Warm-up

Course: SDS 323
Instructor: Nhat Ho
Teaching Assistant: Jiwon Kim

Spring 2022

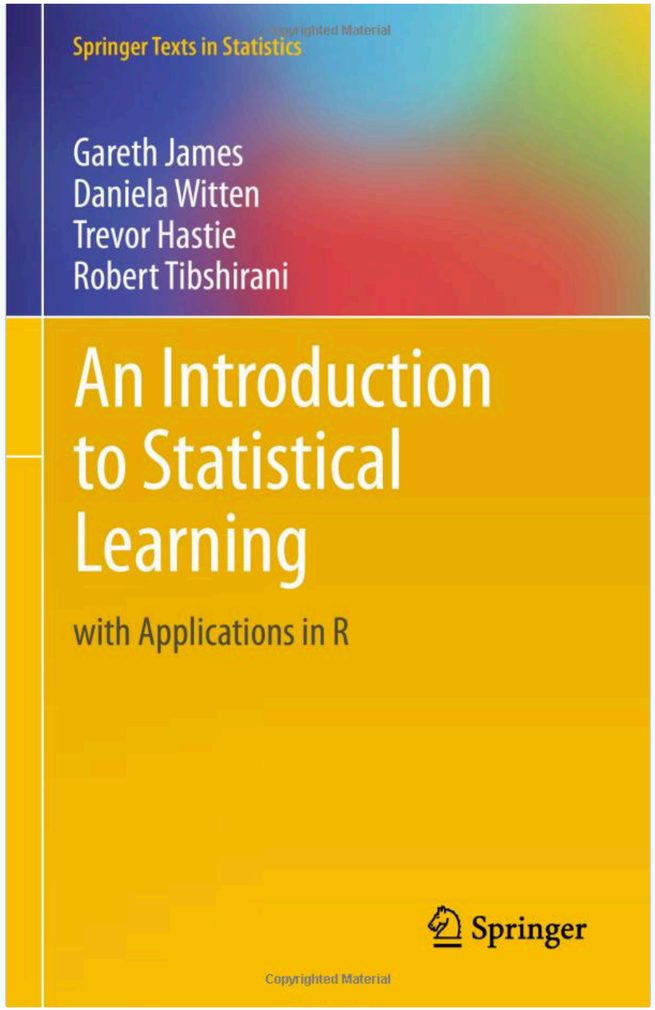
Overview

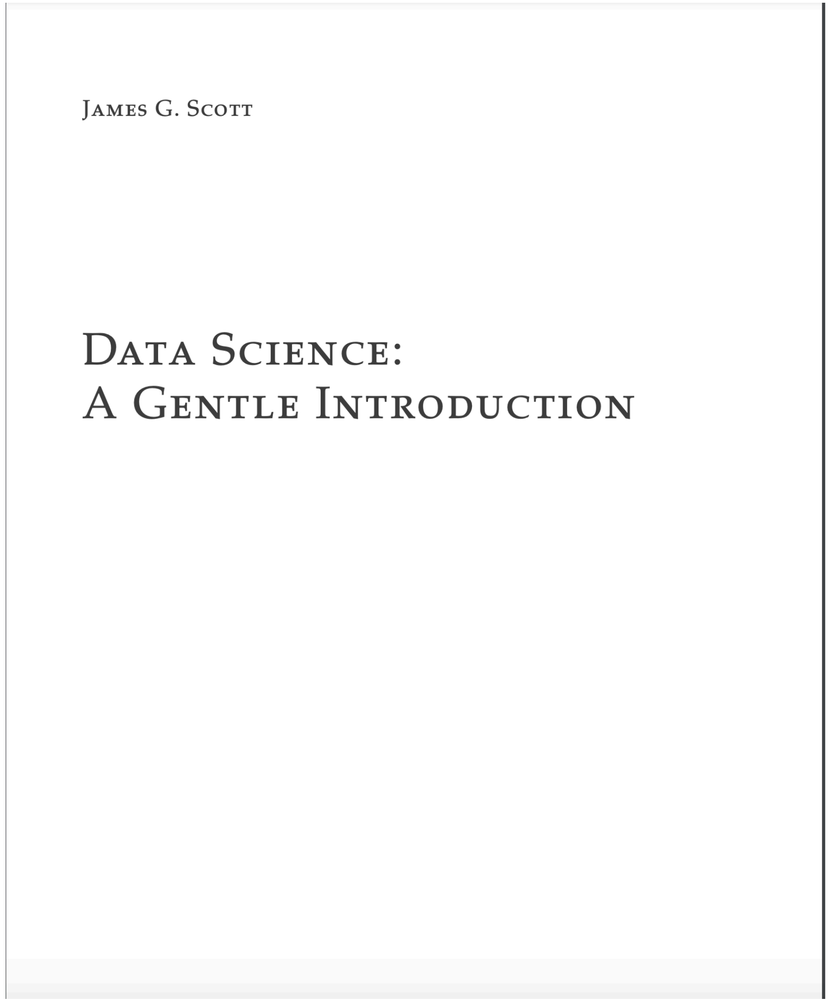
- This course is an introduction to Statistical Learning and Inference
- We study basic concept and tools and less on mathematical formulations
- There are 8 homework (every 1-2 weeks) and one final project (no midterm and final exams)
- You can work in a group of up to 4 people on the homework and final project
    - You can send me or TA members of your group by 01/30/2022
    - After that time, for students that do not have group, I will assign the group for you
- There are bonus questions during class
    - Each partially correct answer gives you one bonus point for the homework/ final project

Final Project

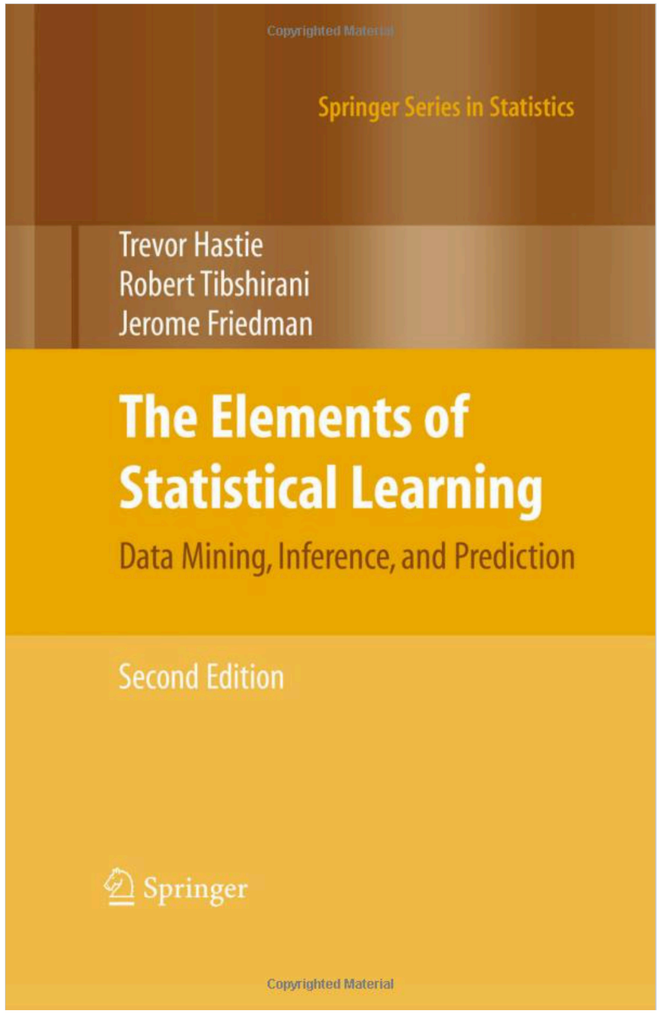
- The final project is due at the final day of the class
- The detailed format of the final project will be in the Canvas (around the end of January)
- For the final project, you are encouraged to choose your own data and analyze it
- You will need to submit the final project proposal around early April
    - Detailed timeline will be in the Canvas
- The evaluation of the final project is based on the technical correctness and quality of the analysis (figures/ plots, etc.)

Course Textbooks



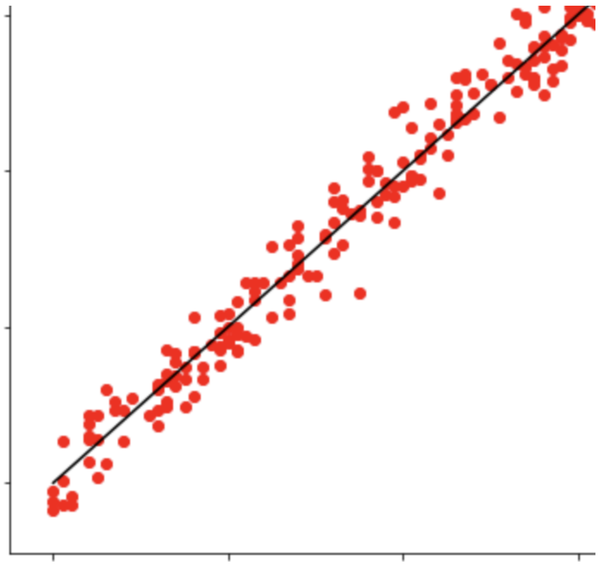


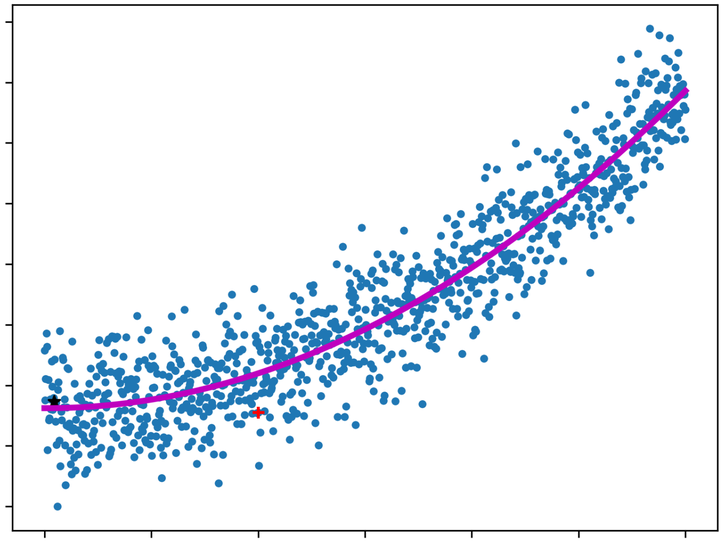
Optional Advanced Textbook



Course Structure: Supervised Learning

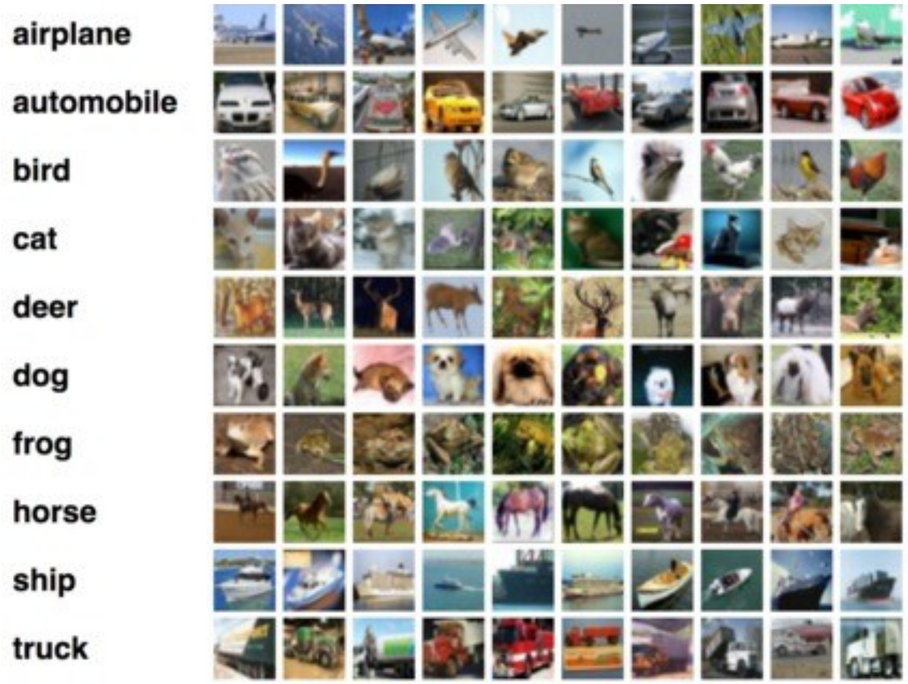
- For the first half of semester, we will focus on **supervised learning**
    - Linear regression (how can we predict the income based on the education?)



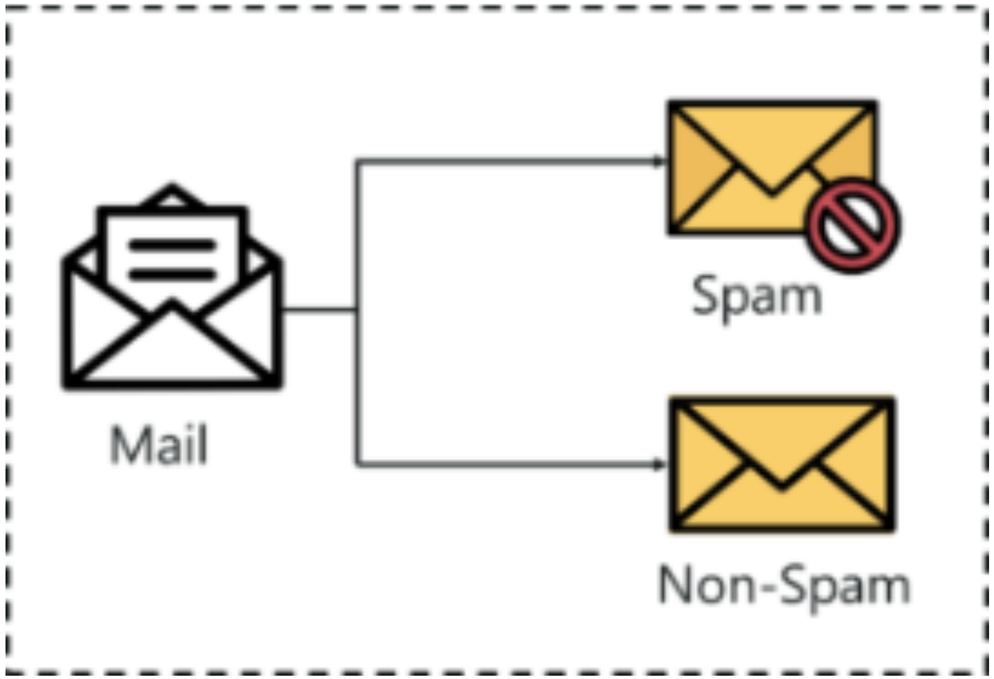


Course Structure: Supervised Learning

- Classification (given several images, can we train the machine to automatically classify the images into different groups, such as animals, trees, cars, etc.)



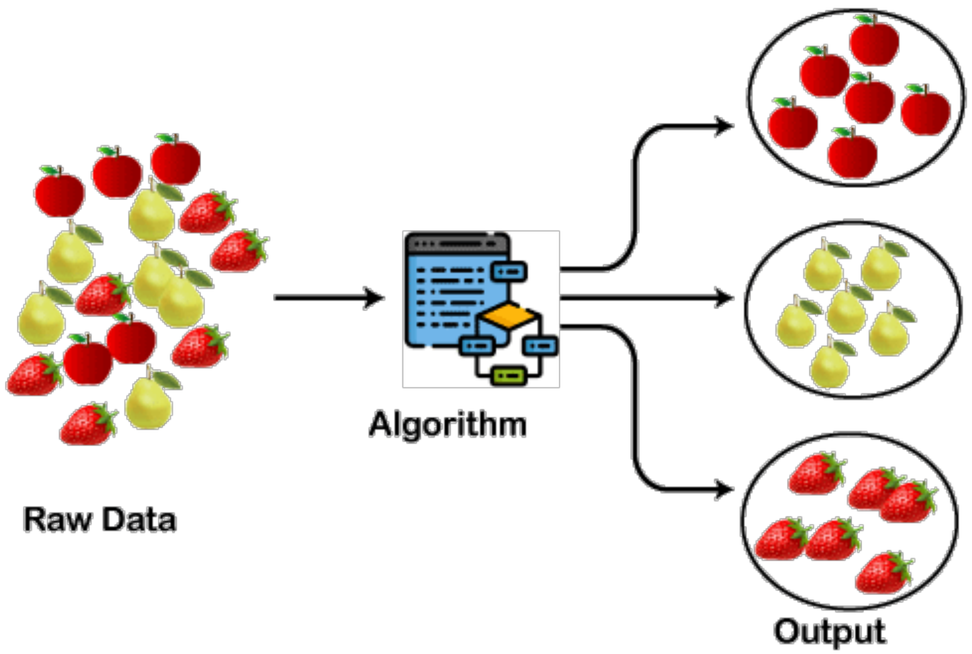
CIFAR10

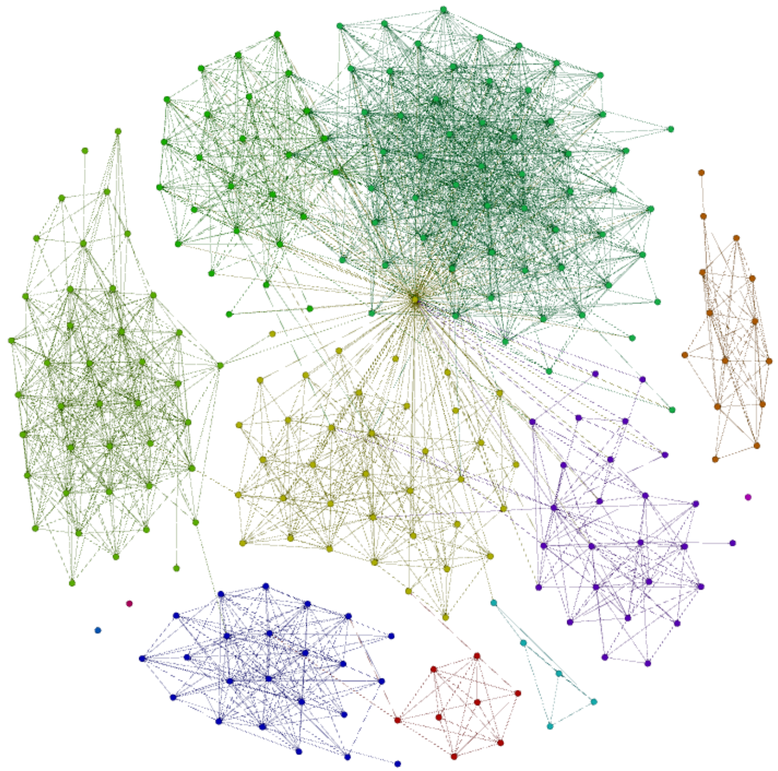


Images taken from web

Course Structure: Unsupervised Learning

- For the second half of semester, we will focus on **unsupervised learning**
    - Clustering (develop algorithms to partition objects in groups without knowing the labels of the objects, etc.)





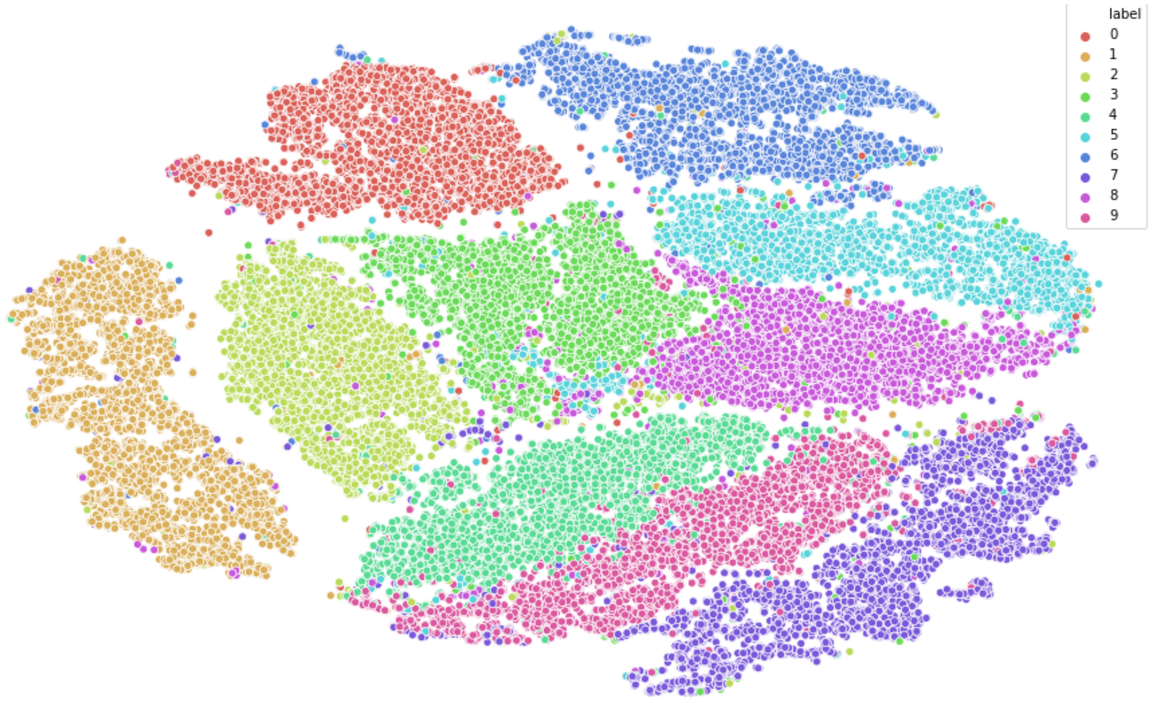
Facebook friendship network

Images taken from web

Course Structure: Unsupervised Learning

- Dimension reduction (Real-world data are very large-scale and high dimension. How can we reduce the dimension of data without losing too much data's information?)

MNIST



Images taken from web

Course Structure: Optional Topics

- If time permits, we will also cover **deep learning/ neural networks (Generative Models, Convolutional Neural Networks, etc.)**
    - Deep learning has been very powerful in computer vision and natural language processing. It can match human's ability in several computation tasks, such as image classification.

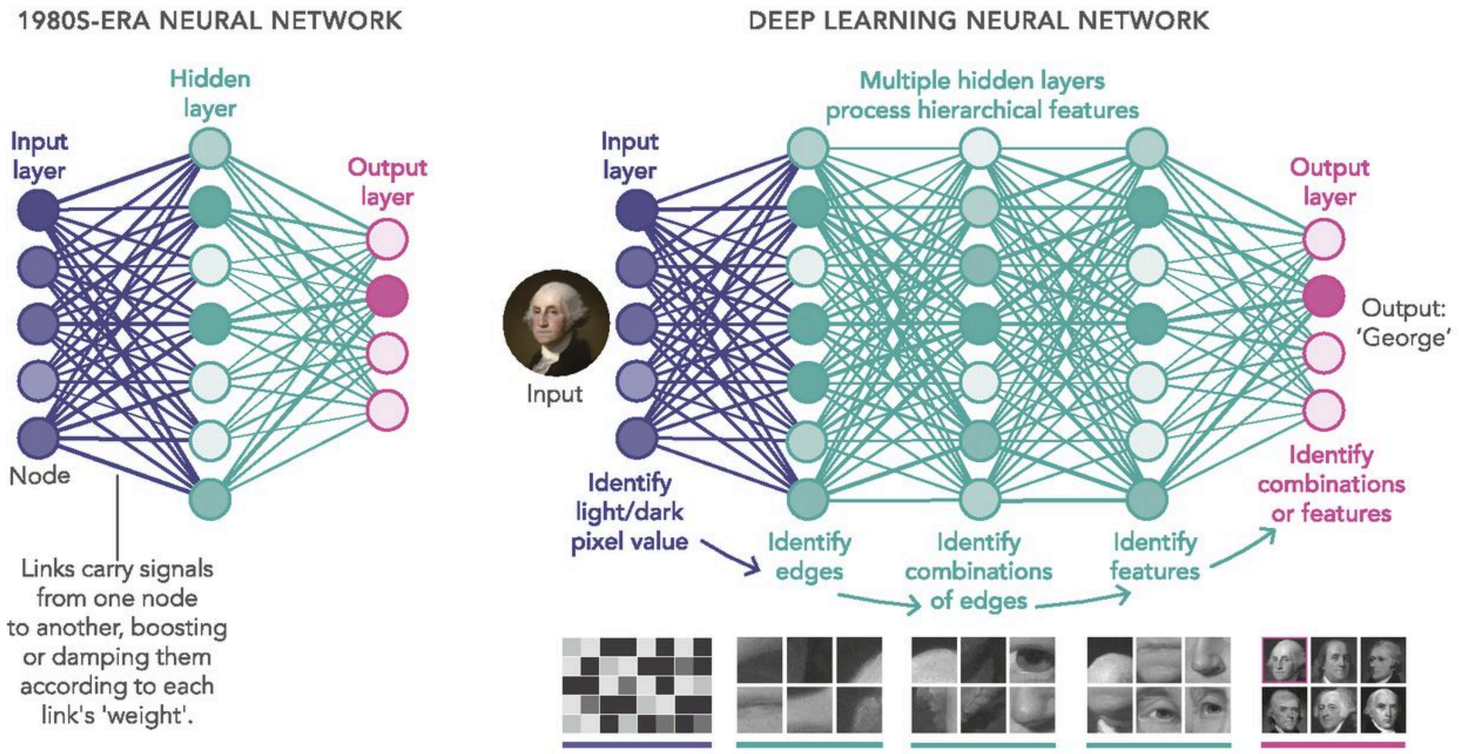
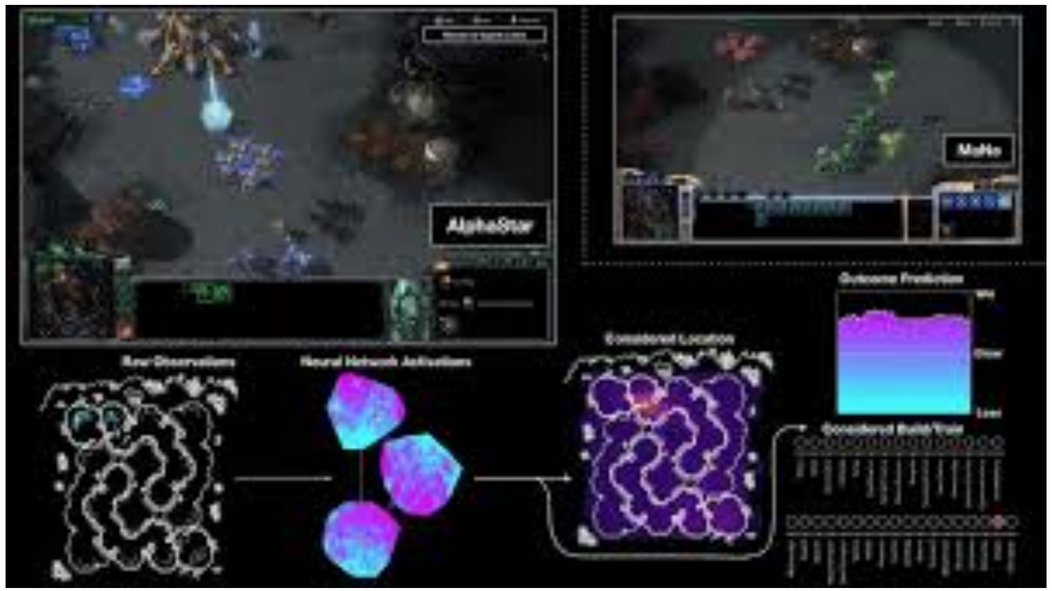


Image taken from Waldrop, PNAS (2019)

Course Structure: Optional Topics

- Reinforcement learning leads to several progress in self-driving/ autonomous driving (Tesla, Waymo, etc.)



Images taken from web

Main Elements of Data Science/ Machine Learning

- Regardless of the labels (Machine Learning/ Data Science or AI), in this course we study **useful concept/ techniques/ algorithms** that have been used successfully with real world data and applications
- There are practically 4 pillars of Machine Learning and Data Science:
    1. Data collection
    2. Data cleaning (preprocessing)
    3. Data analysis
    4. Data summary (Figures/ plots)
- In this course, we will mainly focus on “Data analysis” and “Data summary”

Main Steps of Data Analysis/ Summary

- Given the data, data analysis/ summary involves:

    1. **A (scientific) question**: What do we want to learn from the data?

    2. **Evidence**: well-developed techniques with a set of figures/ plots

    3. **Conclusion**

R Programming Language

- Main teaching programming language in this course is R

- A good introductory material for beginner of R is: https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf
    - This material provides basic commands/ functions to analyze data (how to read data, how to have some simple summary of data, etc.)

Course Data/ R Files

- I will post the data and R files in Canvas (for lecture notes/ homework)
- You are encouraged to play with them before/ after the class to have a good understanding of the concept/ techniques
- If you feel more convenient with Python/ C/ C++/ Java, etc., you can work with the data via these programming languages (There is no constraint on which programming language you can use for your homework/ final project)

Review of Probability

- Basic probability
- Conditional probability
- Bayes' Rule
- Rule of total probability
- Independence
- Discrete random variable
- Probability mass/ density function
- Expectation, Variance

Basic Probability

- Let A denote an event, such as “snowy day in Austin” or “coin lands heads”
- P(A): probability of the event A
  - P(“snowy day in Austin”) = 0.001
  - P(“coin lands heads”) = 0.5

Question: How to compute probability of an event?

Kolmogorov's Axioms

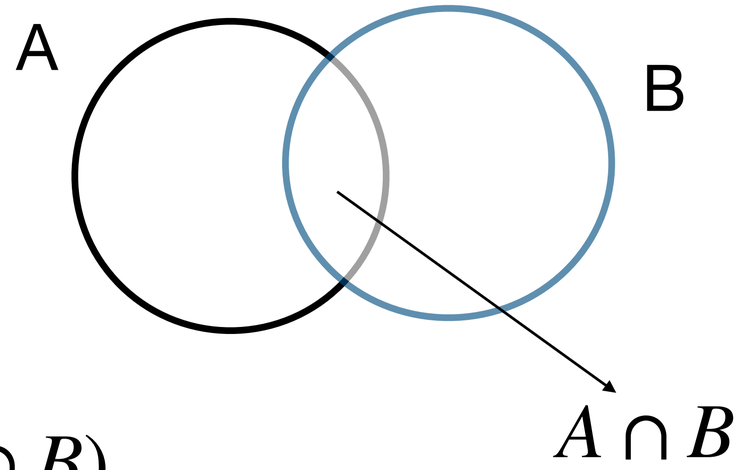
- Denote by Ω the set of all possible outcomes
- Ω = {heads, tails} (if we toss a coin); Ω = {1,2,3,4,5,6} (if we roll a dice)
- P(A): probability of some event A

1. **First axiom**: $P(A) \geq 0$ for all $A \subset \Omega$
2. **Second axiom**: $P(\Omega) = 1$ and $P(\varnothing) = 0$
3. **Third axiom**: If events A and B are disjoint, $P(A \cup B) = P(A) + P(B)$
   - $P(\{1,2\} \cup \{3,4,5\}) = P(\{1,2\}) + P(\{3,4,5\})$

Conditional Probability

- Conditional probability is the chance that one thing happens given that one other thing has already happened
- P(A| B): Probability that event A happens given that event B already happened
- The vertical bar means “given/ conditioned on”
  - P(“snowy day in Austin”| “the weather was below 40 F”) = 0.3
  - We can express that conditional probability as “The probability of snow in Austin today given that the weather was below 40 F is 30%”

Multiplication Rule



- Given two events A and B, we define

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

- $P(A \cap B)$: Joint probability that both events A and B happen
- $P(A \cap B) = P(A \mid B) \times P(B)$: multiplication rule - an axiom for conditional probability

Important Fact

- In general, we have $P(A \mid B) \neq P(B \mid A)$ when $P(A) \neq P(B)$
- It shows that the conditional probability is not symmetric
- An easy example is:
    - A = “I can play soccer”
    - B = “I am a professional soccer player”
- Obviously, $P(A \mid B) = 1$ (as a professional soccer player can certainly play soccer)
- However, $P(B \mid A)$ is nearly 0

Example of Conditional Probability

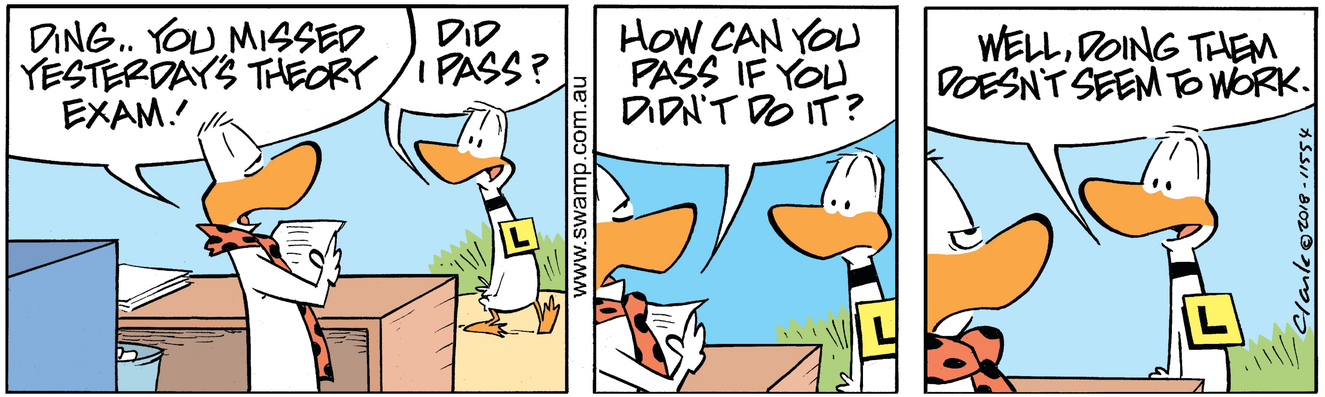
- 90% of students pass the final exam

- 80% of students pass both the final and the midterm exams

Question: What is the percentage of students who passed the final also passed the midterm exam?

Example of Conditional Probability

- Denote A = {pass the final exam}, B = {pass the midterm exam}
- P(A) = 0.9, $P(A \cap B) = 0.8$
- We need to compute P(B| A)



Example of Conditional Probability

- Recall that P(A) = 0.9, $P(A \cap B) = 0.8$
- Conditional probability:

$$P(B \mid A) = \frac{P(B \cap A)}{P(A)}$$

$$= \frac{0.8}{0.9} = \frac{8}{9} \approx 0.89$$

- **Conclusion**: The percentage of students who passed the final also passed the midterm exam is about 89%

Bayes' Rule: The Cornerstone of Data Science

- A is some event that we are interested in and B is data information
- Recall that: $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$ and $P(B \mid A) = \frac{P(A \cap B)}{P(A)}$
- It indicates that

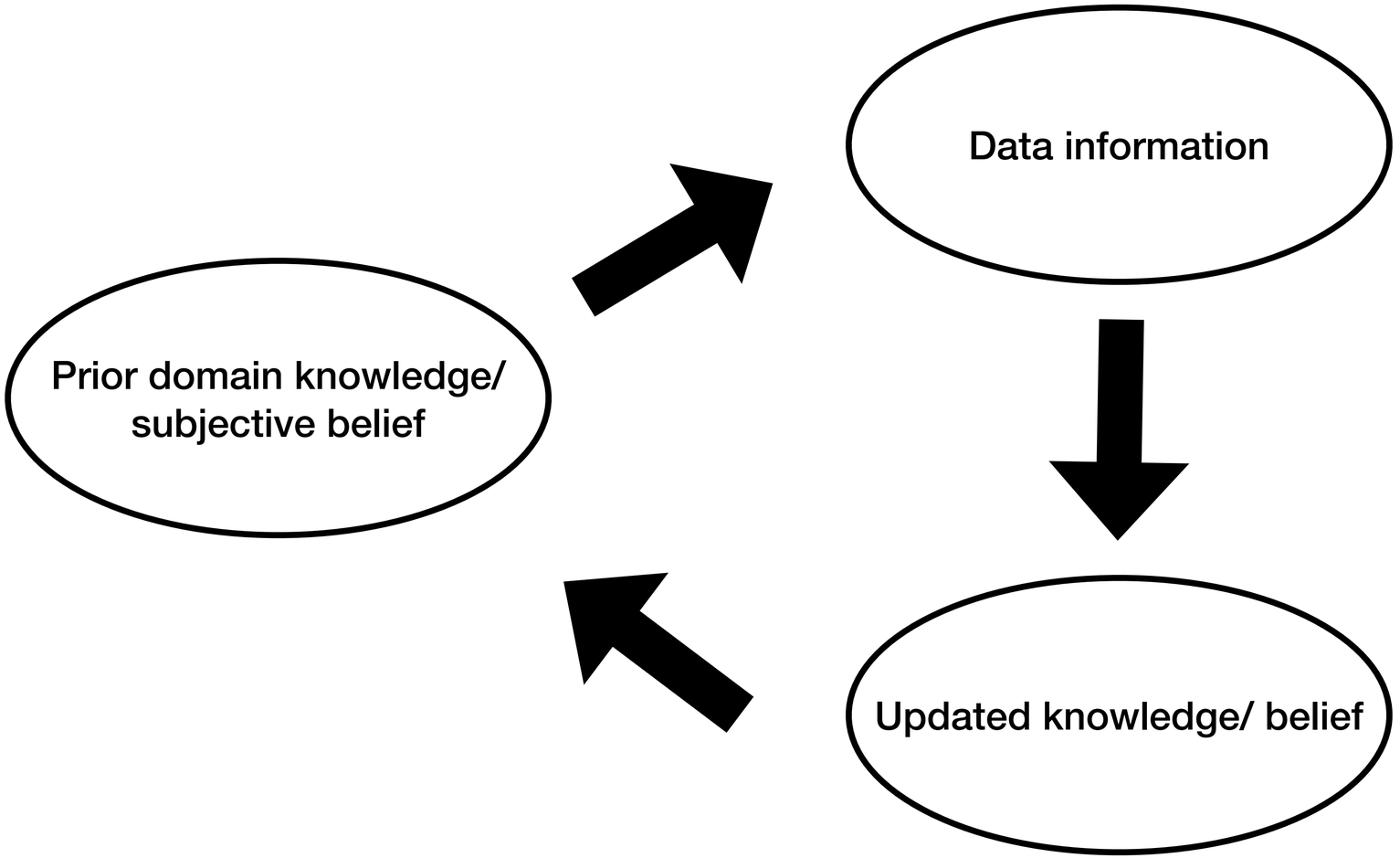
$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

- The above equation is called **Bayes' rule**, an important concept in data science

Bayes' Rule: Interpretation

- $P(A \mid B) = \dfrac{P(B \mid A)P(A)}{P(B)}$
- $P(A)$: the **prior probability**, namely, how probable the event A is before we have seen the data information B
- $P(B \mid A)$: the **likelihood**, namely, conditioned on event A, how likely is that we would see data information B
- $P(B)$: the **marginal probability** of B (regardless of whether event A holds or not)
- $P(A \mid B)$: the **posterior probability**, namely, how probable the event A after we have seen the data information B

Bayes' Rule: Diagram

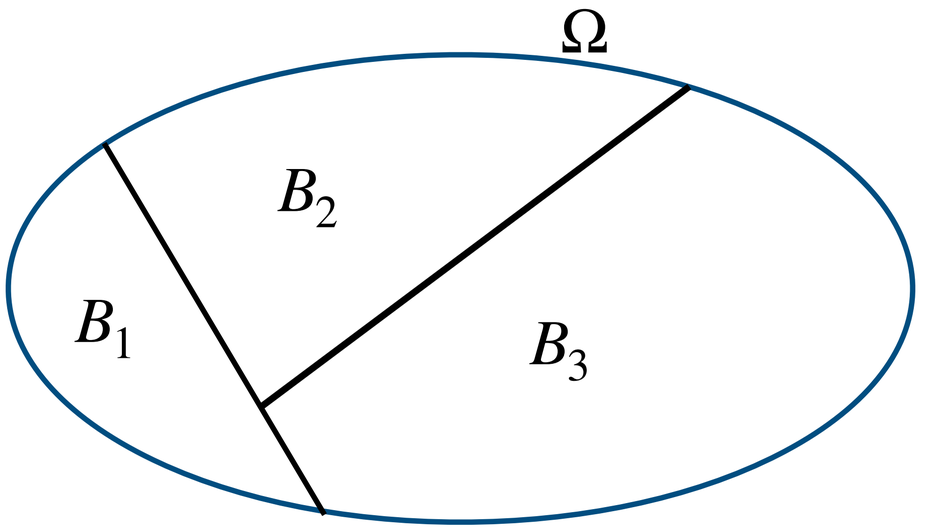


Bayes' Rule: Computation

- Bayes' rule: $P(A \mid B) = \dfrac{P(B \mid A)P(A)}{P(B)}$

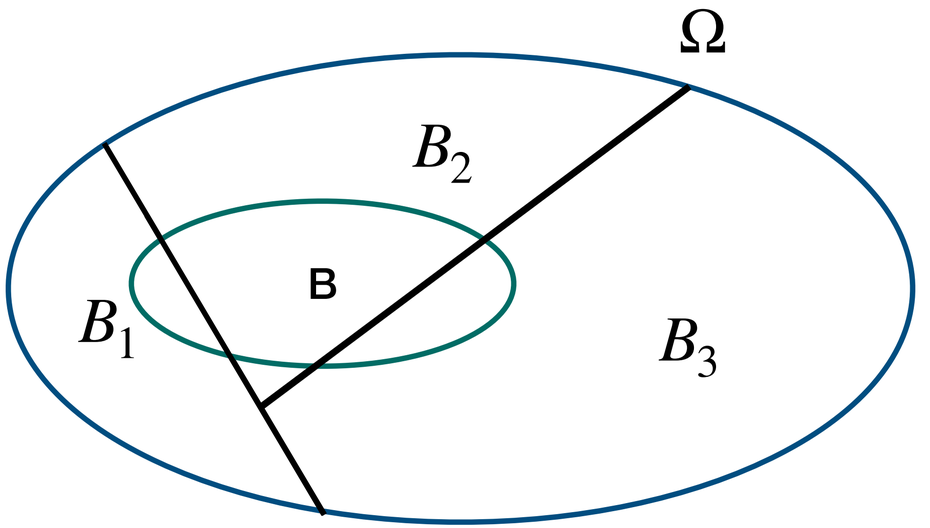
- In practice, the prior probability $P(A)$ and the likelihood $P(B \mid A)$ can be easily calculated based on how we model the data
- The challenge lies in calculating $P(B)$, which is usually **computationally intractable** in complicated problems
- A standard approach to compute $P(B)$ is through **law of total probability**

Law of Total Probability



- Recall that, Ω is the set of all possible outcomes
- Assume that $\Omega = B_1 \cup B_2 \cup \ldots \cup B_m$
- $B_1, B_2, \ldots, B_m$ are pair-wise disjoint events
  - $\Omega = \{1,2,3,4,5,6\}, B_1 = \{1,2\}, B_2 = \{3,4,5\}, B_3 = \{6\}$

Law of Total Probability



- Recall that, $\Omega = B_1 \cup B_2 \cup \ldots \cup B_m$
- **Law of total probability:**

$$P(B) = \sum_{i=1}^{m} P(B \cap B_i)$$

$$= \sum_{i=1}^{m} P(B \mid B_i) \times P(B_i)$$

for any event B

Example of Law of Total Probability: Simpson's Paradox

- We consider the following data on complication rates at a maternity hospital in Cambridge, England:

	Easy cases	Hard cases	Overall rates
Senior doctors	0.052	0.127	0.076
Junior doctors	0.067	0.155	0.072

- Senior doctors have lower complication rates in both easy and hard cases than junior doctors
- However, the overall complication rates of senior doctors are higher than those of junior doctors!!!
- This phenomenon is called **Simpson's paradox** and can be explained via law of total probability

Example of Law of Total Probability: Simpson's Paradox

	Easy cases	Hard cases	Overall rates
Senior doctors	0.052 (213)	0.127 (102)	0.076 (315)
Junior doctors	0.067 (3169)	0.155 (206)	0.072 (3375)

- B = {complication rates of senior doctors}
- $P(B) = P(B \mid \text{easy cases}) \times P(\text{easy cases}) + P(B \mid \text{hard cases}) \times P(\text{hard cases})$

$$= 0.052 \times \frac{213}{315} + 0.127 \times \frac{102}{315} = 0.076$$

Example of Law of Total Probability: Simpson's Paradox

	Easy cases	Hard cases	Overall rates
Senior doctors	0.052 (213)	0.127 (102)	0.076 (315)
Junior doctors	0.067 (3169)	0.155 (206)	0.072 (3375)

- B = {complication rates of junior doctors}
- $P(B) = P(B \mid \text{easy cases}) \times P(\text{easy cases}) + P(B \mid \text{hard cases}) \times P(\text{hard cases})$

$$= 0.067 \times \frac{3169}{3375} + 0.127 \times \frac{206}{3375} = 0.072$$

Example of Law of Total Probability: Simpson's Paradox

	Easy cases	Hard cases	Overall rates
Senior doctors	0.052 (213)	0.127 (102)	0.076 (315)
Junior doctors	0.067 (3169)	0.155 (206)	0.072 (3375)

- From the total law of probability, we observe that the reason for Simpson's paradox is:
  - The junior doctors mostly perform easy cases and those cases have small complication rates
  - The senior doctors have higher fraction of hard cases, which lead to high overall rates

Example of Bayes' Rule: Spam Emails

- $P(A \mid B) = \dfrac{P(B \mid A)P(A)}{P(B)}$
- 50% of emails are spam emails (Prior probability)
- A certain software can detect 99% of spam emails (the likelihood)
- The probability that a non-spam email is detected as spam email is 5%

Question: If an email is detected as spam email, what is the probability that it is in fact a non-spam email?

Example of Bayes' Rule: Spam Emails

- A = an event that an email is not spam
- $A^c =$ an event that an email is spam
- B = an event that an email is detected as spam
- We need to calculate $P(A \mid B)$
- **Hypothesis**: $P(A) = P(A^c) = 0.5$; $P(B \mid A^c) = 0.99, P(B \mid A) = 0.05$

Example of Bayes' Rule: Spam Emails

- Recall that: $P(A) = P(A^c) = 0.5$; $P(B \mid A^c) = 0.99, P(B \mid A) = 0.05$
- Bayes' rule: $P(A \mid B) = \dfrac{P(B \mid A)P(A)}{P(B)}$
- From law of total probability: $P(B) = P(B \mid A)P(A) + P(B \mid A^c)P(A^c) = 0.52$
- $P(A \mid B) = \dfrac{0.05 \times 0.5}{0.52} = \dfrac{5}{104} \approx 0.048$
- **Conclusion**: Given that an email is detected as spam email, the probability that it is in fact a non-spam email is 4.8%

Bayes' Rule in the Real World

- Bayes' rule has huge applications in practice:
    - Search engine (Google, Baidu, etc.)
    - Autonomous driving (Tesla, Waymo, etc.)
    - Computer vision, natural language processing (Alexa, Siri, etc.)
    - Reinforcement learning (AlphaGo, AlphaStar, etc.)
    - Medical testing

Independence

- In practice, we would like to determine whether two events A and B are “independent”
- Mathematically, two events A and B are **independent** if

$$P(A \mid B) = P(A \mid B^c) = P(A),$$

where B^c corresponds to the event that B does not hold.

- A few simple examples of independence:
  - P(“snowy day in Austin”| “I love music”) = P(“snowy day in Austin”)
  - P(“Coronavirus is over in 2021”| “Dog chases cat”) = P(“Coronavirus is over in 2021”)
- The independence means that events A and B do not convey information about each other

Independence

- Recall that, A and B are independent means $P(A \mid B) = P(A)$
- It suggests that $P(A \cap B) = P(A) \times P(B)$
- An application of conditional probability: $P(B \mid A) = \dfrac{P(A \cap B)}{P(A)} = P(B)$
- In practice, a simple way to check whether two events are independent is by checking the equation $P(A \cap B) = P(A) \times P(B)$

Conditional Independence

- Another useful notion of independence is **conditional independence**
- Assume that A, B, C are three events. We say that events A and B are **independent** given event C if

$$P(A, B \mid C) = P(A \mid C) \times P(B \mid C)$$

- This notion means that once we know event C, the events A and B do not convey information about each other
- A few important fact:
  - "Independence" $\neq$ "Conditional independence"
  - "Events A and B are independent" does not mean that "A and B are independent given event C"
  - Events A and B can be dependent and yet they are independent given an event C

Conditional Independence

- Consider the following events:
    - A = “Tiffany has black eyes”
    - B = “Stephanie has black eyes”
    - C = “Tiffany and Stephanie are sisters”
- It is clear that **A and B are independent events**, namely, the fact that Tiffany has black eyes does not convey any information about the event that Stephanie has black eyes
- However, given the information that Tiffany and Stephanie are sisters, the event that Tiffany has black eyes does imply the event that Stephanie has black eyes as well. Therefore, the events **A and B are conditionally dependent given event C**

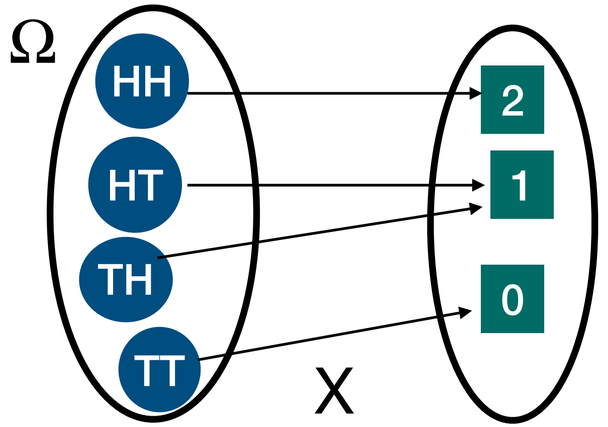
Conditional Independence

- Consider another example: Bryan and his wife commute to work at the same time via the metro
    - A = “Bryan is late for work”
    - B = “His wife is late for work”
    - C = “The weather is fine and the metro is on time”
- Since Bryan is late for work, it is possible that the weather is bad (too much rain, etc.) or the metro is late. Given that, his wife is also likely to be late for work too. It demonstrates that **events A and B are dependent**
- If we know that the weather is fine and the metro is on time and Bryan is late for work, then it is hard to know why his wife is late for work. It shows that **events A and B are conditionally independent given event C**

Conditional Independence

- Other simple examples of conditional independence:
    - P(“lung cancer”, “yellow teeth”| “smoking”) = P(“lung cancer”| “smoking”) P(“yellow teeth”| “smoking”)
    - *Markov’s assumption*: P(“future”, “past”| “present”) = P(“future”| “present”) P(“past”|”present”)
- Conditional independence is very important in machine learning and data science as it greatly **reduces the complexity of the models**

Discrete Random Variable



- Ω: a set of possible outcomes
- A **discrete random variable** X is a function from Ω to a set of finite values
- A simple example:
  - Toss a fair coin twice and $\Omega = \{HH, HT, TH, TT\}$ where H = “head”, T = “tail”
  - X: number of heads
  - X({HH}) = 2, X({HT}) = X({TH}) = 1, X({TT}) = 0

Discrete Random Variable

- P(X = 2) = P({HH}) = 1/ 4 (as only one event {HH} has two heads)
- P(X = 1) = P({HT}, {TH}) = 2/ 4 (two events {HT} and {TH} have one head)
- P(X = 0) = P({TT}) = 1/ 4 (only one event {TT} has no head)

Beyond Discrete Random Variable

- A few remarks:
    - When X has infinite values and not countable (imagine the height or weight of people), the random variable X is called **continuous random variable**
    - Discrete and continuous random variables are important notion to define **probability mass function/ probability density function**

Probability Mass Function

- Discrete random variable X takes values in $\{x_1, \ldots, x_n\}$
- Define $p(x_1) = P(X = x_1), p(x_2) = P(X = x_2), \ldots, p(x_n) = P(X = x_n)$
- The function p is called the **probability mass function**

- Recall the previous example where we toss a coin twice:
  $\Omega = \{HH, HT, TH, TT\}$, the number of heads X takes values in {0, 1, 2}
- p(0) = P(X = 0) = 1/4, p(1) = P(X = 1) = 1/ 2, p(2) = P(X = 2) = 1/4
- It is not hard to check that p(0) + p(1) + p(2) = 1

Probability Mass Function

- In general, if the random variable X takes values in $\{x_1, \ldots, x_n\}$ and p is probability mass function, then we have

$$p(x_1) + p(x_2) + \ldots + p(x_n) = 1$$

- In general, the forms of probability mass function represent the distribution of the random variable X

- A few popular examples of probability mass function include:

    - **Bernoulli distribution**: an example is that you toss a coin one time and X is the number of heads. Then, X takes value in {0, 1} and p(0) = P(X = 0), p(1) = P(X = 1). The value of p(0) or p(1) represents the distribution of X.

    - **Binomial distribution:** an example is that you toss a coin more than 1 time and X is the number of heads.

    - **Poisson distribution:** an example is that we count the number of births per hour in a given day

Expectation of Discrete Random Variable

- X: discrete random variable takes values in $\{x_1, \ldots, x_n\}$
- p: probability mass function
- The **expectation** of random variable X is defined as

$$E(X) = x_1 p(x_1) + x_2 p(x_2) + \ldots + x_n p(x_n)$$

- The expectation represents the **weighted average** of the possible values that X can take

Expectation of Discrete Random Variable

- We toss a coin twice: $\Omega = \{HH, HT, TH, TT\}$, the number of heads X takes values in {0, 1, 2}
- p(0) = P(X = 0) = 1/4, p(1) = P(X = 1) = 1/ 2, p(2) = P(X = 2) = 1/4
- E(X) = $0 \times P(X = 0) + 1 \times P(X = 1) + 2 \times P(X = 2)$

$$= 0 \times \frac{1}{4} + 1 \times \frac{1}{2} + 2 \times \frac{1}{4} = 1$$

- It shows that the expected value of the number of heads we observe will be 1

Expectation of Discrete Random Variable

Question: Assume that we toss a fair coin three times. What is the expected value of the number of heads?

- *A few hints*:
  - The outcome space
    $\Omega = \{HHH, HHT, HTH, HTT, THH, TTH, THT, TTT\}$
  - The number of heads X takes values in $\{0,1,2,3\}$
  - $P(X=0) = \frac{1}{8}, P(X=1) = \frac{3}{8}, P(X=2) = \frac{3}{8}, P(X=3) = \frac{1}{8}$

Expectation of Discrete Random Variable

- $P(X=0)=\frac{1}{8}, P(X=1)=\frac{3}{8}, P(X=2)=\frac{3}{8}, P(X=3)=\frac{1}{8}$
- $E(X) = 0 \times P(X=0) + 1 \times P(X=1) + 2 \times P(X=2) + 3 \times P(X=3)$

$$= 0 \times \frac{1}{8} + 1 \times \frac{3}{8} + 2 \times \frac{3}{8} + 3 \times \frac{1}{8} = \frac{3}{2}$$

- **Conclusion**: The expected value of the number of heads when we toss a fair coin three times is 1.5

Properties of Expectation

- Assume that X and Y are discrete random variables
- **Additive property**: $E(X + Y) = E(X) + E(Y)$
- **Multiplication property**: For any number a,

$$E(a \times X) = a \times E(X)$$

- The additive and multiplication properties are useful for understanding several concept of statistical learning that we study later in the course

Variance of Discrete Random Variable

- Another important notion is variance
- X: discrete random variable takes values in $\{x_1, \ldots, x_n\}$
- p: probability mass function
- The **variance** of random variable X is defined as

$$\begin{aligned} \text{var}(X) &= E(X^2) - E^2(X) \\ &= \left[x_1^2 p(x_1) + \ldots + x_n^2 p(x_n)\right] - E^2(X) \end{aligned}$$

- The variance measures the amount that the random variable differs from its expectation

Variance of discrete random variable

- We toss a coin twice: $\Omega = \{HH, HT, TH, TT\}$, the number of heads X takes values in {0, 1, 2}
- p(0) = P(X = 0) = 1/4, p(1) = P(X = 1) = 1/ 2, p(2) = P(X = 2) = 1/4
- E(X) = $0 \times \frac{1}{4} + 1 \times \frac{1}{2} + 2 \times \frac{1}{4} = 1$
- $E(X^2) = 0^2 \times P(X = 0) + 1^2 \times P(X = 1) + 2^2 \times P(X = 2)$

$$= 0^2 \times \frac{1}{4} + 1^2 \times \frac{1}{2} + 2^2 \times \frac{1}{4} = 3/2$$

- var(X) = $E(X^2) - E^2(X) = \frac{3}{2} - 1^2 = \frac{1}{2}$

Independent discrete random variables

- Assume that X and Y are discrete random variables
- X and Y take values in $\{x_1, \ldots, x_n\}$
- X and Y are **independent** random variables if

$$P(X = x, Y = y) = P(X = x) \times P(Y = y)$$

for any $x, y \in \{x_1, \ldots, x_n\}$

- The intuition of independence notion is that knowing about X gives us no information about Y and vice versa

Independent Discrete Random Variable

- We toss a coin twice
- X = the number of heads in the first toss
- Y = the number of heads in the second toss

Question: Are X and Y independent random variables?

Independent Discrete Random Variable

- $X, Y \in \{0,1\}$
- X({H}) = Y({H}) = 1, X({T}) = Y({T}) = 0
- $P(X = 0, Y = 0) = P(\{TT\}) = \frac{1}{4}$ (as we only have 4 possible outcomes from two tosses)
- Therefore, $P(X = 0, Y = 0) = P(X = 0) \times P(Y = 0)$
- Similarly, we can check that $P(X = x, Y = y) = P(X = x) \times P(Y = y)$ for all $(x, y) \in \{0,1\}$
- It indicates that X and Y are independent random variables

(In)dependent Random Variables in Data Science

- A lot of instances of (in)dependent random variables in data science:
    - A collection of images, videos (computer vision)
    - Time series data (stock market prices, natural language processing, etc.)
    - Spatial data (ecology, geography, etc.)