values, attitudes, and preferences.
- Humans, animals, some plants, and **some machines** have ability to learn.

What is Machine learning?

Herbert Simon

Learning is any process by which a system improves performance from experience.

Tom Mitchell

A computer program is said to "learn" from experience E with respect to some task T and performance measure P if its performance on T as measured by P improves with experience E.

Arthur Samuel

Machine Learning is the field of study that gives the computer the ability to learn without being explicitly programmed.

When Machine learning?

- **Complex rules and functions:**

  You cannot code the rules due to their complexity, i.e., they are hard to define explicitly through traditional programming.

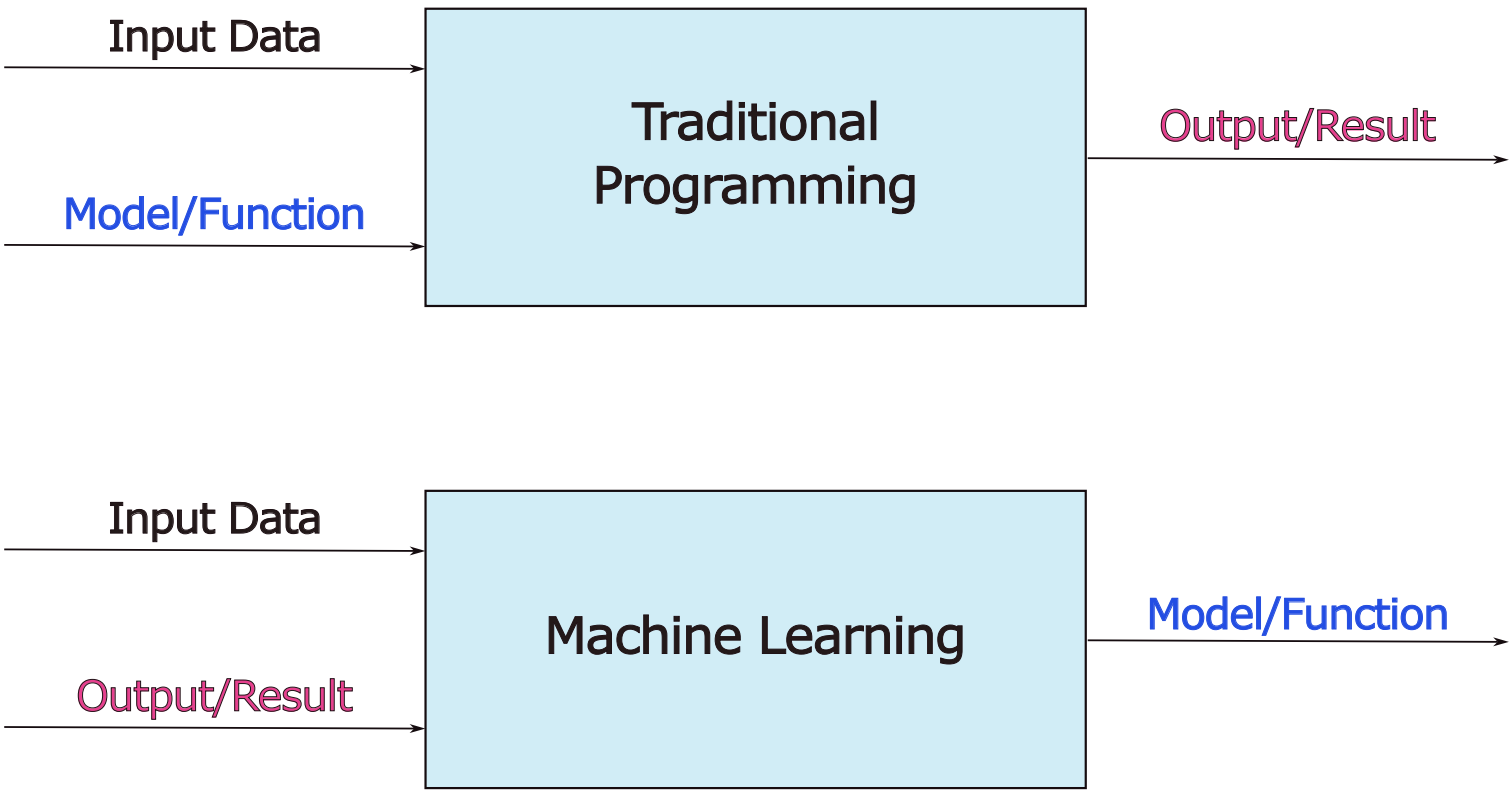
- **Lack of Exact Mathematical Function:**

  When the relationship between input variables and output predictions is not easily captured by a mathematical function or formula, machine learning can be useful.

- **Scalability challenges with traditional methods:**

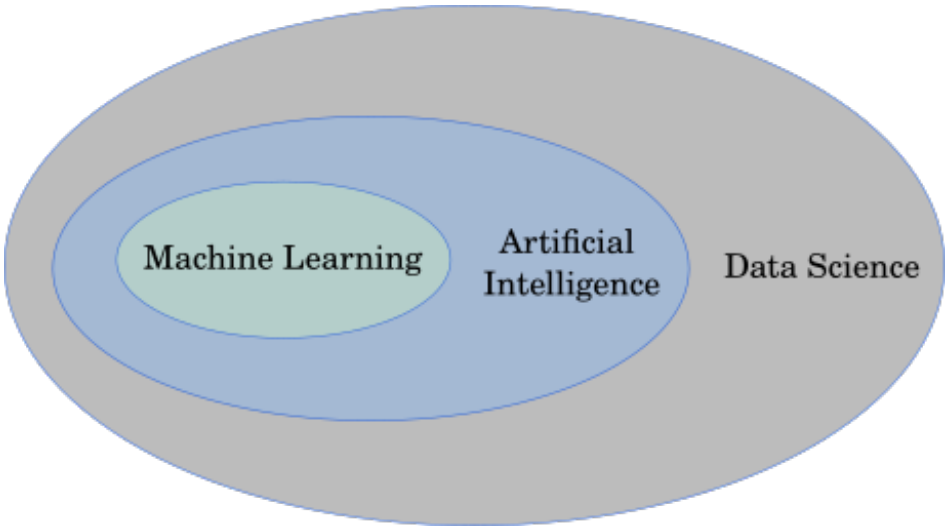
  Traditional programming methods might struggle to handle large-scale problems efficiently due to their reliance on fixed rules and explicit algorithms.

Traditional program vs ML



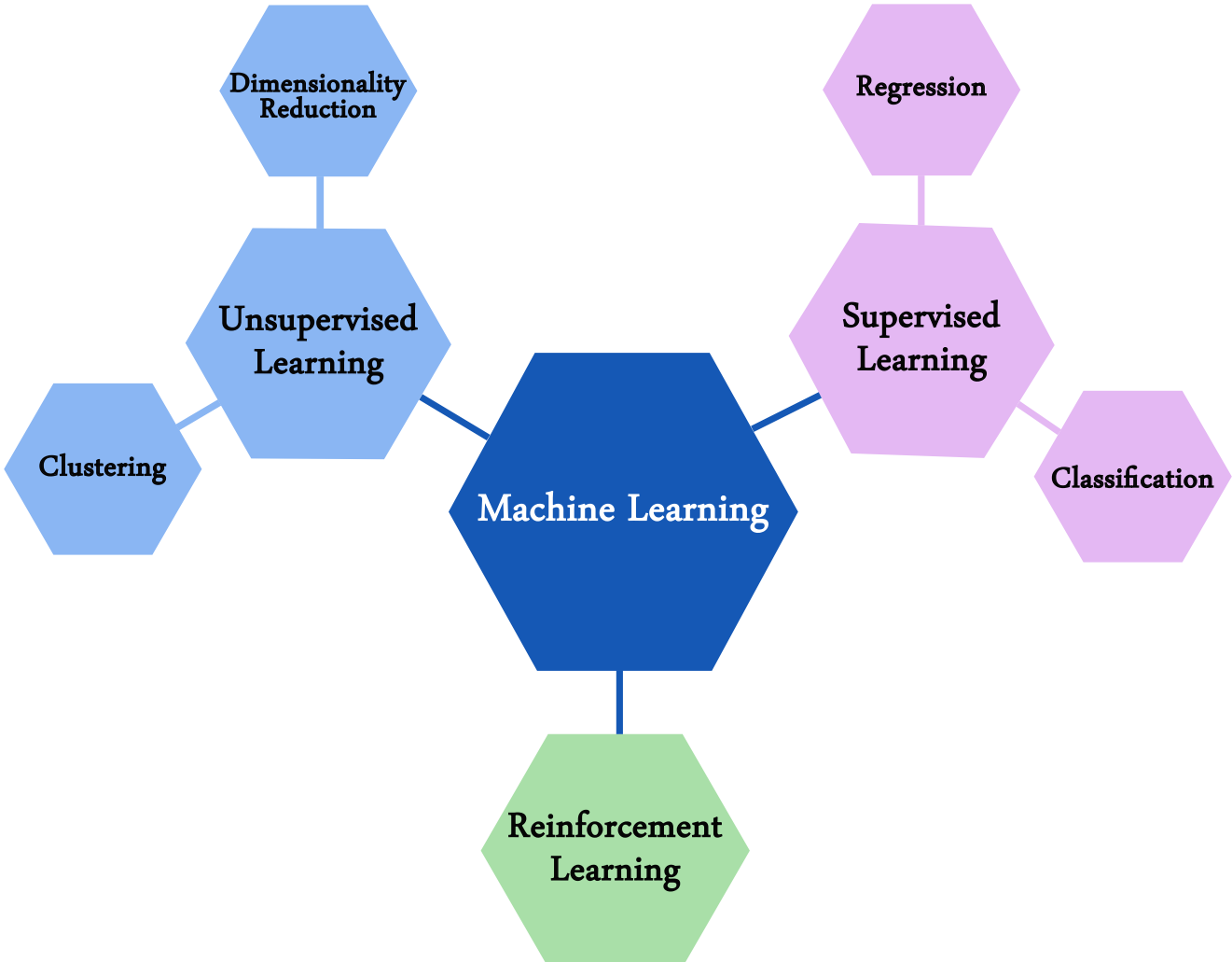
Overlapping?

ML vs AI vs DS



- "AI is the study of agents that receive percepts from the environment and perform actions. An agent is anything that can be viewed as perceiving its environment through sensors and acting upon that environment through actuators.", *by Stuart Russell and Peter Norvig*
- DS is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data, and apply knowledge and actionable insights from data across a broad range of application domains.

Types of ML


Dimensionality Reduction
Regression
Unsupervised Learning
Supervised Learning
Clustering
Machine Learning
Classification
Reinforcement Learning

Types of ML

- **Supervised learning:** labeled data, i.e. input-output pairs
  - Disease gene identification, email spam classification
  - Decision tree, Random forest, support vector machine, neural networks, etc
- **Unsupervised learning:** unlabeled data, i.e, data contain input only
  - Customer grouping
  - K-means clustering, DBSCAN, Gaussian Mixture Model, etc
- **Reinforcement learning:** designing machines and software agents that can automatically determine the ideal behavior within a specific context to maximize its performance.
  - Chess, autonomous driving

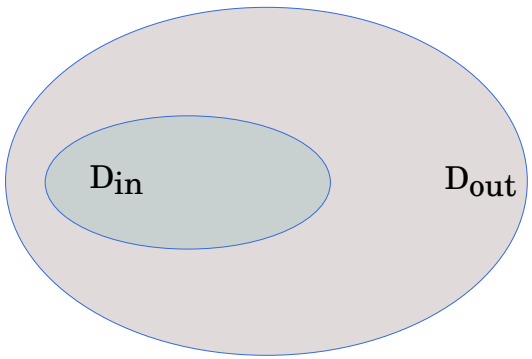
Supervised learning

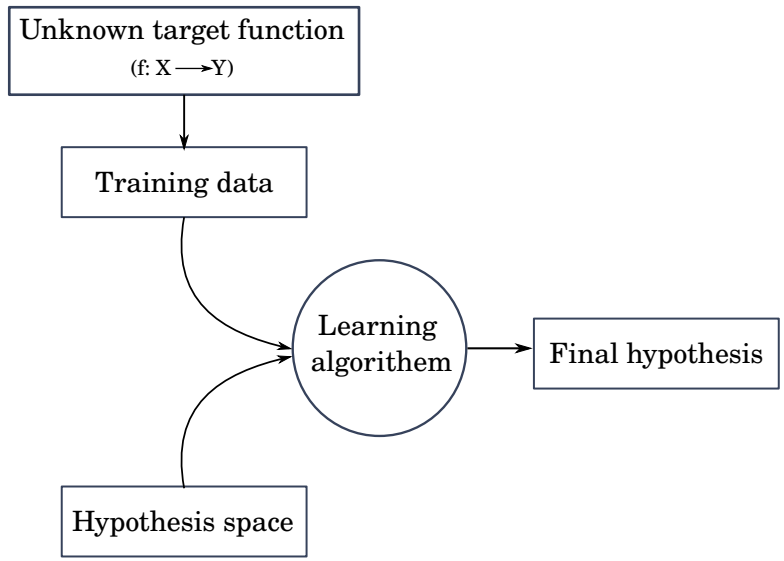
- $\mathbb{D} = \{(\mathbf{x}_i, \mathbf{y}_i) \mid \mathbf{x}_i \in \mathbb{X},\ \mathbf{y}_i \in \mathbb{Y}\}$
  - Each instance x_i represents an input object, e.g., an image.
  - Vector is the most common representation form.
  - $\mathbb{X}$: set of all instances or instance space, often $\mathbb{X} \subseteq \mathbb{R}^d$
  - Each instance $\mathbf{x}_i$ is associated with a label, $y_i \in Y$, e.g., spam or not spam. Y: label set.
- The target or true function

$$f : \mathbb{X} \longrightarrow \mathbb{Y} \text{ s.t } f(x_i) = y_i$$

  - It maps each input $\mathbf{x}_i$ with a label y_i
  - f is an unknown function.
- Often, we observe or have access to only a subset of data from the domain, $\mathbf{D}_{in} \subset \mathbf{D}$

Supervised learning





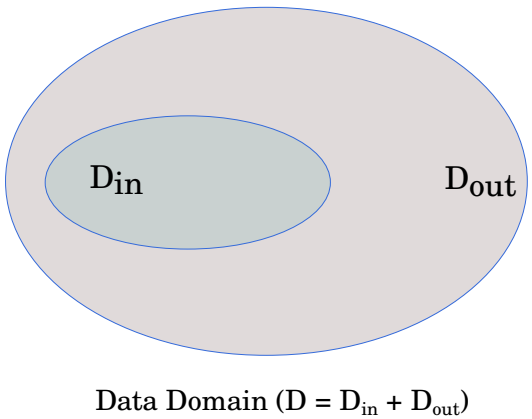
Let $\mathcal{H}$ be a hypothesis space (set of functions). Given a train set, $\mathbf{D}_{\text{in}}$, a supervised learning method aims at learning a function $h(x) \in \mathcal{H}$ s.t $h \approx f$. The estimated function will be used to predict for unseen data (generalization), $\mathbf{D}_{\text{out}}$.

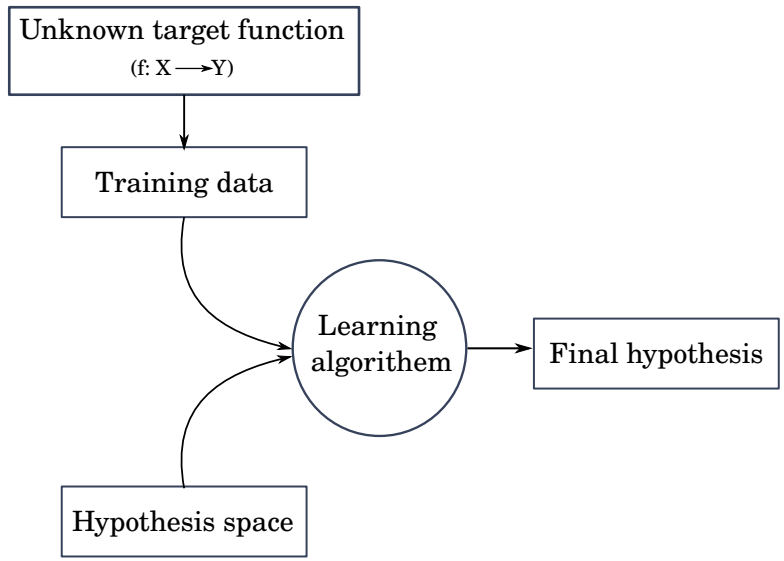
$$E_{\text{in}} \longrightarrow \min$$

Targeting to have $E_{\text{out}} \longrightarrow \min$. E_{out} (error on unseen data) indicates the generalization

$$(E_{\text{in}} = \frac{1}{N} \sum_{x_i \in \mathbb{D}_{\text{in}}} \mathcal{L}(f(\mathbf{x}_i), h(\mathbf{x}_i)); \quad E_{\text{out}} = \frac{1}{N} \sum_{x_i \in \mathbb{D}_{\text{out}}} \mathcal{L}(f(\mathbf{x}_i), h(\mathbf{x}_i)))$$

Supervised learning





Supervised learning tasks:

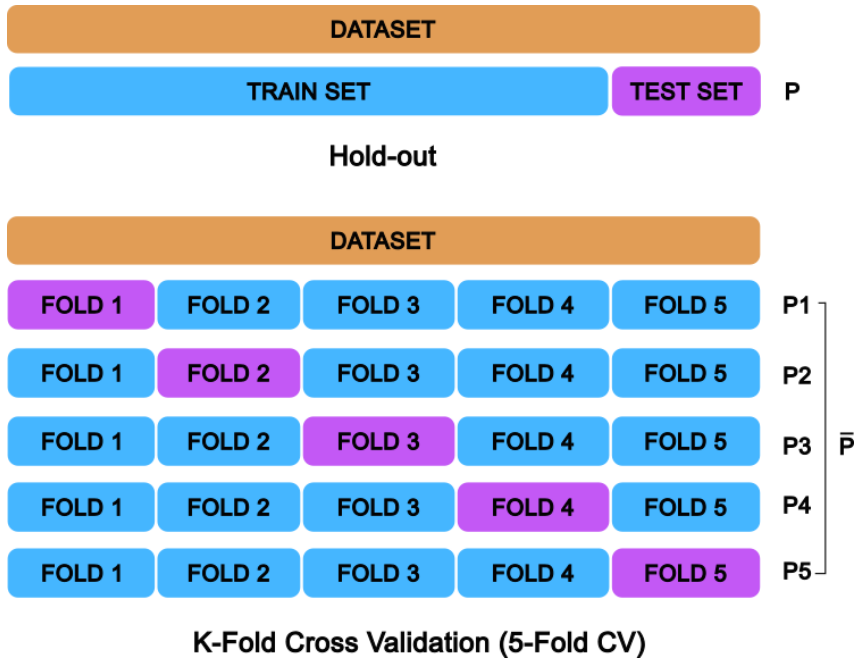
- $\mathbb{Y}$: categorical labels
  - $\mathbb{Y} = \{0, 1\}$: Binary classification
  - $\mathbb{Y} = \{0, 1, \ldots, N\},\ N > 2$: Multi-class classification
  - $f : \mathbb{X} \longrightarrow 2^{\mathbb{Y}}$: Multi-label, multi-class classification
- $\mathbb{Y} \subseteq R$: Regression

Model performance assessment

We need an unbiased evaluation since it helps to accurately reflect the model's performance on unseen data without distortions or unfair influences. It includes:

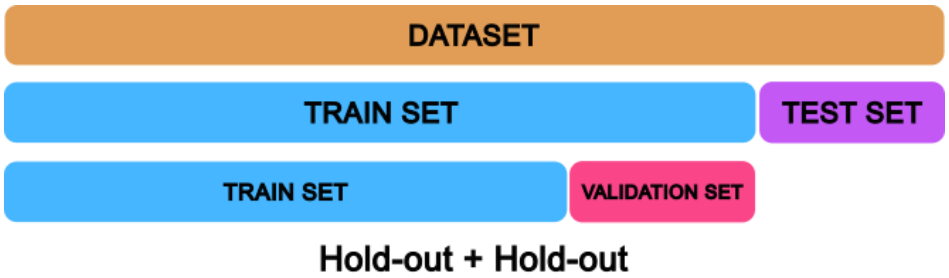
- Proper data splitting:
  - Training, validation, and test sets: train for model training, validation for hyperparameter tuning and test for final evaluation of the model.
  - Statified splitting: ensures the distribution of classes is preserved across train, validation and test.
- Avoiding Data Leakage, data contamination

Evaluation: Data splitting

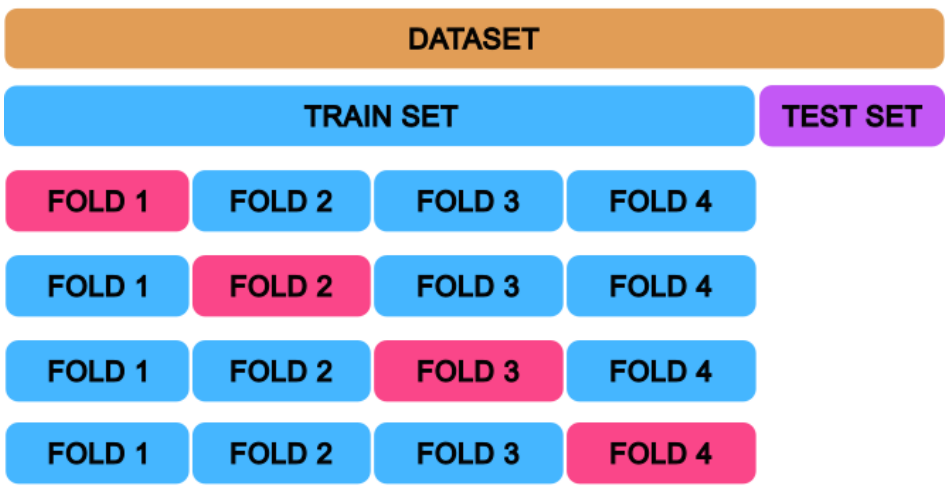


- Hold-out:
  - Model is trained on the train set and is used to predict on the test set.
  - The model assessment can be biased due to the selection of the test set.
- Cross-validation:
  - Less variance in performance estimates
  - Can be computationally expensive if the number of folds, K, is high.

Evaluation: Hyperparameter estimate

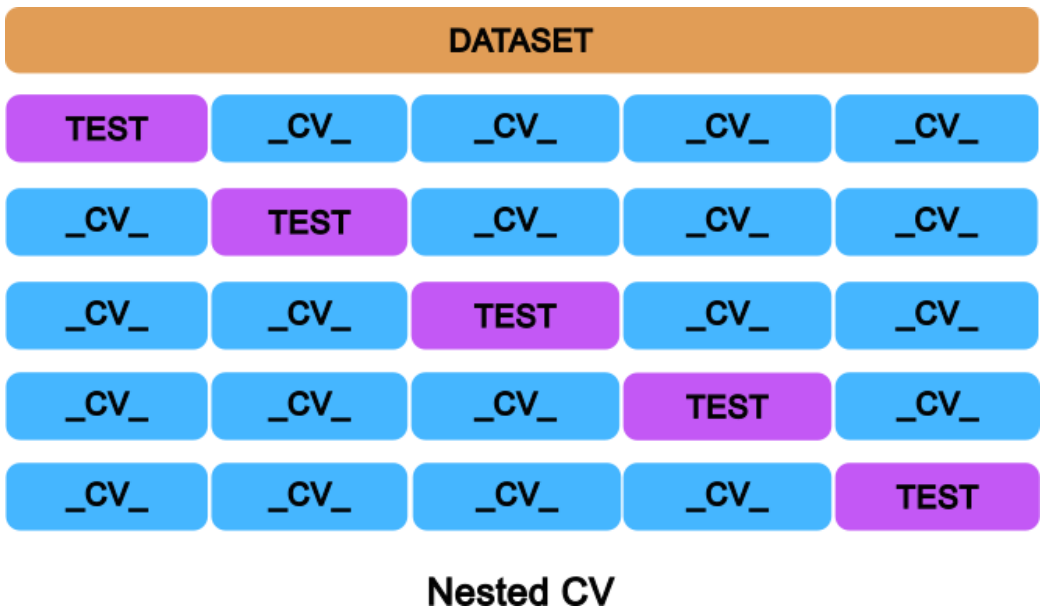


Hold-out + Hold-out



Hold-out + CV

Evaluation: Hyperparameter estimate



Nested CV

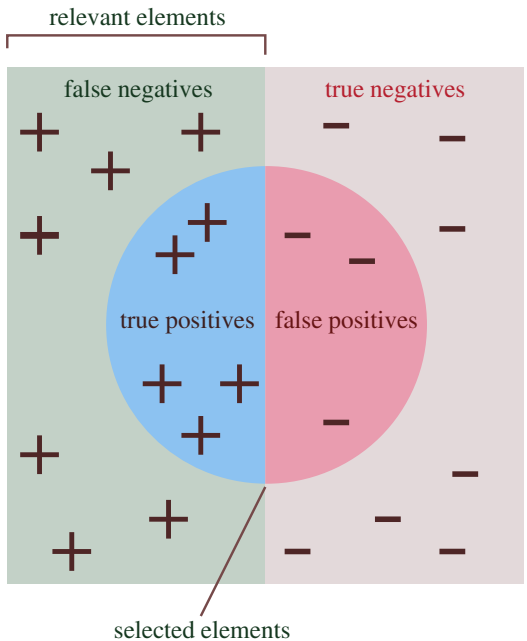
- In practice, K = 5 or K = 10 are commonly used.
- When K = # training examples, we have Leave-one-out CV.
- A high value of K results in lower bias and variance, while a low value of K leads to an increase in bias and variance.

Evaluation metrics

- A binary classifier $\mathcal{M}$ estimates $P(y = 1|x_i)$, i.e $\mathcal{M}$ returns a score $\forall x_i$.
- If $\mathcal{M}(x_i) > \sigma$, $\hat{y} = 1$, else 0; where σ: a threshold

In the figure, we assume that the model predicts as positive with instances in the circle.

- True positive (TP):
  - True label: Positive
  - Predicted label: Positive
- False positive (FP):
  - True label: negative
  - Predicted label: Positive
- True negative (TN):
  - True label: Negative
  - Predicted label: Negative
- False negative (FN):
  - True label: Positive
  - Predicted label: Negative



Evaluation metrics

- $\text{TPR} = \text{Sensitivity} = \text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$
- $\text{FPR} = \frac{\text{FP}}{\text{FP+TN}}$
- $\text{TNR} = \text{Specificity} = \frac{\text{TN}}{\text{TN+FP}}$
- $\text{FNR} = \frac{\text{FN}}{\text{FN+TP}}$
- $\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FP+TN+FN}}$
- $\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$
- $\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision+Recall}}$

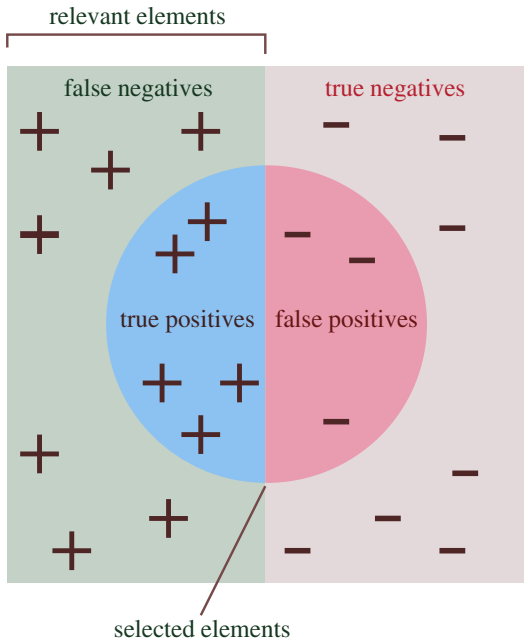
Given the confusion matrix on the right, we can calculate:

$\text{TPR} \approx 0.83$

$\text{TNR} \approx 0.95$

$\text{Accuracy} \approx 0.91$

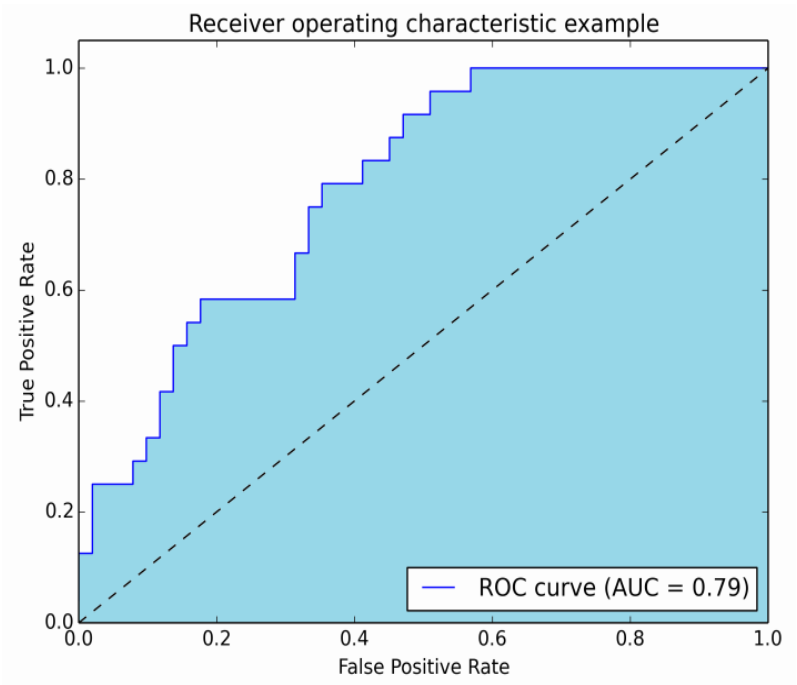
$\text{F1} \approx 0.87$

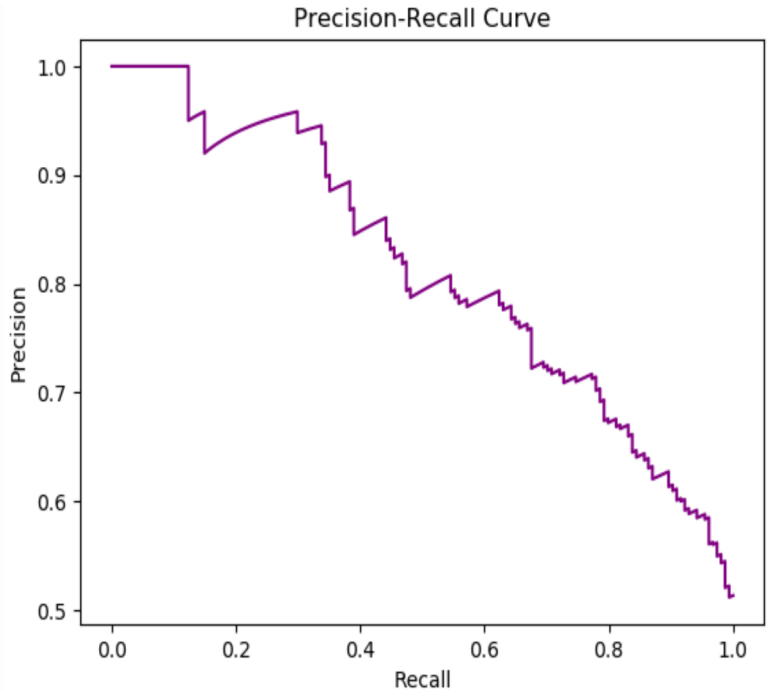


Predicted Labels \ True Labels	Positive	Negative
Positive	TP (50)	FP (5)
Negative	FN (10)	TN (100)

Evaluation metrics

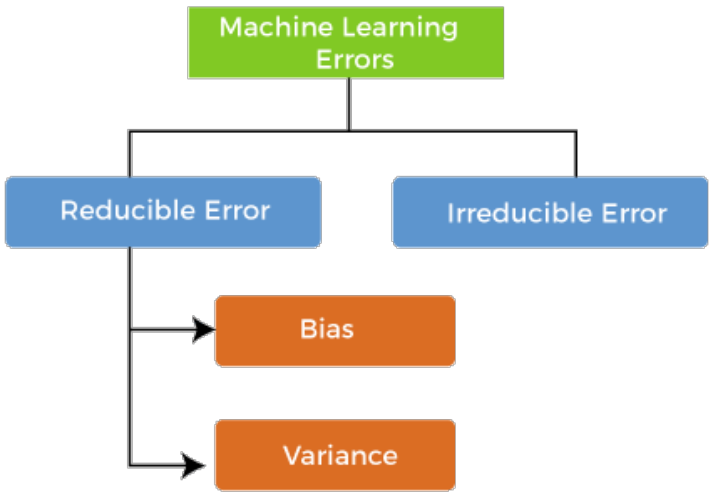
- The above metrics are used to calculate the performance for a given threshold
- Defining a threshold can be sensitive
- Need metrics that consider all possible thresholds





Illustrations of *Left:* area under receiver operating characteristic curve, AUROC and *Right:* area under the precision and recall curve, AUPRC.

Bias and Variance



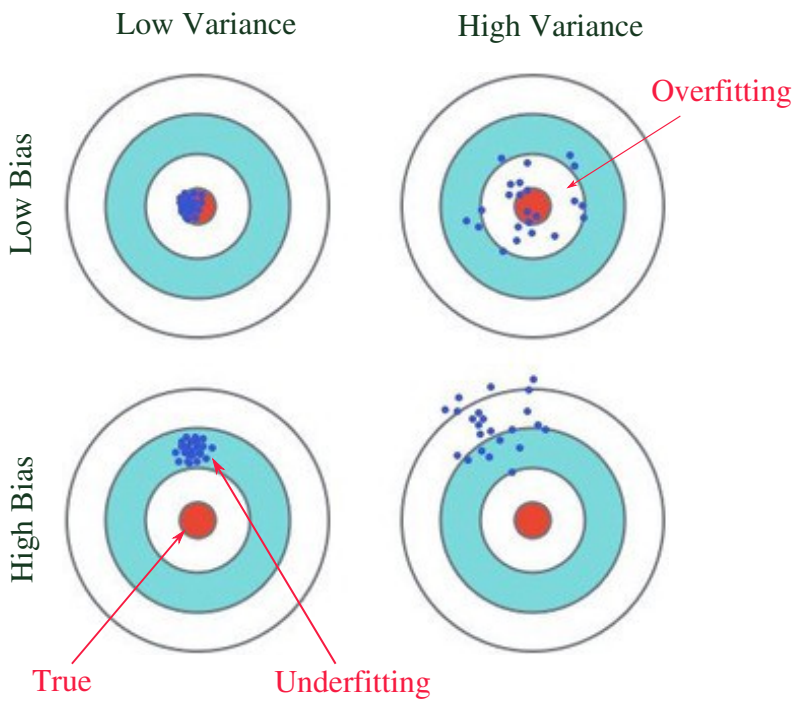
- **Reducible error:**
  - can be reduced by improving the model
  - associated with the model's ability to learn from the data.
- **Irreducible error:**
  - cannot be reduced by any model, no matter how sophisticated.
  - It is inherent in the data and arises due to factors that cannot be predicted or controlled.

Bias and Variance

Let the true and predicted values be y and $\hat{y}$.

- **Bias:** $E(\hat{y}) - y$
  - Difference between the expected predicted and true values.
  - Caused by simplifying assumptions made by a model to make the target function easier to learn.
  - Low bias suggests fewer assumptions about the form of the target function.
  - High-Bias: suggests more assumptions about the form of the target function.
- **Variance:** $E(\hat{y} - E(\hat{y})^2)$
  - Variance of estimated values predicted by models trained on different subsets of data.
  - Low Variance suggests small changes to the estimate of the target function with changes to the training dataset.
  - High Variance: suggests large changes to the estimate of the target function with changes to the training dataset.

Bias and Variance



Bias and Variance trade-off

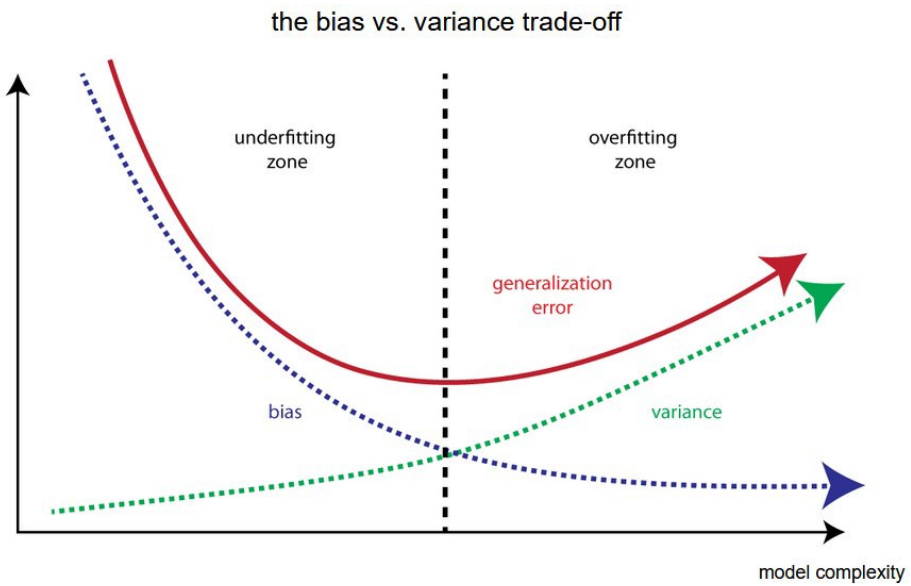


Image by Seth Mottaghinejad

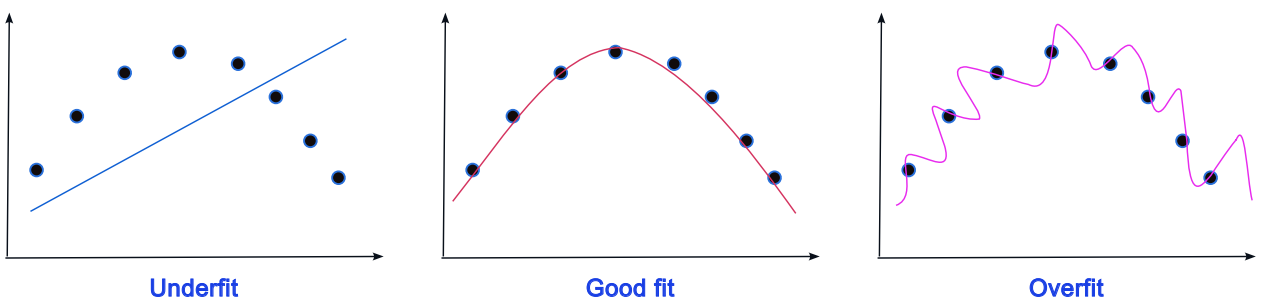
- Goal: low bias, low variance
- Increasing the bias will decrease the variance
- Increasing the variance will decrease the bias

$\Longrightarrow$ Trade-off between bias and variance

Underfitting

Underfitting occurs when a model is too simple to capture the underlying patterns in data.

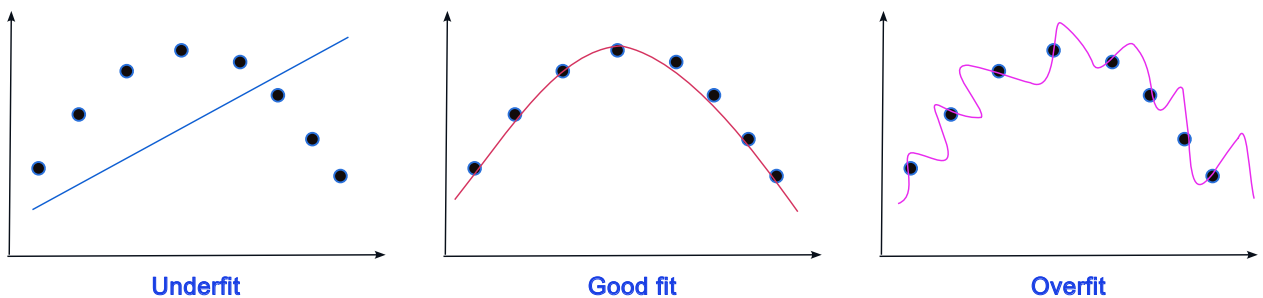
- Symptoms: low accuracy on both the training and test data.
- Causes:
  - Oversimplified models
  - Insufficient training: Training the model for too few epochs
  - Poor Feature Selection: Not using enough relevant features or using features that do not adequately represent the underlying problem.
- Solutions:
  - Increase model complexity
  - Feature engineering: create more informative features
  - Longer training



Overfitting

Overfitting occurs when a model learns not only the underlying patterns in the training data but also the noise and random fluctuations.

- Symptoms: high accuracy on train, but significantly lower accuracy on test
- Causes:
  - Model is too complex.
  - Small training set
  - Too many features
- Solutions:
  - Simplify model
  - Regularization
  - Increase the train data
  - Cross-validation



References

- Hastie, Trevor, et al. The elements of statistical learning: data mining, inference, and prediction. Vol. 2. New York: springer, 2009.
- Machine learning mastery
- Towards Data Science