Lights, Data, Action!

Intro to Data Science and ML in Python

Christopher Perez

CS50 Seminar Fall 2024

Outline ✅

1. What is data science?
2. What is machine learning?
3. Interactive example with real dataset
4. Looking ahead: resources and next steps

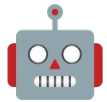
Data Science 

Data Science 

- The process of **gathering**, **analyzing**, and **interpreting** data to uncover patterns, make informed **decisions**, and predict future **outcomes**.
- A multidisciplinary field that combines **statistics**, **data analysis**, and ***machine learning***.
- Lots of real-world applications! Route optimization, revenue forecasting, election forecasting, etc…

Typical Data Science Workflow 🔄

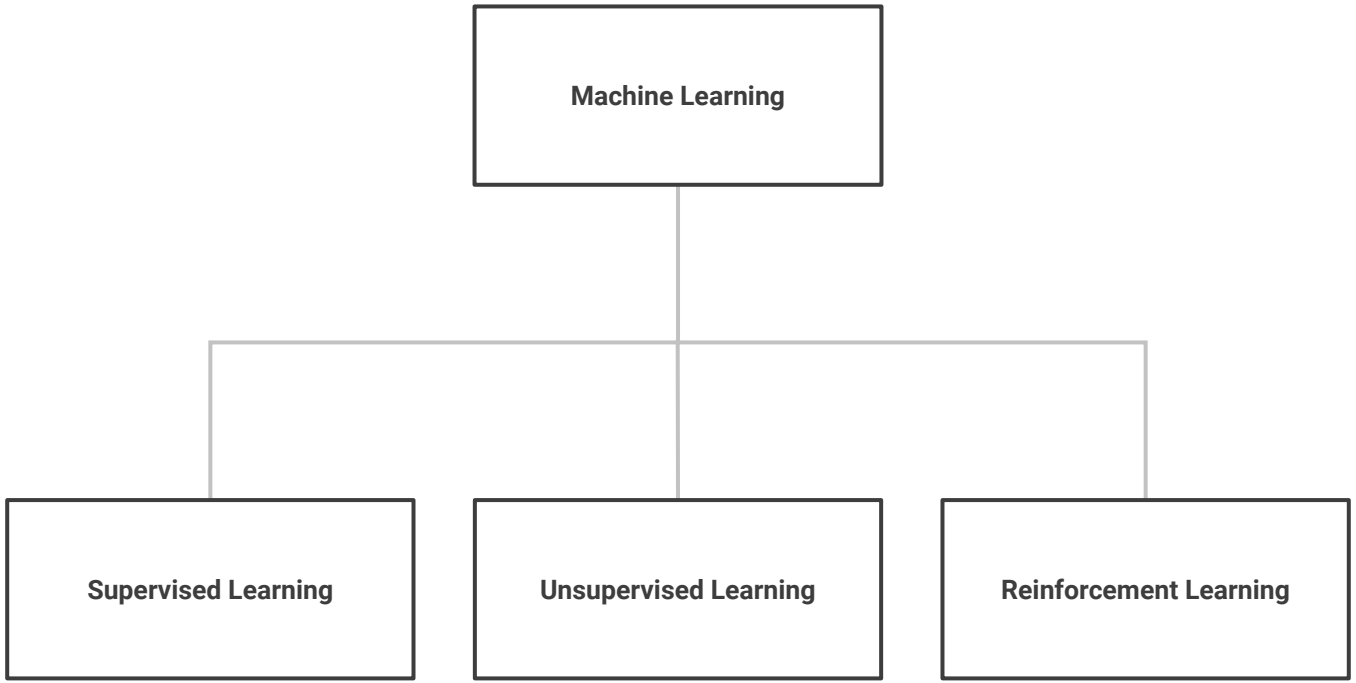
1. Define the problem
2. Data collection
3. Data cleaning and preparation
4. Exploratory data analysis (EDA)
5. Feature engineering
6. Model development
7. Model evaluation
8. Deployment
9. Monitoring and iteration

Machine Learning 

Machine Learning 🤖

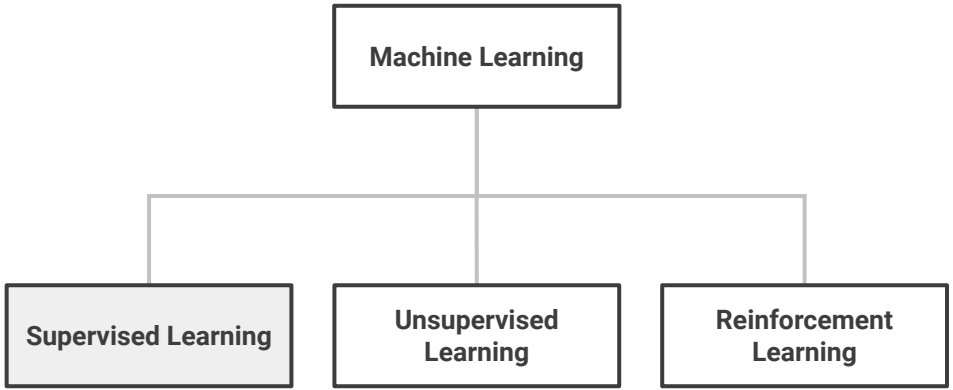
- **Goal**: Imitate the way that humans learn to gradually improve accuracy.
- The application of statistical, mathematical, and numerical techniques to derive some form of knowledge from data.
- Dependent on human intervention (e.g., determining the set of features, understanding data input)
- Powers many innovative technologies used today: personalized recommendations (e.g., Netflix, Spotify), fraud detection in banking, predictive healthcare analysis)

Machine Learning 🤖

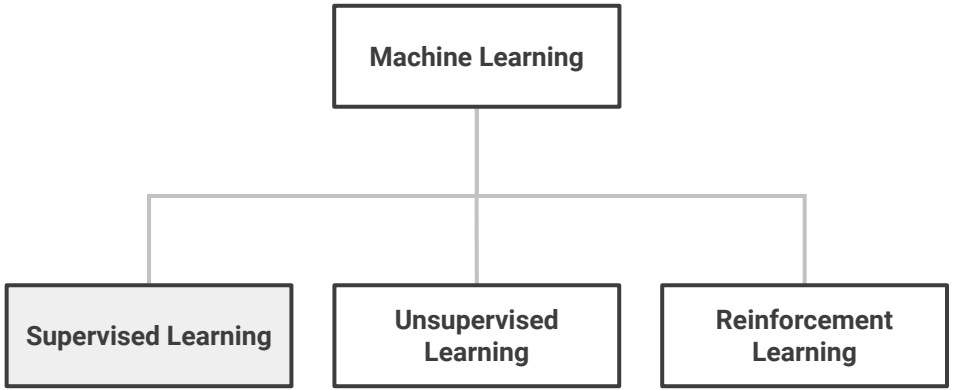


Machine Learning 🤖

Supervised Learning Overview

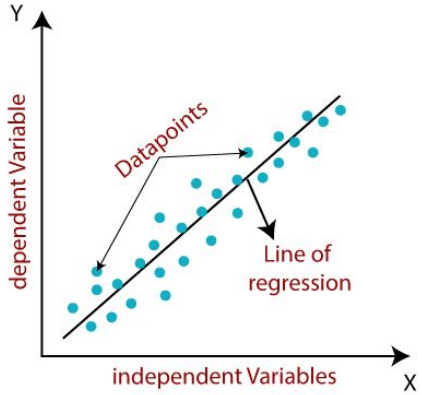


Machine Learning 🤖



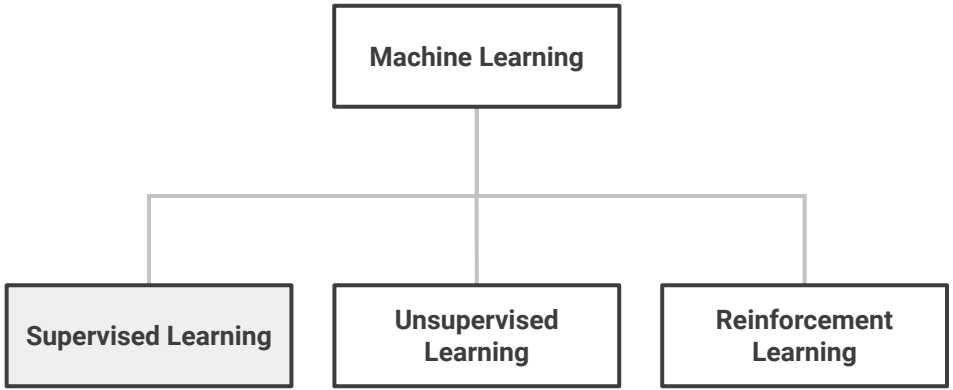
Supervised Learning Methods

1) Regression



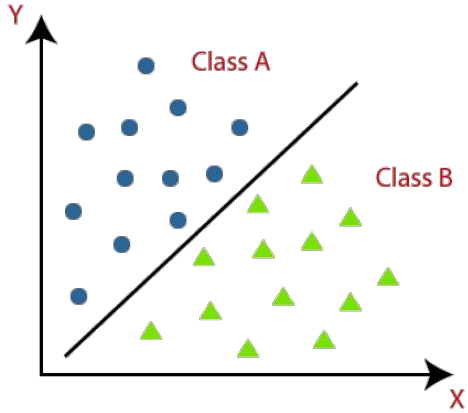
goal: minimize least squares loss

Machine Learning 🤖



Supervised Learning Methods

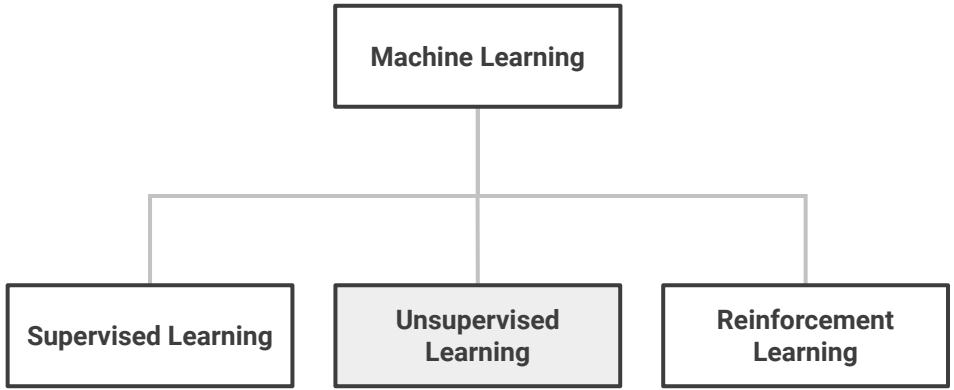
2) Classification



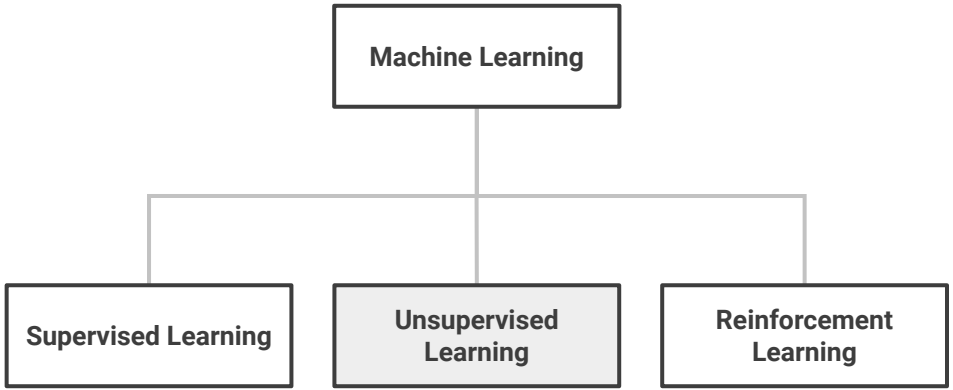
goal: 0/1 loss

Machine Learning 🤖

Unsupervised Learning Overview

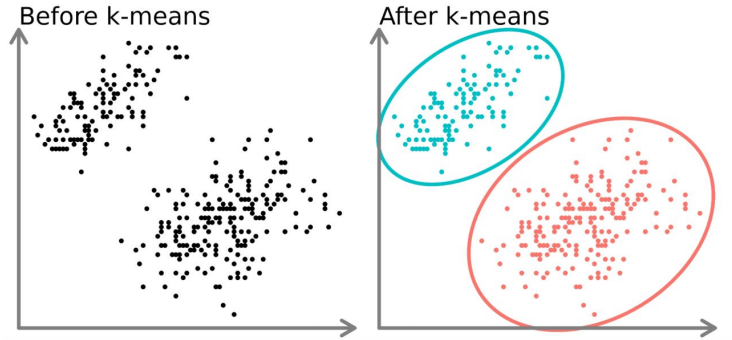


Machine Learning 🤖

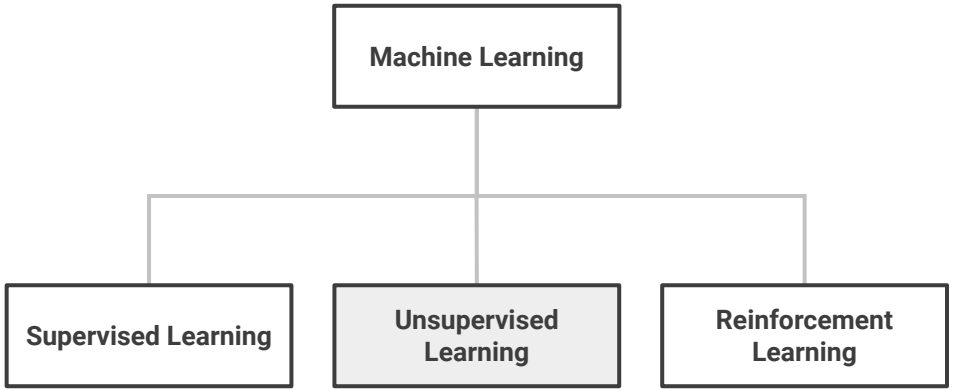


Unsupervised Learning Methods

1) Clustering

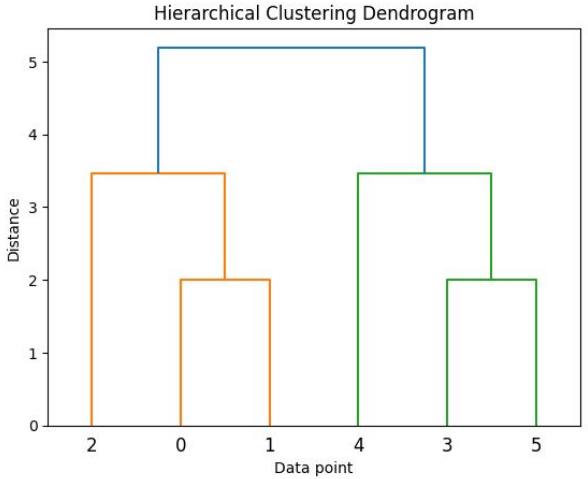


Machine Learning 🤖

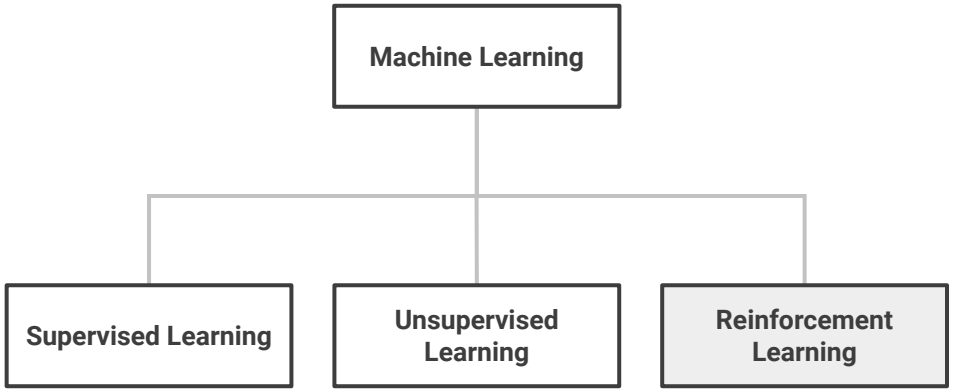


Unsupervised Learning Methods

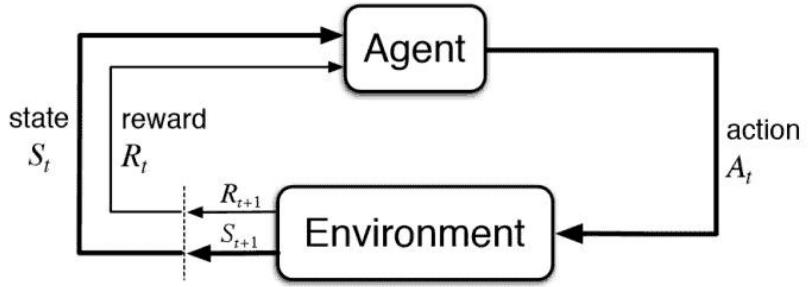
2) Hierarchical Agglomerative Clustering (HAC)



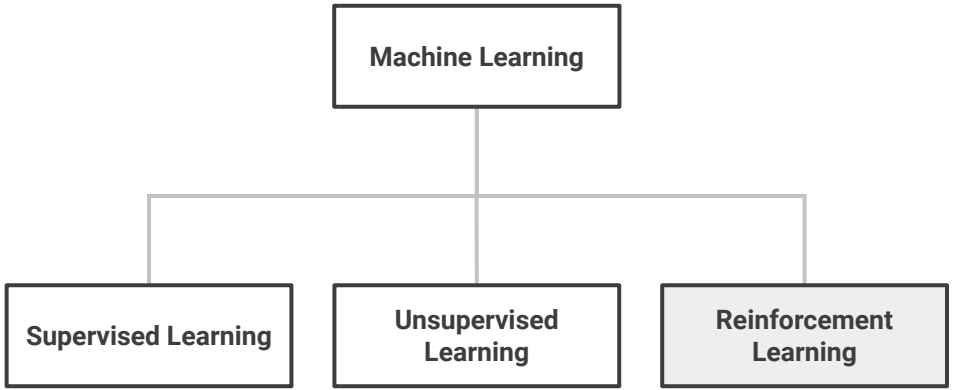
Machine Learning 🤖



Reinforcement Learning Overview

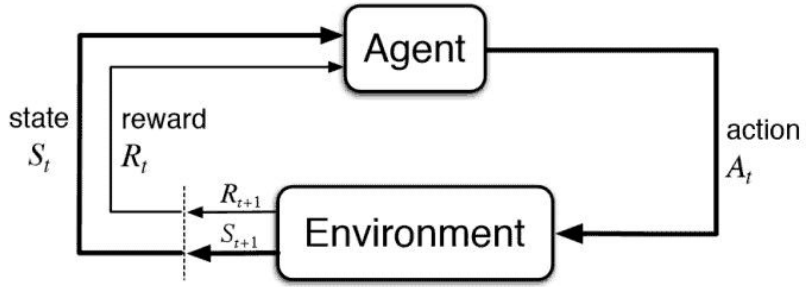


Machine Learning 🤖



Reinforcement Learning Methods

1) Q-Learning



Hands-on Example 🎵

(Spotify Most Streamed Songs)

Looking ahead 

Resources

- Pandas Documentation
  - Official Docs: https://pandas.pydata.org/docs/
  - Learn how to manipulate and analyze data effectively
- Scikit-Learn (sklearn) Documentation
  - Official Docs: https://scikit-learn.org/stable/user_guide.html
  - Explore tools for building and evaluating ML models
- Weights and Biases (W&B) Documentation
  - Official Docs: https://docs.wandb.ai/
  - A platform to track, visualize, and optimize ML experiments
- Interactive Learning Resources:
  - Kaggle Datasets and Tutorials: https://www.kaggle.com/

Thanks for Watching!