Introduction to Machine Learning

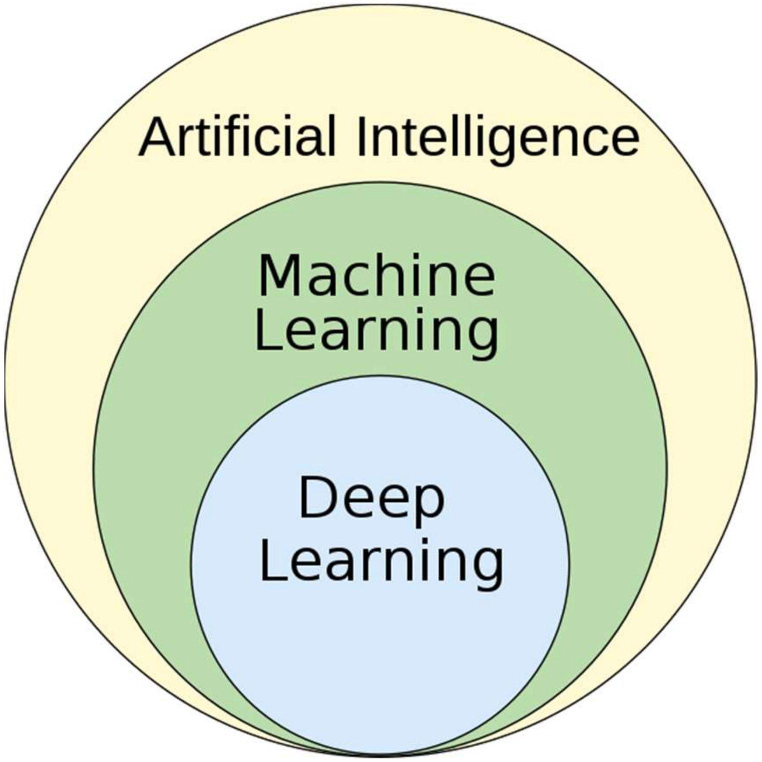
Outline

- What is machine learning?
- Machine learning applications
- Types of machine learning
    - Supervised learning
    - Unsupervised learning
    - Reinforcement learning
- Notation and conventions
- Machine learning terminology
- Machine learning in predictive modeling workflow
- Installing Python and packages

What is machine learning?

- “Machine learning as a field of study that gives computers the ability to learn without being explicitly programmed” - Arthur Samuel (1959)
- “A computer program is said to learn from experience *E* with respect to some class of tasks *T* and performance measure *P* if its performance at tasks in *T*, as measured by *P*, improves with experience *E*.” Tom M. Mitchell (1997)
- Machine learning is about designing algorithms that learn from data.

Artificial Intelligence, Machine Learning, and Deep Leaning



Source: https://en.wikipedia.org/wiki/File:AI_hierarchy.svg

Machine learning applications

- Credit card fraud detection: collect customer transactions to learn typical customer behaviour, then use this model to detect anomaly transactions.
- Recommender systems: providing better product recommendations by predicting user preferences based on preferences of similar users (collaborative filtering techniques)
- Sentiment analysis: collect voice of the customer materials for applications, for example, determine the attitude of a customer to a product based on their reviews
- And many more ...

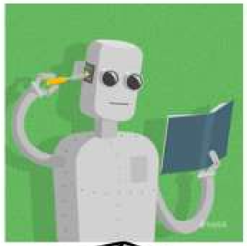
Types of machine learning

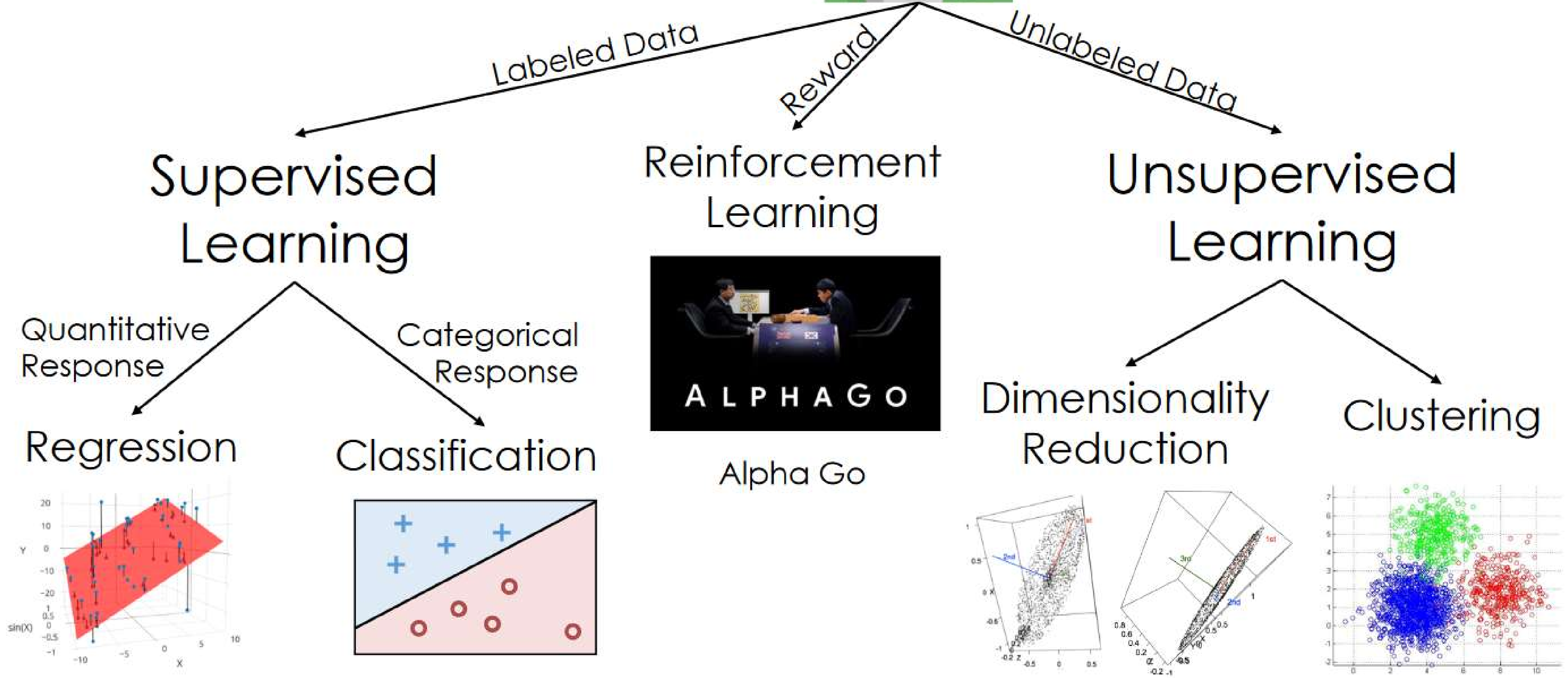
- Predictive or supervised learning: we learn a function to predict an output variable Y based on input variable X.
  - This function is learned based on labeled data $\{(x_i , y_i)\}_{i=1..N}$, which we call the training data.

- Descriptive or unsupervised learning: we are given only inputs $\{x_i\}_{i=1..N}$, and the goal is to find interesting patterns in this data.

- Reinforcement learning: develop a system (agent) to maximize the reward through a series of interactions with the environment.

Taxonomy of Machine Learning





Labeled Data

Reward

Unlabeled Data

Supervised Learning

Reinforcement Learning

Unsupervised Learning

Quantitative Response

Categorical Response

Regression

Classification

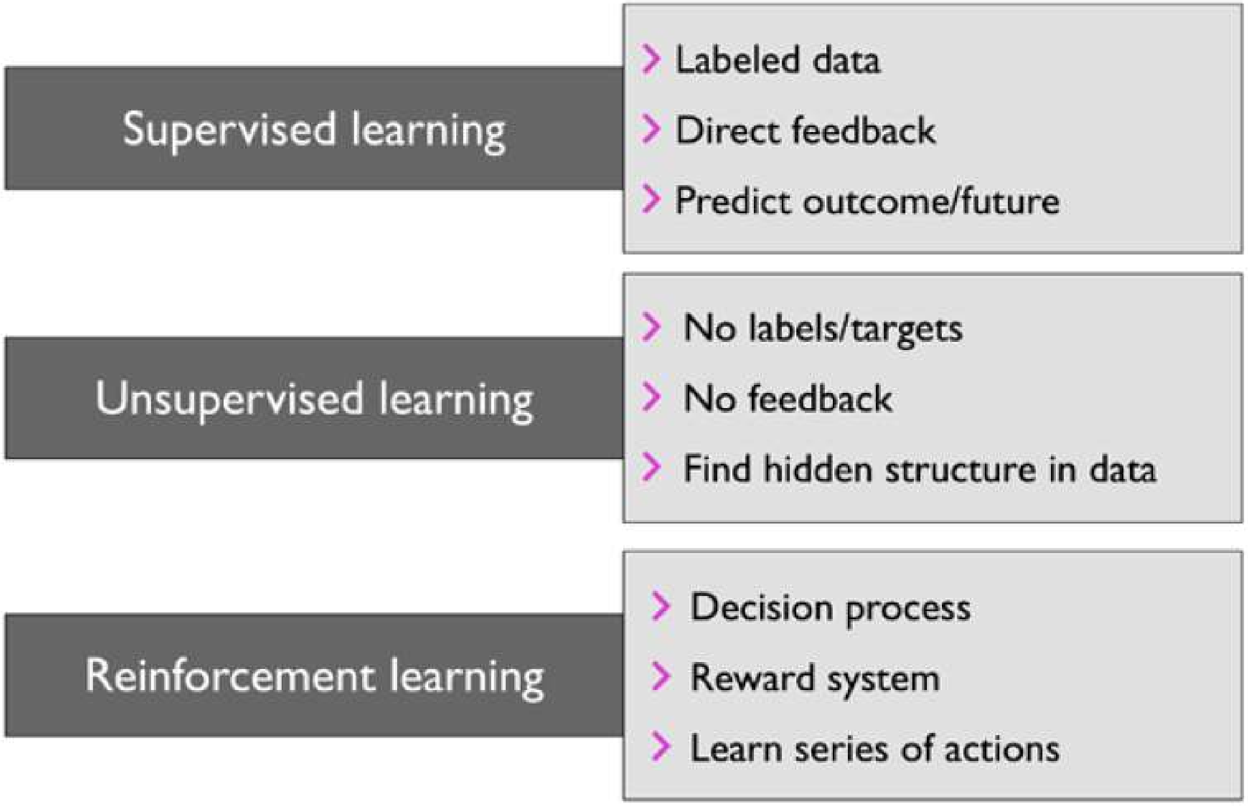
Dimensionality Reduction

Clustering

Alpha Go

Source: Joseph E. Gonzalez, *AI-Systems Big Ideas*, 2019.

The three different types of machine learning



Supervised learning

- Labeled data
- Direct feedback
- Predict outcome/future

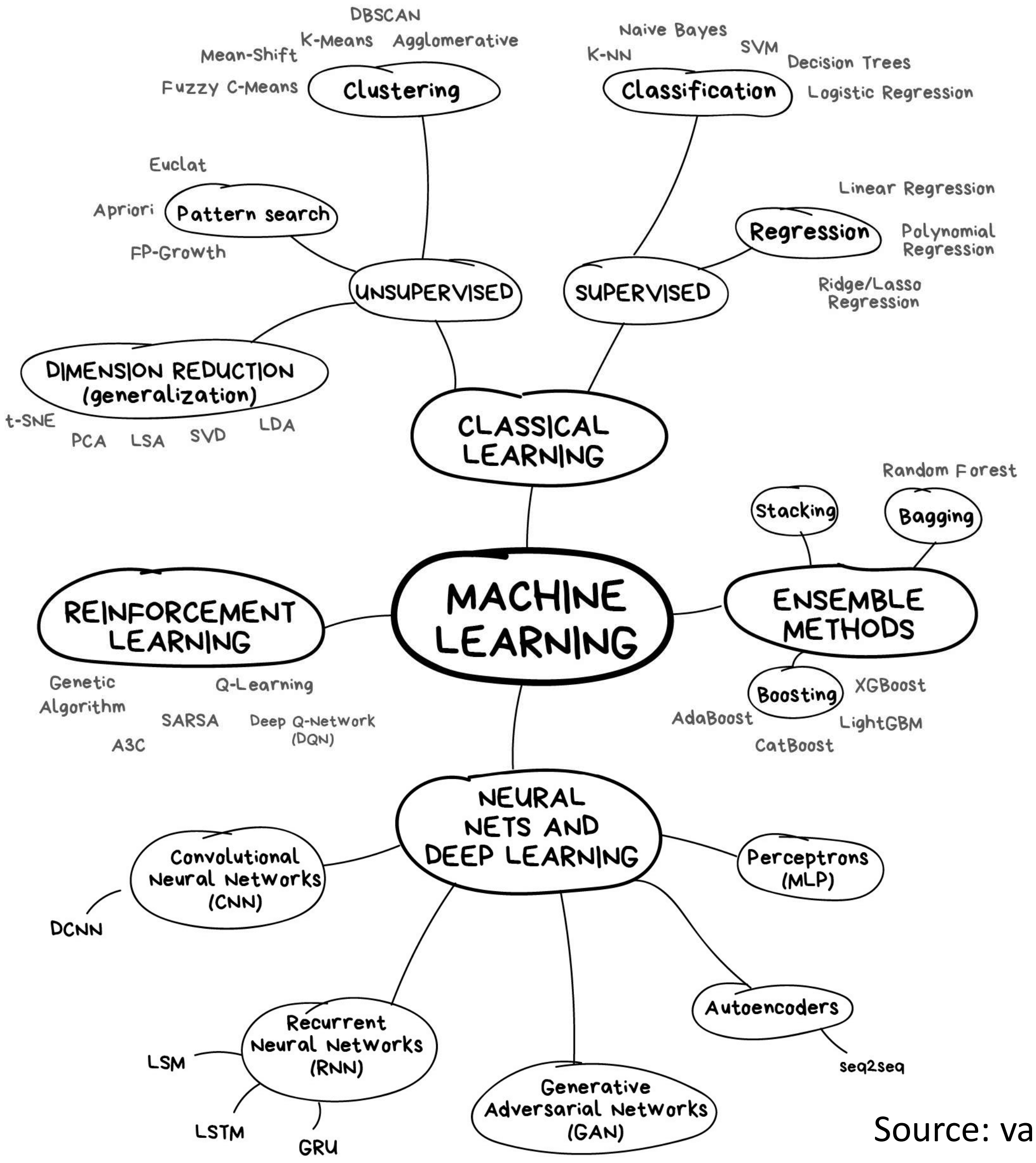
Unsupervised learning

- No labels/targets
- No feedback
- Find hidden structure in data

Reinforcement learning

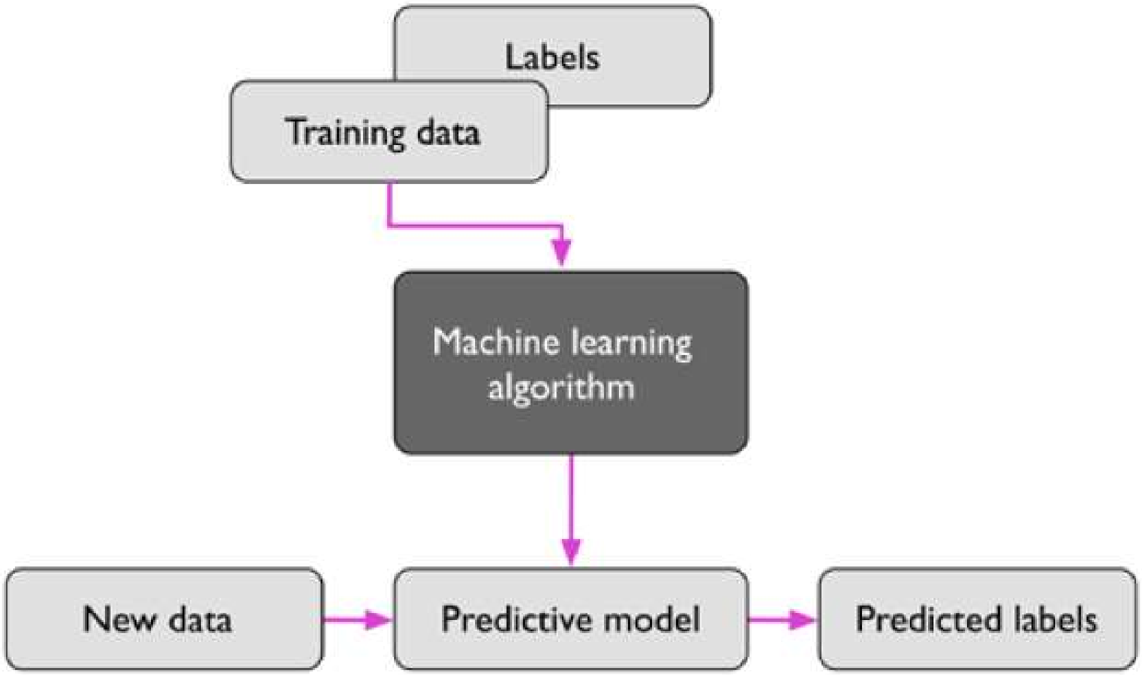
- Decision process
- Reward system
- Learn series of actions

Machine learning algorithms

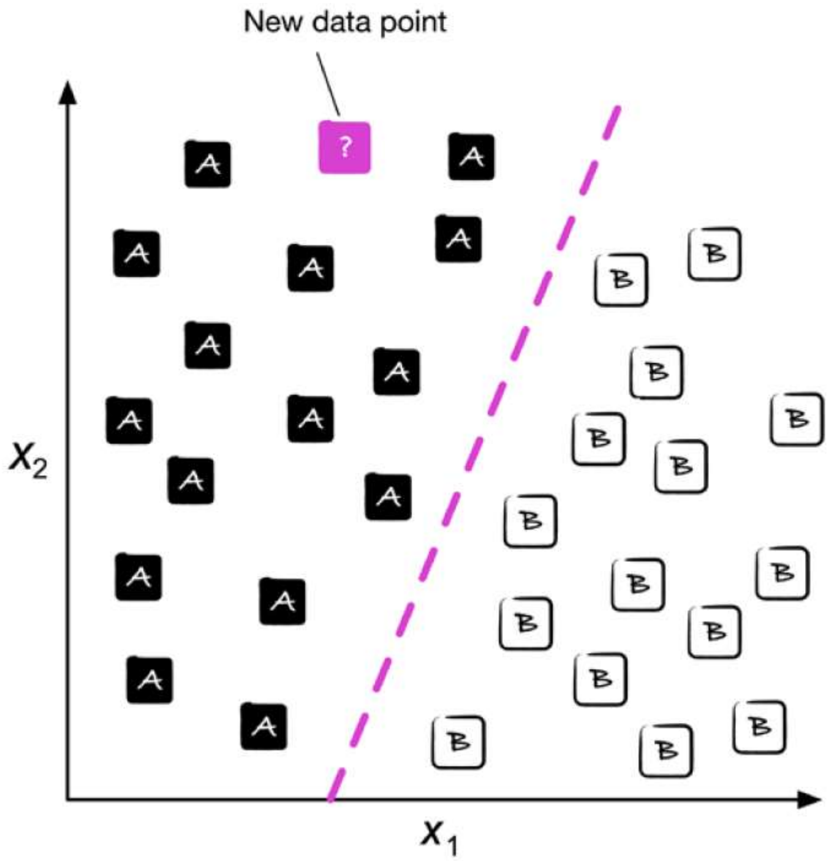


Source: vas3k.com/blog/machine_learning

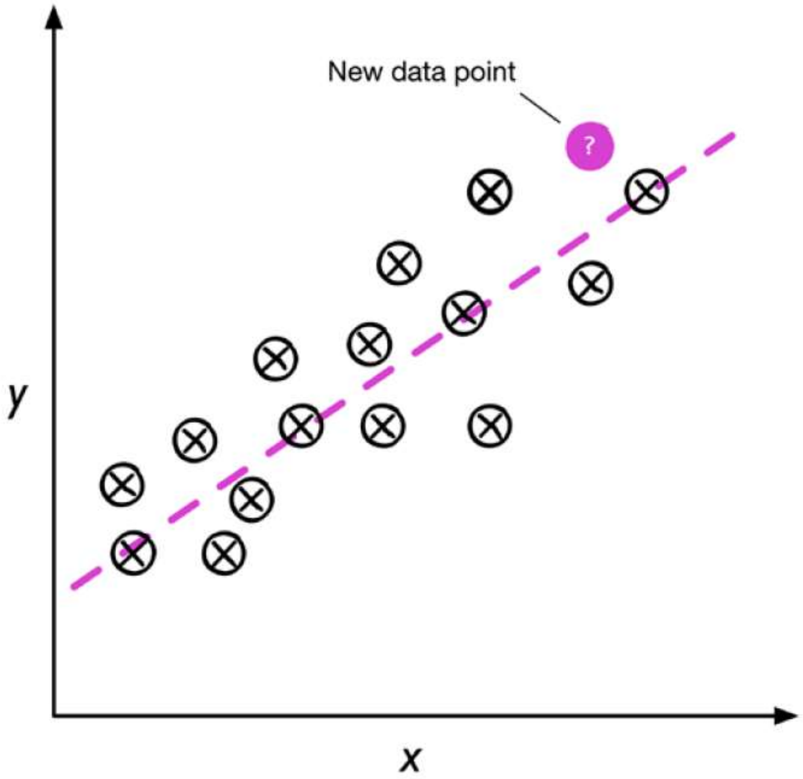
Supervised learning process



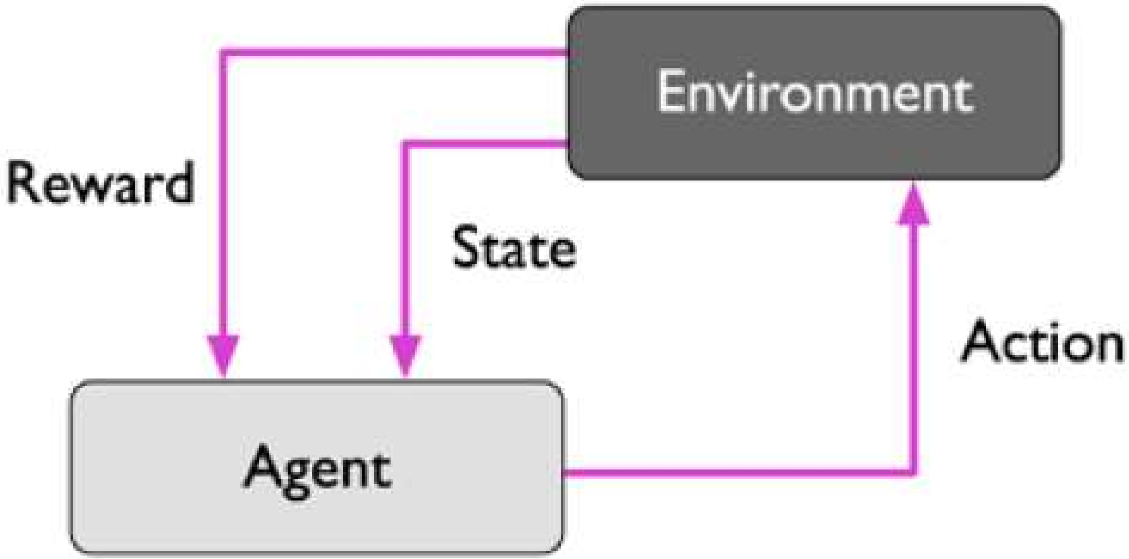
Classification - predicting class labels



Regression - predicting continuous outcomes

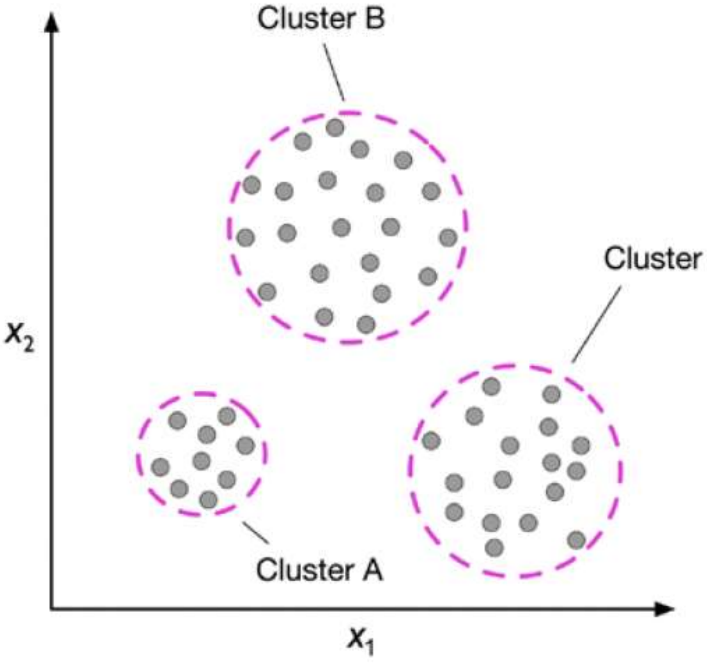


Reinforcement learning - solving interactive problems



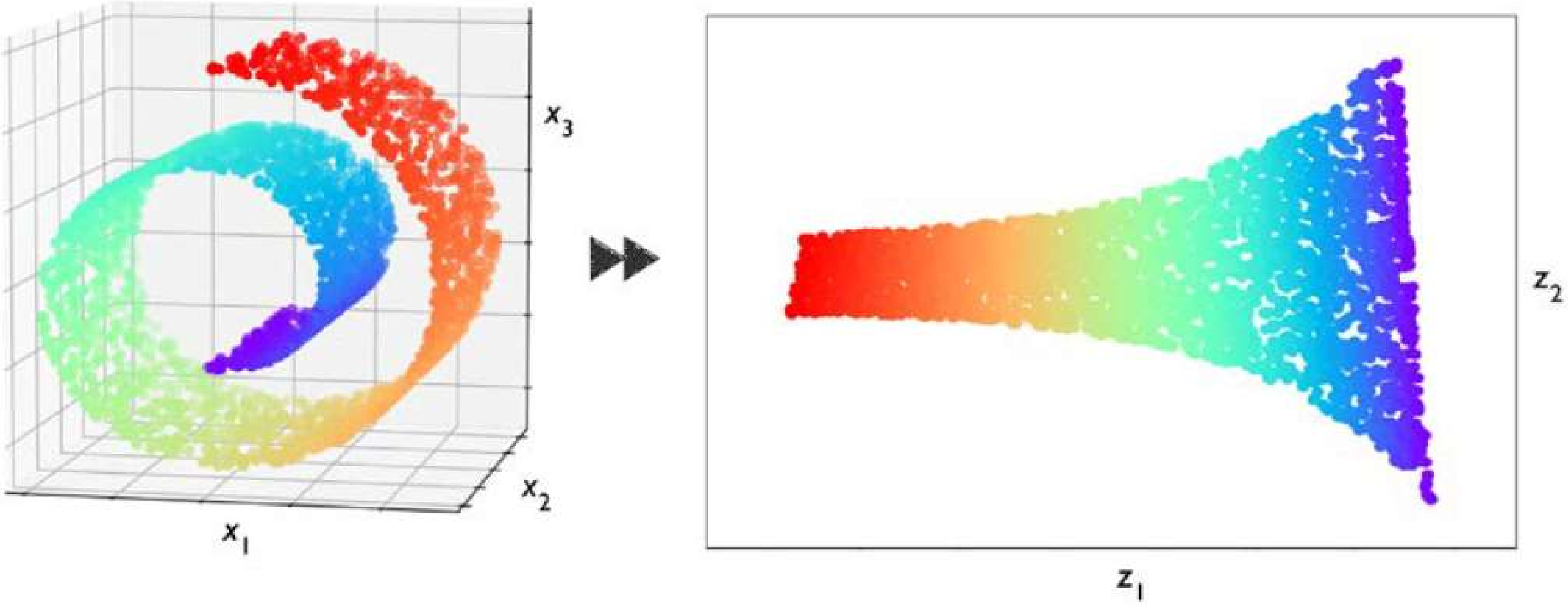
Unsupervised learning - discovering hidden structures

- Clustering - finding subgroups

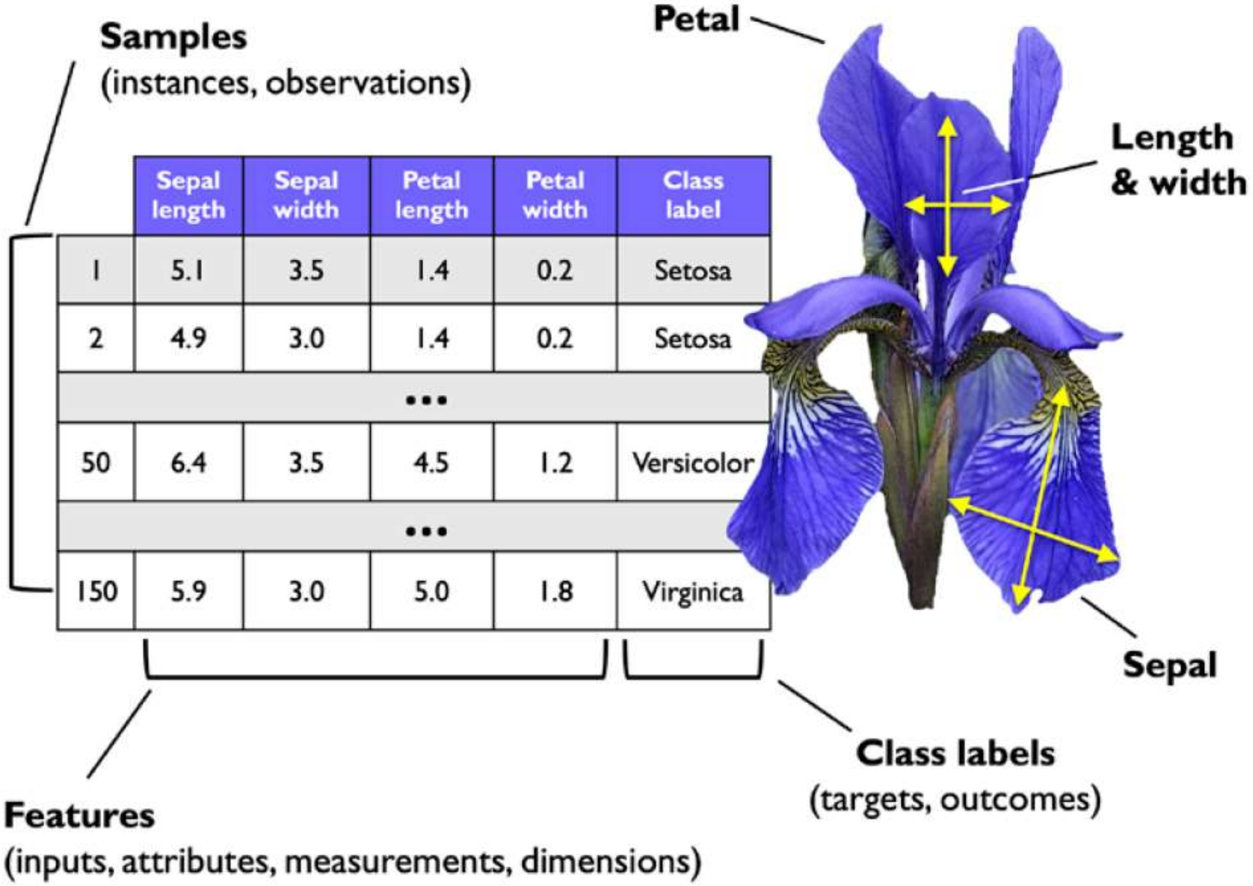


Unsupervised learning - discovering hidden structures

- Dimensionality reduction - data compression



Notation and conventions - the Iris dataset



	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
•••					
50	6.4	3.5	4.5	1.2	Versicolor
•••					
150	5.9	3.0	5.0	1.8	Virginica

Samples
(instances, observations)
Petal
Length
& width
Sepal
Class labels
(targets, outcomes)
Features
(inputs, attributes, measurements, dimensions)

Notation and conventions - the Iris dataset

- The Iris dataset: 150 examples and 4 features, $\boldsymbol{X} \in R^{150\times4}$

$$\begin{bmatrix} x_1^{(1)} & x_2^{(1)} & x_3^{(1)} & x_4^{(1)} \\ x_1^{(2)} & x_2^{(2)} & x_3^{(2)} & x_4^{(2)} \\ \vdots & \vdots & \vdots & \vdots \\ x_1^{(150)} & x_2^{(150)} & x_3^{(150)} & x_4^{(150)} \end{bmatrix}$$

- Vectors ($\boldsymbol{x} \in R^{n\times1}$): lowercase, bold-face letters
- Matrices ($\boldsymbol{X} \in R^{n\times m}$): uppercase, bold-face letters
- Single elements in a vector or matrix ($x^{(n)}, x_m^{(n)}$): letters in italics

Notation and conventions - the Iris dataset

- The Iris dataset: 150 examples and 4 features, $\boldsymbol{X} \in R^{150\times 4}$

$$\begin{bmatrix} x_1^{(1)} & x_2^{(1)} & x_3^{(1)} & x_4^{(1)} \\ x_1^{(2)} & x_2^{(2)} & x_3^{(2)} & x_4^{(2)} \\ \vdots & \vdots & \vdots & \vdots \\ x_1^{(150)} & x_2^{(150)} & x_3^{(150)} & x_4^{(150)} \end{bmatrix}$$

Class labels: $\mathrm{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(150)} \end{bmatrix}, y^{(1)} \in \{Setosa, Versicolor, Virginica\}$

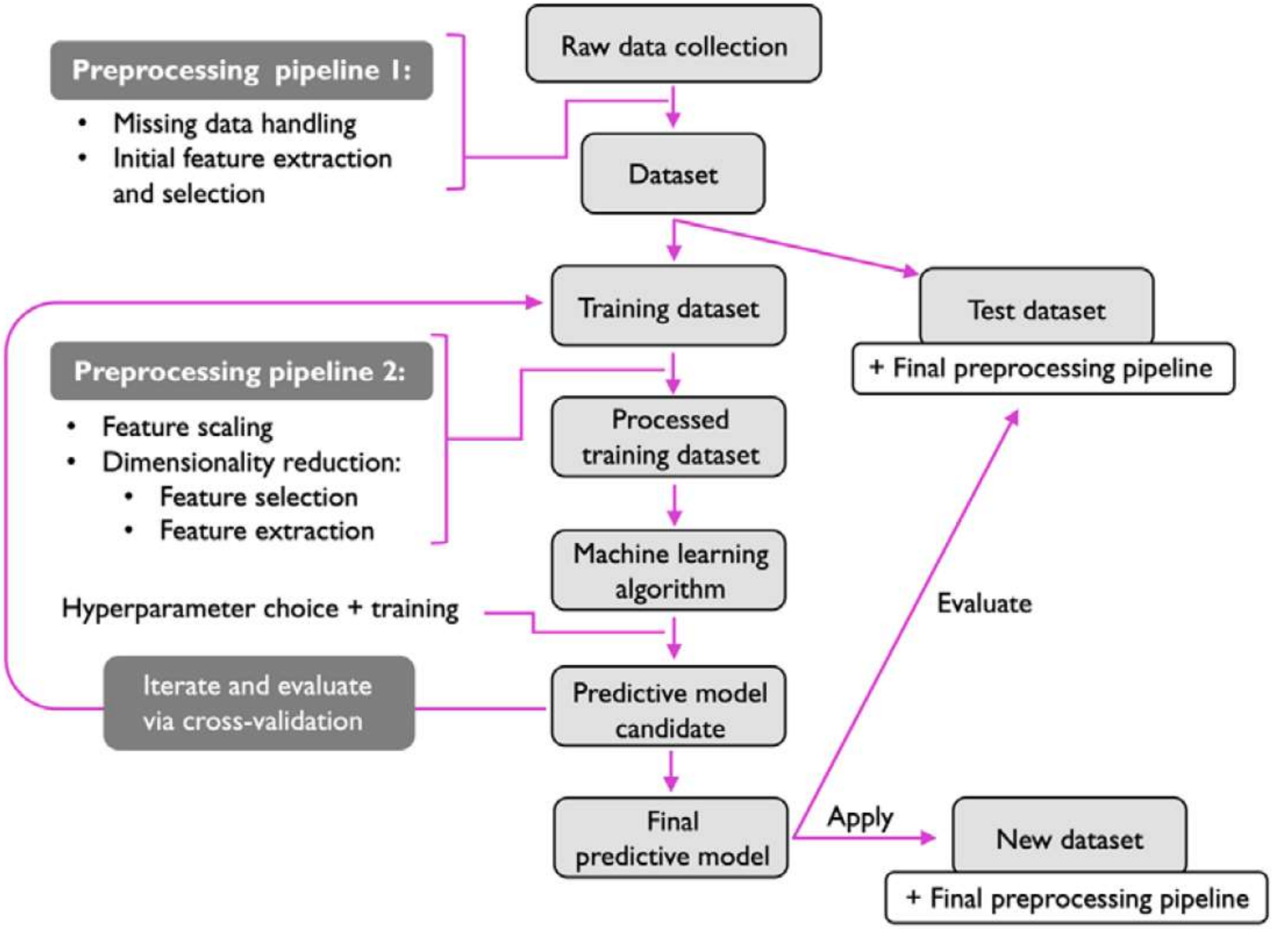
- Row vector: $x^{(i)} \in R^{1\times 4} = [x_1^{(i)}, x_2^{(i)}, x_3^{(i)}, x_4^{(i)}]$

- Column vector: $X^{(j)} \in R^{150\times 1} = \left[x_j^{(1)}, x_j^{(2)}, \cdots, x_j^{(150)}\right]^T = \begin{bmatrix} x_j^{(1)} \\ x_j^{(2)} \\ \vdots \\ x_j^{(150)} \end{bmatrix}$

Machine learning terminology

- Training example: A row in a table representing the dataset and synonymous with an observation, record, instance
- Training: Model fitting
- Feature, abbrev. x: A column in a data table or data matrix. Synonymous with predictor, variable, input, attribute, or covariate.
- Target, abbrev. y: Synonymous with outcome, output, response variable, dependent variable, (class) label, and ground truth.
- Loss function: measured for a single data point. Sometimes, also called a error function
- Cost function: The loss (average or summed) over the entire dataset

Machine learning in predictive modeling workflow



Preprocessing pipeline 1:
• Missing data handling
• Initial feature extraction and selection
Raw data collection
Dataset
Training dataset
Test dataset
+ Final preprocessing pipeline
Preprocessing pipeline 2:
• Feature scaling
• Dimensionality reduction:
• Feature selection
• Feature extraction
Processed training dataset
Machine learning algorithm
Hyperparameter choice + training
Iterate and evaluate via cross-validation
Predictive model candidate
Evaluate
Final predictive model
Apply
New dataset
+ Final preprocessing pipeline

Installing Python

- Anaconda: comes with many scientific computing packages pre-installed
- Miniconda: similar to Anaconda but without any packages pre-installed
- Miniforge: similar to Miniconda but community-maintained and uses a different package repository (conda-forge) from Miniconda and Anaconda

Install new Python packages

- `conda install SomePackage`
- `conda update SomePackage`
- `conda install SomePackage --channel conda-forge`
- `pip install SomePackage`

Packages for scientific computing, data science, and machine learning

- numpy
- scipy
- scikit-learn
- matplotlib
- pandas

Create and activate a virtual environment

- `conda create -n pyml python=3.9`
- `conda activate pyml`
- `conda deactivate`

References

- Sebastian Raschka, et al., *Giving Computers the Ability to Learn from Data*, In *Machine Learning with PyTorch and Scikit-Learn* (pp. 1–18), Packt Publishing, 2022. https://github.com/rasbt/machine-learning-book