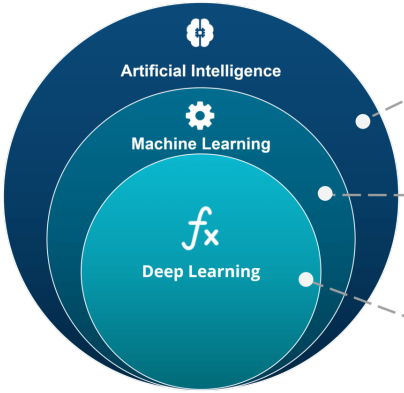
Machine Learning Problems

CSCI/DSCI 575 Advanced Machine Learning

Department of Computer Science
Colorado School of Mines

AI *vs.* ML *vs.* Deep Learning



ARTIFICIAL INTELLIGENCE
A technique which enables machines to mimic human behaviour

MACHINE LEARNING
Subset of AI technique which use statistical methods to enable machines to improve with experience

DEEP LEARNING
Subset of ML which make the computation of multi-layer neural network feasible

Image source: Edureka!

Types of learning problems (not exhaustive)

- **Supervised learning**: explicit feedback in the form of examples and target labels
  - goal to make predictions based on examples (classify them,predict prices, etc)
- **Unsupervised learning**: only examples, no explicit feedback
  - goal to reveal structure in the observed data
- **Semi-supervised learning**: limited explicit feedback, mostly only examples
  - tries to improve predictions based on examples by making use of the additional *unlabeled* examples
- **Reinforcement learning**: delayed and partial feedback, no explicit guidance
  - goal to minimize the cost of a sequence of actions (policy)

Types of learning problems (not exhaustive)

- **Supervised learning**: explicit feedback in the form of examples and target labels
  - goal to make predictions based on examples (classify them,predict prices, etc)
- **Unsupervised learning**: only examples, no explicit feedback
  - goal to reveal structure in the observed data
- **Semi-supervised learning**: limited explicit feedback, mostly only examples
  - tries to improve predictions based on examples by making use of the additional *unlabeled* examples
- **Reinforcement learning**: delayed and partial feedback, no explicit guidance
  - goal to minimize the cost of a sequence of actions (policy)

Types of learning problems (not exhaustive)

- **Supervised learning**: explicit feedback in the form of examples and target labels
  - goal to make predictions based on examples (classify them,predict prices, etc)
- **Unsupervised learning**: only examples, no explicit feedback
  - goal to reveal structure in the observed data
- **Semi-supervised learning**: limited explicit feedback, mostly only examples
  - tries to improve predictions based on examples by making use of the additional *unlabeled* examples
- **Reinforcement learning**: delayed and partial feedback, no explicit guidance
  - goal to minimize the cost of a sequence of actions (policy)

Types of learning problems (not exhaustive)

- **Supervised learning**: explicit feedback in the form of examples and target labels
  - goal to make predictions based on examples (classify them,predict prices, etc)
- **Unsupervised learning**: only examples, no explicit feedback
  - goal to reveal structure in the observed data
- **Semi-supervised learning**: limited explicit feedback, mostly only examples
  - tries to improve predictions based on examples by making use of the additional *unlabeled* examples
- **Reinforcement learning**: delayed and partial feedback, no explicit guidance
  - goal to minimize the cost of a sequence of actions (policy)

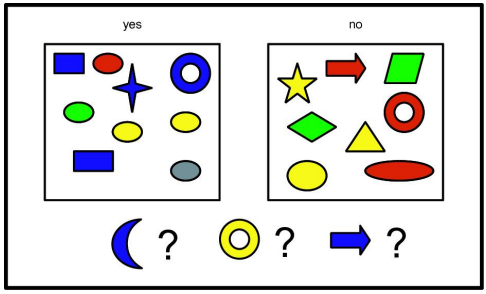
Supervised Learning

- In **supervised learning**,
  - the goal is to learn a mapping from inputs $\mathbf{x}$ to outputs y,
  - given a labeled set of input-output pairs $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$.
    - Here $\mathcal{D}$ is called the training set,
    - and N is the number of training examples.

- In the simplest setting, each training input $\mathbf{x}_i$ is a D-dimensional vector of numbers, representing, say, the height and weight of a person. These are called features, attributes or covariates.
  - In general, however, $\mathbf{x}_i$ could be a complex structured object, such as an image, a sentence, an email message, a time series, a molecular shape, a graph, etc.

Supervised Learning

Left: Some labeled training examples of colored shapes, along with 3 unlabeled test cases. Right: Representing the training data as an $N \times D$ design matrix. Row i represents the feature vector $\mathbf{x}_i$. The last column is the label, $y_i \in \{0, 1\}$.



D features (attributes)

N cases

Color	Shape	Size (cm)
Blue	Square	10
Red	Ellipse	2.4
Red	Ellipse	20.7

Label
1
1
0

Image source: Murphy book, Figure 1.1

Supervised Learning

The form of the output or *response variable* can in principle be anything, but most methods assume that y_i is

- a **categorical** or **nominal** variable from some finite set, $y_i \in \{1, \ldots, C\}$ (such as male or female),
- or that y_i is a **real-valued** scalar (such as income level).

- When y_i is categorical, the problem is known as *classification* or *pattern recognition*,
- and when y_i is real-valued, the problem is known as *regression*.
- Another variant, known as *ordinal regression*, occurs where label space Y has some natural ordering, such as grades A — F.

Unsupervised learning

- In unsupervised learning, we are only given inputs **x without** the outputs labels y.
- The goal is to discover "interesting structure" in the data, which is sometimes called *knowledge discovery*.

Unsupervised learning

- In unsupervised learning, we are only given inputs **x without** the outputs labels y.
- The goal is to discover "interesting structure" in the data, which is sometimes called *knowledge discovery*.

Left: The height and weight of some people. Right: A possible clustering using $K = 2$ clusters.

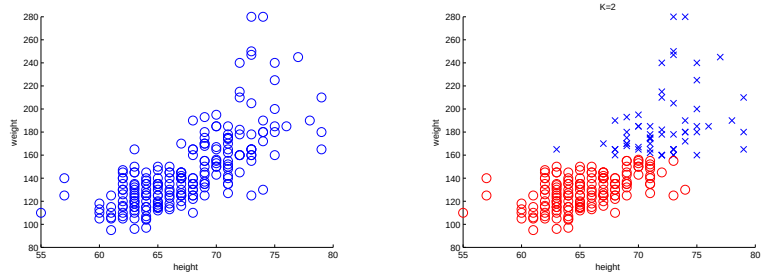


Image source: Murphy book, Figure 1.8

Semi-supervised learning

- **General idea**: learning from both labeled and unlabeled data
- **Semi-supervised Classification/Regression**
  - Given: Labeled training data $\mathcal{L} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{L}$, unlabeled data $\mathcal{U} = \{(\mathbf{x}_i)\}_{i=L+1}^{L+U}$ (usually $U \gg L$).
  - Goal: Learning a classifier f **better than using labeled data alone**.

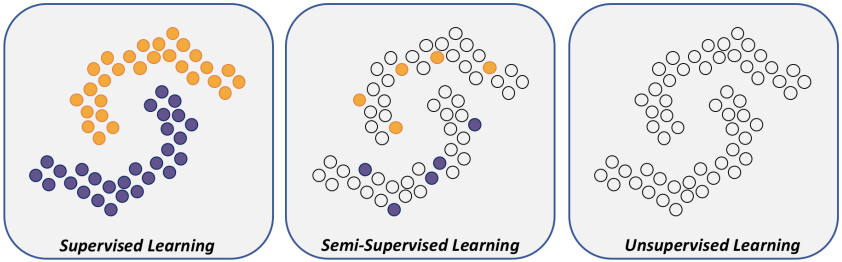


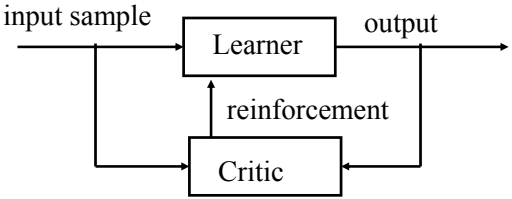
Image source: KDnuggets

Semi-supervised learning

- **General idea**: learning from both labeled and unlabeled data
- **Semi-supervised Classification/Regression**
  - Given: Labeled training data $\mathcal{L} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{L}$, unlabeled data $\mathcal{U} = \{(\mathbf{x}_i)\}_{i=L+1}^{L+U}$ (usually $U \gg L$).
  - Goal: Learning a classifier f **better than using labeled data alone**.
- **Semi-Unsupervised Learning**
  - Given: Unlabeled data $\{(\mathbf{x}_i)\}_{i=1}^{N}$ and the goal could be to do clustering or dimensionality reduction.
  - Additionally given: Some constraints on the data.
    - E.g., for clustering: two points must be in the same cluster, or two points must not be in the same cluster; for dimensionality reduction: two points must be close after the projection.
- In this course and most literature: **Semi-supervised Learning (SSL)** refer to semi-supervised classification/regression.

Reinforcement learning

- We want to learn: $f : X \to Y$.
- We see samples of $\mathbf{x}$ but not labels y.
- Instead of y we get a feedback (reinforcement) from a critic about how good our output was.



- The goal is to select outputs that lead to the best reinforcement.

Machine learning problems

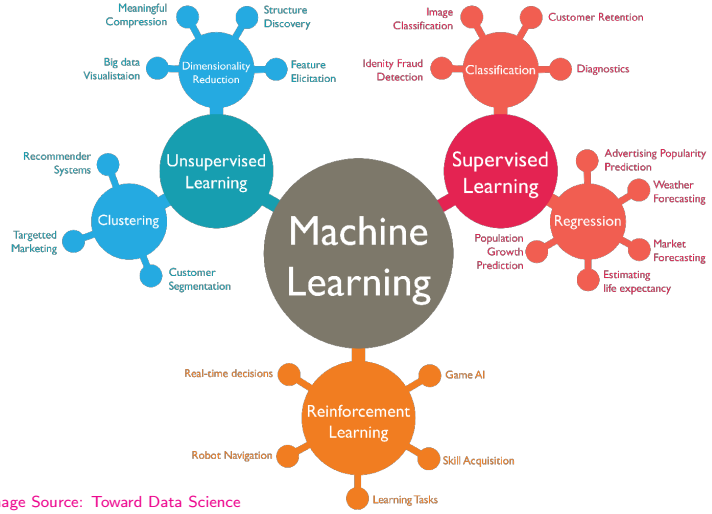


Image Source: Toward Data Science

Copyright statement

This slide set is designed only for teaching CSCI/DSCI-575 Advanced Machine Learning at the Department of Computer Science of Colorado School of Mines.

Some contents in this slide set are obtained from Internet and maybe copyright sensitive. Copyright and all rights therein are retained by the respective authors or by other copyright holders. Distributing or reposting the whole or part of this slide set not for academic purpose is HIGHLY prohibited, unless the explicit permissions from all copyright holders are granted.