DSCI/CS 372(Winter 2024): Machine Learning for Data Science

Lecture 1: Introduction

Thanh H. Nguyen

Course Information

- Course website: https://classes.cs.uoregon.edu/24W/cs372m/
- Instructor: Thanh H. Nguyen (thanhhng@cs.uoregon.edu)
  - Office hour: Room 303 Deschutes, Wednesdays and Fridays 2:30 pm - 3:30 pm
- TA: Aliza Lisan (alisan@uoregon.edu)
  - Office hour: Mondays (2 pm – 4 pm) and Tuesdays (12 pm – 2 pm)
  - Room 207 Deschutes
- Coursework:
  - 3 programming projects: 42% (3 * 14% = 42%)
  - 4 written assignments: 28% (4 * 7% = 28%)
  - 1 final exam: 30%

Late Policy

- You can ask for one extension at most.*
- The earlier you ask, the better. Don't wait until the last minute.
- I will probably say yes.

- Send email to:
  - Instructor: thanhhng@cs.uoregon.edu
  - Email title: "DSCI/CS 372…"

Academic Honesty

Submit your own work:

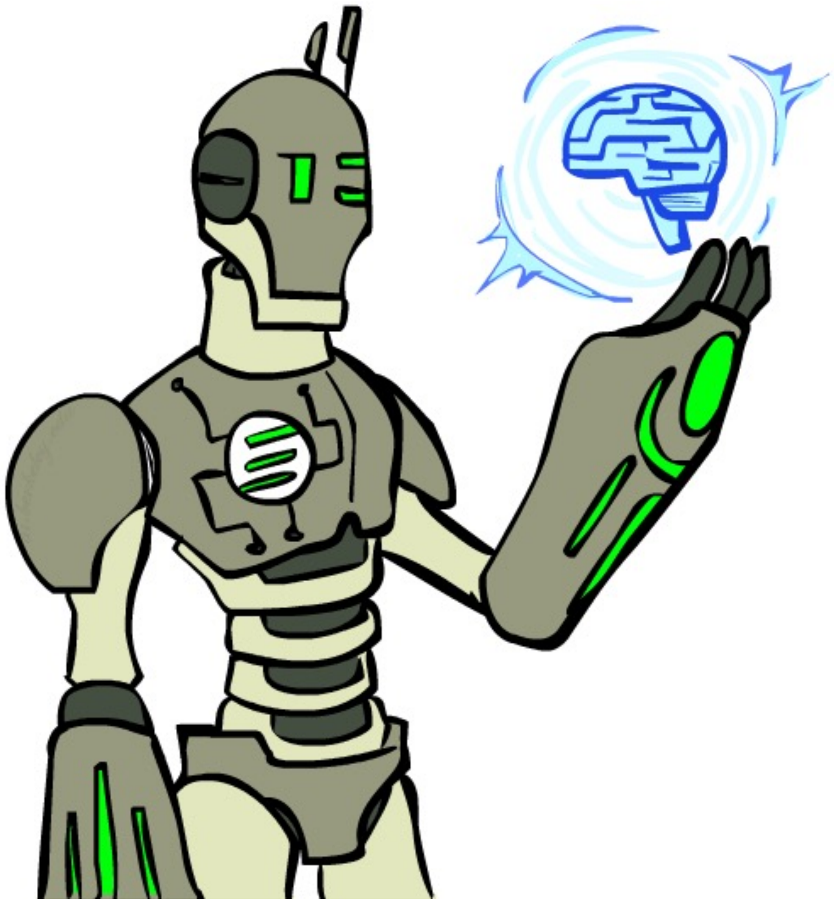
- Write up homework solutions individually

Follow rules for collaboration:

- No notes (written or electronic) from study groups
- Acknowledge all collaborations

Today: Introduction and Overview

- What is Machine Learning?
- What can Machine Learning do?
- What is this course?
- Data Preprocessing



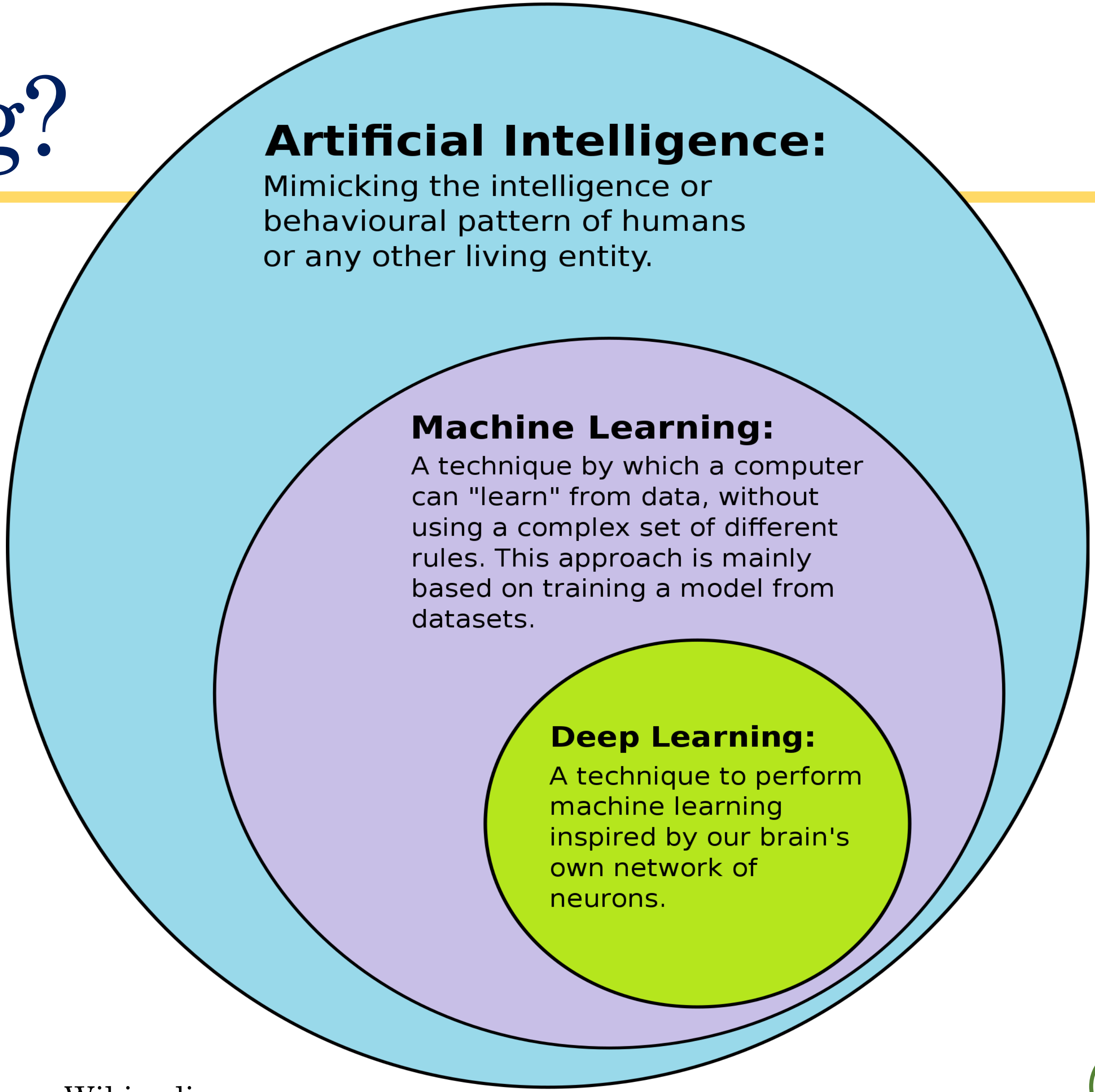
What is Machine Learning?

- "*Machine learning (ML) is the study of computer algorithms that can improve automatically through experience and by the use of data. It is seen as a part of artificial intelligence. Machine learning algorithms build a model based on sample data, known as training data, in order to make predictions or decisions without being explicitly programmed to do so.*" –Source: Wikipedia

- "*Machine learning is a branch of Artificial Intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.*" –Source: IBM

- "*Machine learning is a subfield of artificial intelligence, which is broadly defined as the capability of a machine to imitate intelligent human behavior … Machine learning starts with data — numbers, photos, or text, like bank transactions … From there, programmers choose a machine learning model to use, supply the data, and let the computer model train itself to find patterns or make predictions*" –Source: MIT

What is Machine Learning?



Source: Wikipedia

What is Machine Learning?

Data + Algorithms

- Find patterns
- Make predictions
- Provide suggestions
- …

Functions of a Machine Learning System

Descriptive	• The system uses the data to explain what happened
Predictive	• The system uses the data to explain what will happen
Prescriptive	• The system will use the data to make suggestions about what actions to take

Credit: T. Malone, D. Rus, R. Laubacher | MIT

Machine Learning Paradigms

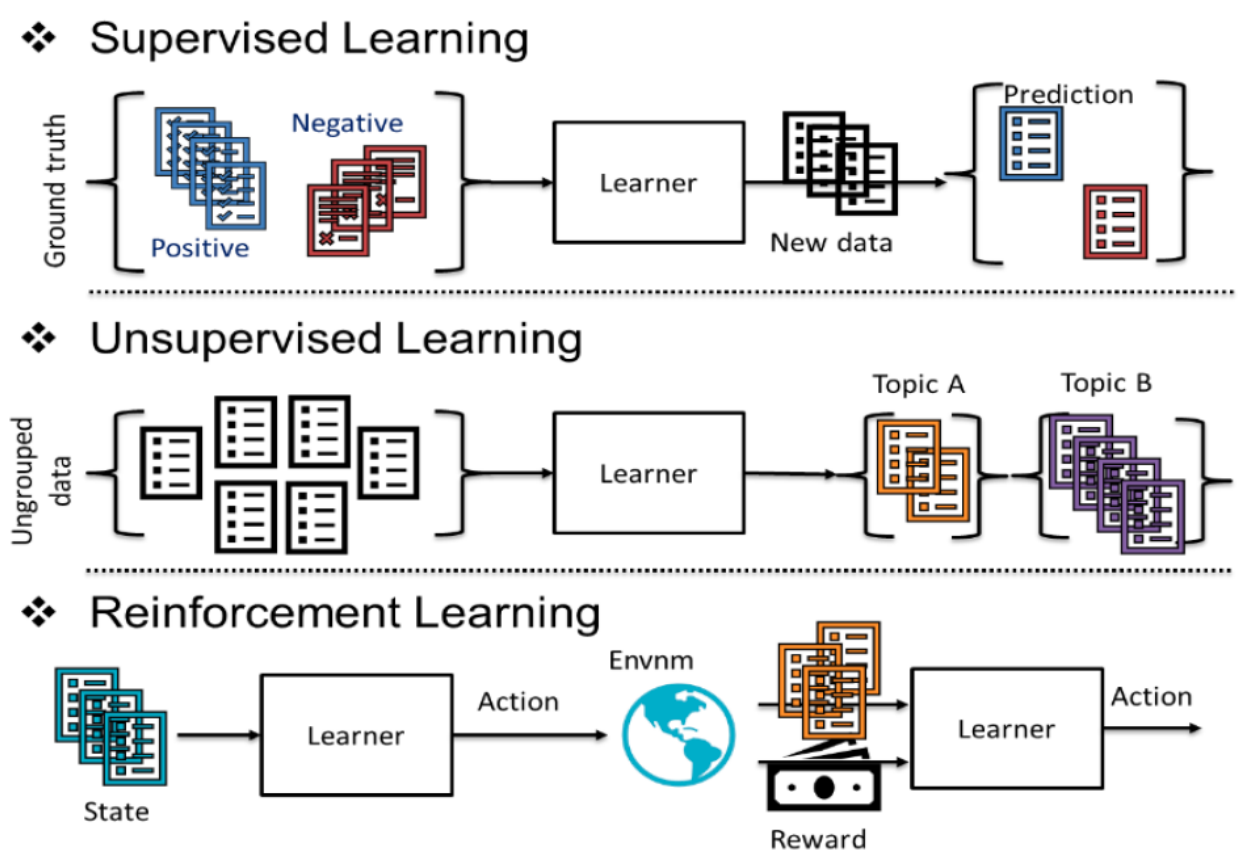
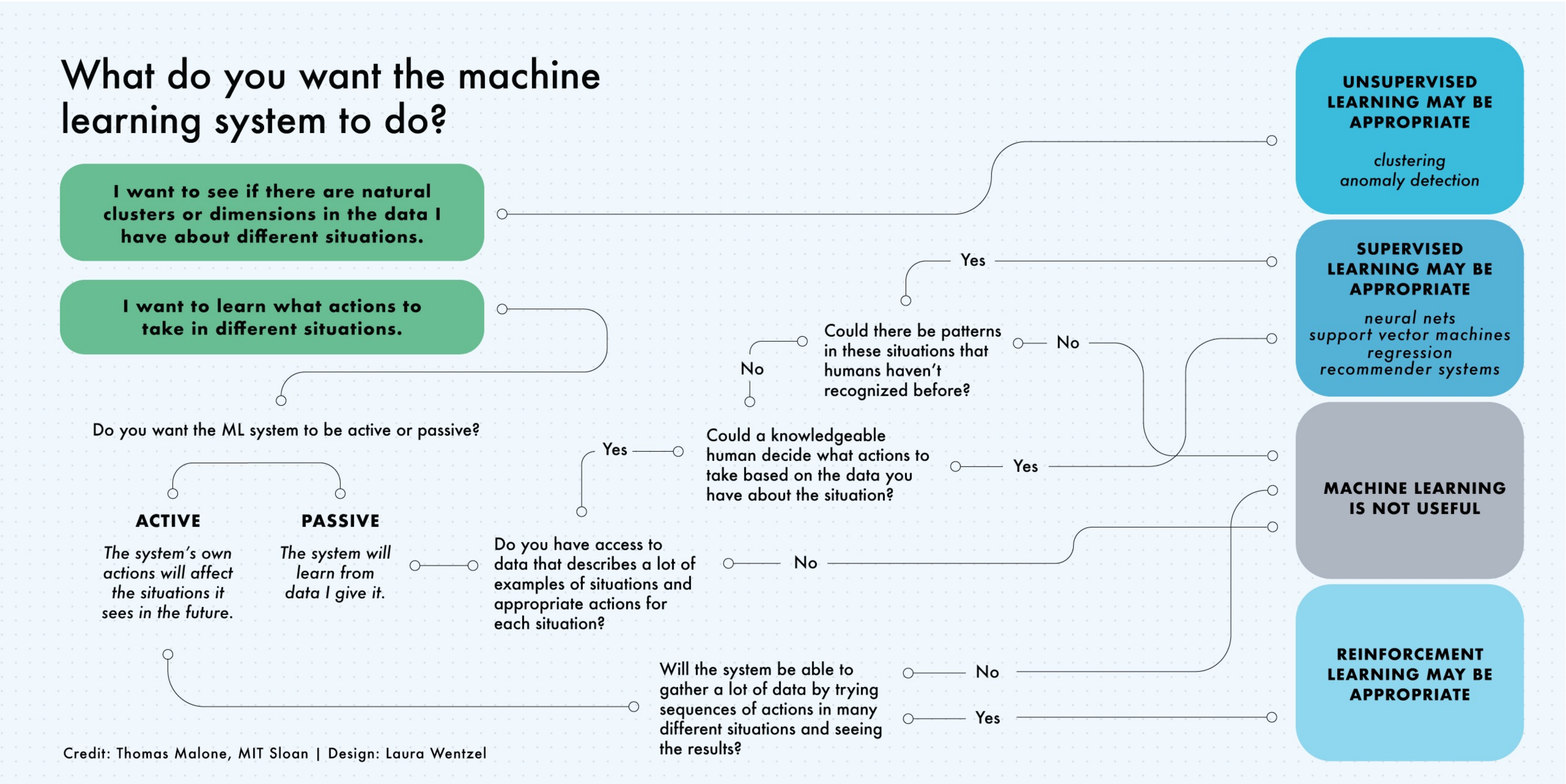


Figure 3: Main learning paradigms of Machine Learning.

- Supervised learning
  - Example: fraud detection, email spam detection, image classification, stock prediction
- Unsupervised learning
  - Example: Face recognition, network community detection
- Reinforcement learning
  - Robot navigation, gaming
- Others: active learning, online learning, etc.

Source: Adriano et al. "Algorithms in future capital markets." Available at SSRN 3527511 (2020).

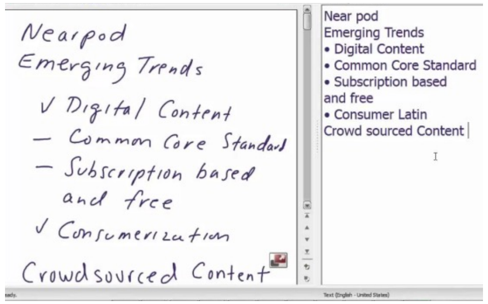


Source: https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained

Applications of Machine Learning

Personal Assistants

Recommendation Systems



Text Scanning

Advertising

Face Recognition

Anomaly Detection

Language Translation

Music Search

3D Modeling

Image Detection and Manipulation

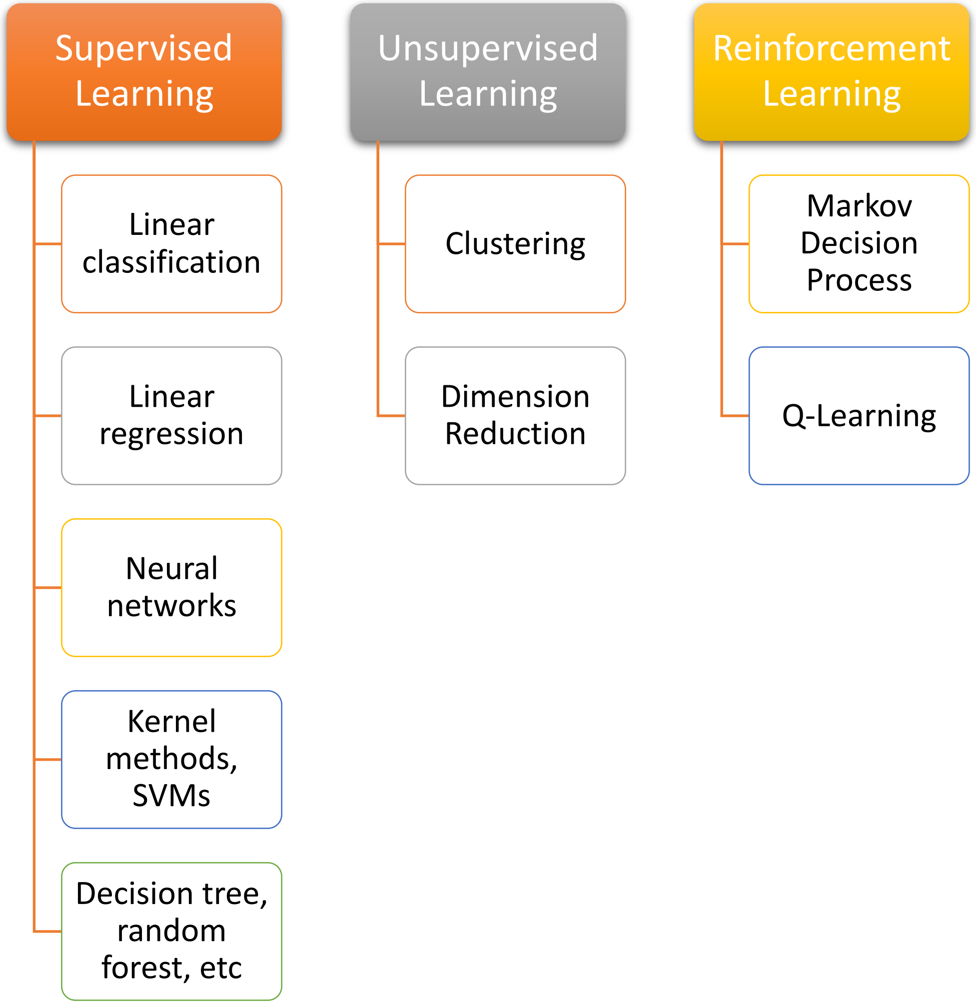
Speech to Text

Route Planning

What is This Course?

- Topics on Machine Learning

- Applications of Machine Learning



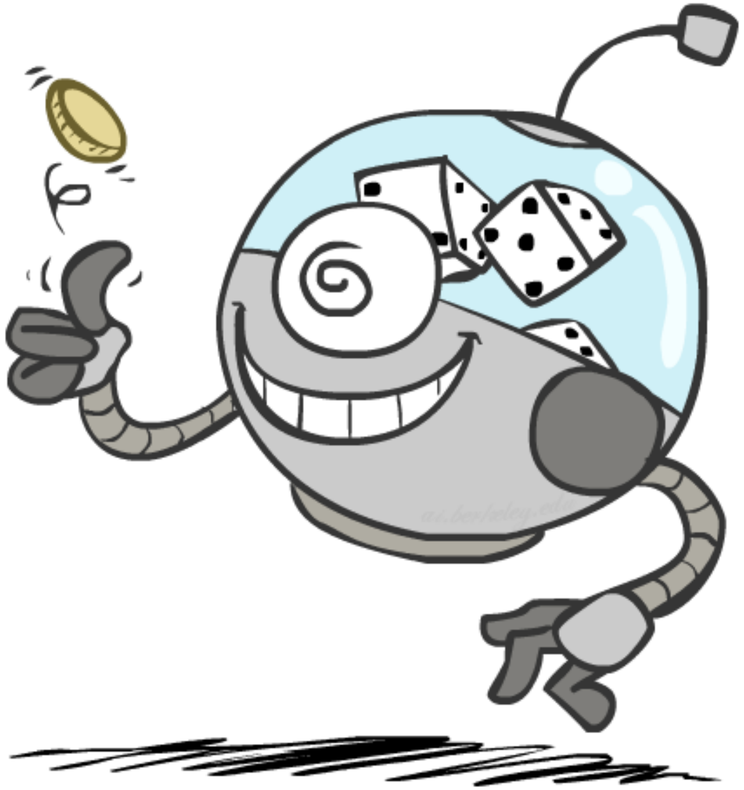
Data Preprocessing and Analysis

- Pandas:
  - Link: https://pandas.pydata.org/docs/getting_started/install.html
  - Data exploration and transformation: open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for Python programming language
- Scikit-learn:
  - Link: https://scikit-learn.org/stable/install.html
  - AI and machine learning: open source, BSD-licensed library providing simple and efficient tools for predictive data analysis (machine learning in Python)
- Matplotlib and seaborn
  - Matplotlib link: https://matplotlib.org/stable/users/installing/index.html
  - Seaborn link: https://seaborn.pydata.org/installing.html
  - Visualization: A library for creating static, animated, and interactive visualizations in Python
- Installation:
  - Recommendation: use Anaconda to install python, pandas, scikit-learn, and matplotlib
  - Conda is an open-source package and environment management system that runs on Windows, macOS, and Linux. Conda quickly installs, runs, and updates packages and their dependencies
  - Download Anaconda: https://www.anaconda.com/download

Recap: Random Variables

- A random variable is some aspect of the world about which we (may) have uncertainty
  - R = Is it raining?
  - T = Is it hot or cold?
  - D = How long will it take to drive to work?
  - L = Where is the ghost?

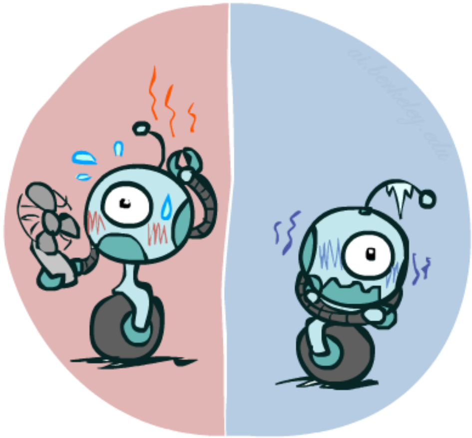
- Random variables have domains
  - R in {true, false}   (often write as {+r, -r})
  - T in {hot, cold}
  - D in $[0, \infty)$
  - L in possible locations, maybe {(0,0), (0,1), …}



Probability Distributions

- Associate a probability with each value

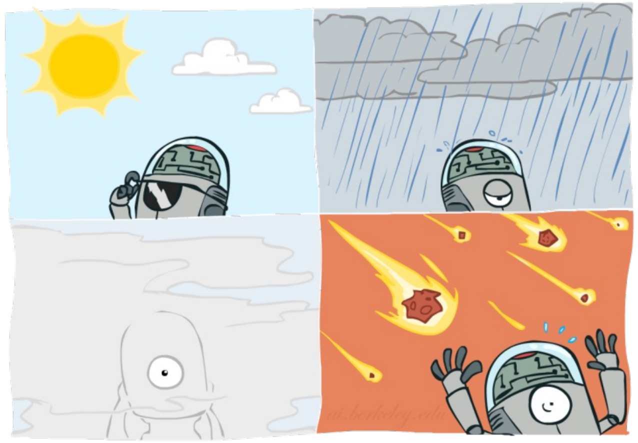
  - Temperature:



$P(T)$

T	P
hot	0.5
cold	0.5

  - Weather:



$P(W)$

W	P
sun	0.6
rain	0.1
fog	0.3
meteor	0.0

Probability Distributions

- Unobserved random variables have distributions

$P(T)$

T	P
hot	0.5
cold	0.5

$P(W)$

W	P
sun	0.6
rain	0.1
fog	0.3
meteor	0.0

Shorthand notation:

$$P(hot) = P(T = hot),$$
$$P(cold) = P(T = cold),$$
$$P(rain) = P(W = rain),$$
$$\ldots$$

OK *if* all domain entries are unique

- A distribution is a TABLE of probabilities of values

- A probability (lower case value) is a single number

$$P(W = rain) = 0.1$$

- Must have: $\forall x \;\; P(X = x) \geq 0$ and $\sum_x P(X = x) = 1$

Joint Distributions

- A *joint distribution* over a set of random variables: $X_1, X_2, \ldots X_n$ specifies a real number for each assignment (or *outcome*):

$$P(X_1 = x_1, X_2 = x_2, \ldots X_n = x_n)$$

$$P(x_1, x_2, \ldots x_n)$$

  - Must obey: $$P(x_1, x_2, \ldots x_n) \geq 0$$

$$\sum_{(x_1, x_2, \ldots x_n)} P(x_1, x_2, \ldots x_n) = 1$$

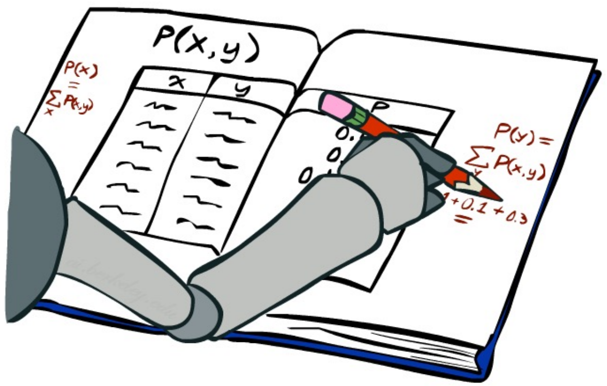
$P(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

- Size of distribution if n variables with domain sizes d?
  - For all but the smallest distributions, impractical to write out!

Marginal Distributions

- Marginal distributions are sub-tables which eliminate variables
- Marginalization (summing out): Combine collapsed rows by adding



$P(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$P(t) = \sum_s P(t, s)$$

$P(T)$

T	P
hot	0.5
cold	0.5

$$P(s) = \sum_t P(t, s)$$

$P(W)$

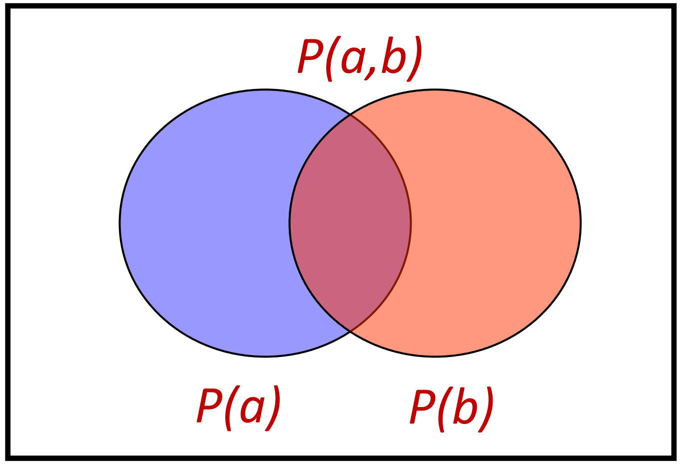
W	P
sun	0.6
rain	0.4

$$P(X_1 = x_1) = \sum_{x_2} P(X_1 = x_1, X_2 = x_2)$$

Conditional Probabilities

- A simple relation between joint and marginal probabilities
  - In fact, this is taken as the *definition* of a conditional probability

$$P(a|b) = \frac{P(a,b)}{P(b)}$$



$$P(T,W)$$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

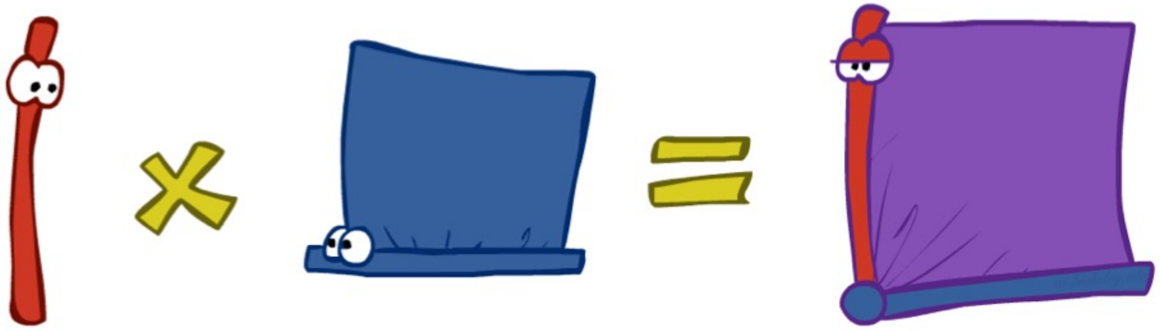
$$P(W=s|T=c) = \frac{P(W=s,T=c)}{P(T=c)} = \frac{0.2}{0.5} = 0.4$$

$$= P(W=s,T=c) + P(W=r,T=c)$$
$$= 0.2 + 0.3 = 0.5$$

The Product Rule

- Sometimes have conditional distributions but want the joint

$$P(y)P(x|y) = P(x,y) \quad \Longleftrightarrow \quad P(x|y) = \frac{P(x,y)}{P(y)}$$



The Product Rule

$$P(y)P(x|y) = P(x,y)$$

- Example:

$P(W)$

R	P
sun	0.8
rain	0.2

$P(D|W)$

D	W	P
wet	sun	0.1
dry	sun	0.9
wet	rain	0.7
dry	rain	0.3

$\Leftrightarrow$

$P(D,W)$

D	W	P
wet	sun	
dry	sun	
wet	rain	
dry	rain	

The Chain Rule

- More generally, can always write any joint distribution as an incremental product of conditional distributions

$$P(x_1, x_2, x_3) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2)$$

$$P(x_1, x_2, \ldots x_n) = \prod_i P(x_i|x_1 \ldots x_{i-1})$$

- Why is this always true?

Bayes Rule

- Two ways to factor a joint distribution over two variables:

$$P(x, y) = P(x|y)P(y) = P(y|x)P(x)$$

- Dividing, we get:

$$P(x|y) = \frac{P(y|x)}{P(y)}P(x)$$

- Why is this at all helpful?
  - Lets us build one conditional from its reverse
  - Often one conditional is tricky but the other one is simple

- In the running for most important AI equation!

That’s my rule!

Mean and Variance of Random Variables

- Mean: the expected value or mean is computed as:

$$\mu = E[X] = \sum_{x \in D} x \cdot P(X = x)$$

  - where $P(X = x)$ is the probability that variable X has value $x \in D$
- Alternatively, given samples $(x_1, x_2, \cdots, x_n)$, then

$$\mu = \frac{(x_1 + x_2 + \cdots + x_n)}{n}$$

Mean and Variance of Random Variables

- Variance: the variance of a random variable is the average of the squared deviations of the random variable from its mean

$$\sigma^2 = Var(X) = E[(X-\mu)^2] = \sum_x P(X=x)(x-\mu)^2$$

- Alternatively, given samples $(x_1, x_2, \cdots, x_n)$, then

$$\sigma^2 = \frac{1}{n}\sum_i (x_i - \mu)^2$$

Continuous Variables

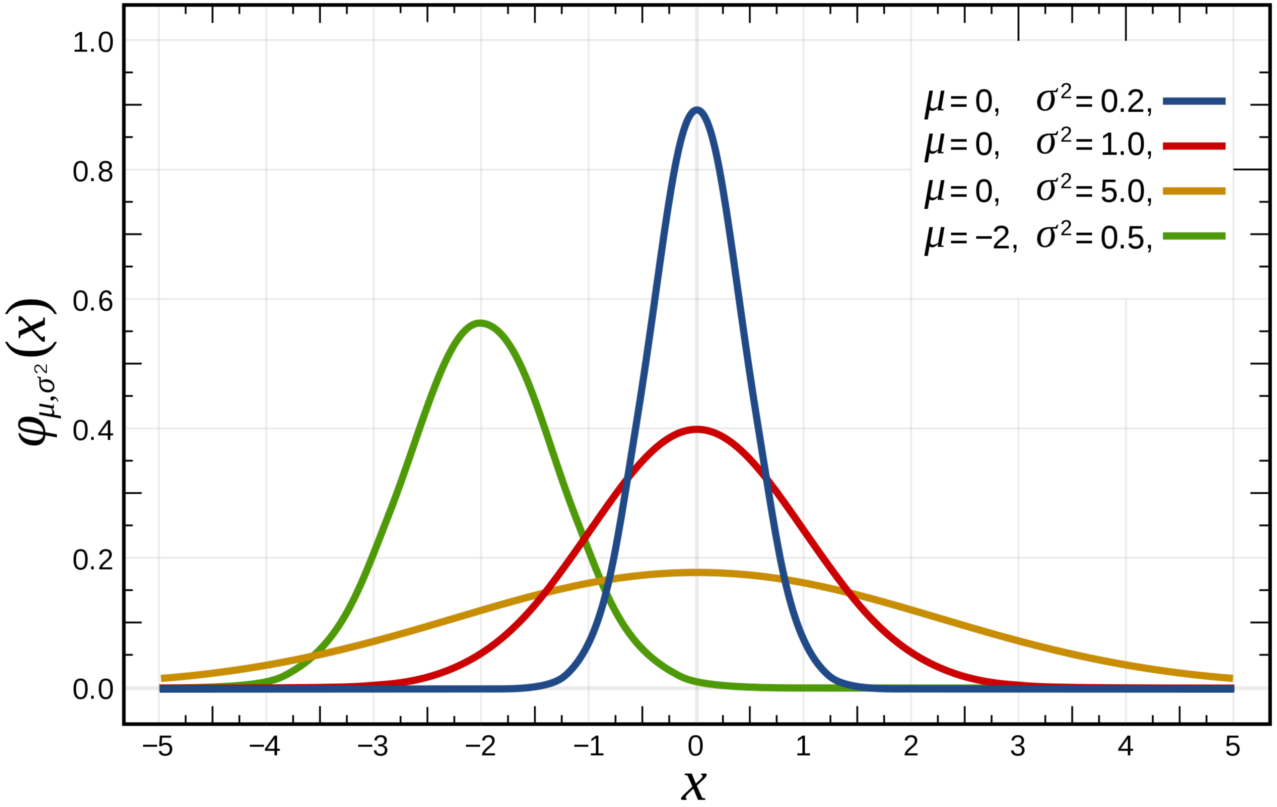
- A random variable X has a cumulative distribution function (CDF) $F(\cdot)$, which is a function from the sample space S to the interval $[0, 1]$
  - $F(x) = P(X \leq x)$ for any given $x \in S$
  - $0 \leq F(x) \leq 1$ for any $x \in S$ and $F(a) \leq F(b)$ for all $a \leq b$

- $F(\cdot)$ has an associated function f(·) that is referred to as a probability mass function (PMF) or probability density function (PDF)
  - PMF (discrete): $f(x) = P(X = x)$ for all $x \in S$
  - PDF (continuous): $\int_a^b f(x)dx = F(b) - F(a) = P(a < X < b)$

Example: Normal Distribution

- Mean and variance: (μ, σ^2)
- Probability density function:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



Graph source: Wikipedia