Mathematics for Machine Learning

Lecture 6
(23.05.2024)

# Probability Theory

Mahammad Namazov

# Table of contents

# Introduction

WHY?

Uncertainty

Philosophy

# Why Probability?

- “Probability theory is nothing more than common sense reduced to calculation.” (Laplace)
- Life is full of observable and incomplete scenarios;
- Now put projection of it onto AI applications: Tons of uncertainty to deal with! (Non-determinism)
- We are here to understand how things work:
  - How do we decide?
  - How AI models decide?
- The art of “Formulation of decision-making process”
- What is uncertainty that AI tries to model and how?

# Uncertainty

- Three main sources of uncertainty:
  - Inherent Stochasticity:
    - ✓ Hypothetical Card Game
  - Incomplete Observability:
    - ✓ Monty Hall problem;
  - Incomplete Modelling:
    - ✓ Discarding some relevant information

- Simple but uncertain is better than complex but certain
  - Who decides?
  - What is better?

# Philosophy

- We are waiting for a friend, where 3 possibilities can occur:
  - H1: He/She is on time;
  - H2: Delay because of traffic;
  - H3: Alien abduction;

- 3 mathematical criteria by E. T. Jaynes (1922-1998):
  - The degrees of plausibility are represented by real numbers;
  - These numbers must be based on the rules of common sense;
  - The resulting reasoning must be consistent, where consistency must be defined in following meanings:
    a. Consistency or non-contradiction;
    b. Honesty;
    c. Reproducibility;

# Set Theory

Sets

Set Operations

Countability

# Union

- Suppose we have 2 sets $A$ (red) and $B$ (blue)

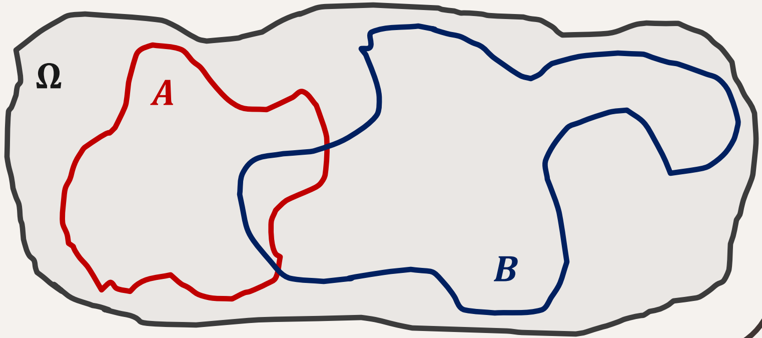
$$A = \{a_1, a_2, \dots, a_m\}$$
$$B = \{b_1, b_2, \dots, b_N\}$$

- The union of these sets is a set C which includes all elements of both sets:

$$C = A \cup B = \{c \in A \; or \; c \in B\}$$

  - $c$ is any element that belongs to $C$: $c \in C$

- In programming, it corresponds to **OR**



# Union

- Suppose we have 2 sets $A$ (red) and $B$ (blue)

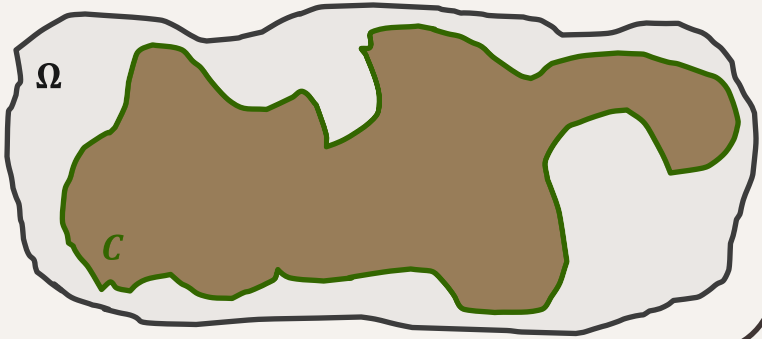
$$A = \{a_1, a_2, \dots, a_m\}$$
$$B = \{b_1, b_2, \dots, b_N\}$$

- The union of these sets is a set C which includes all elements of both sets:

$$C = A \cup B = \{c \in A \; or \; c \in B\}$$

  - $c$ is any element that belongs to $C$: $c \in C$

- In programming, it corresponds to **OR**



# Intersection

- Suppose we have 2 sets $A$ (red) and $B$ (blue)

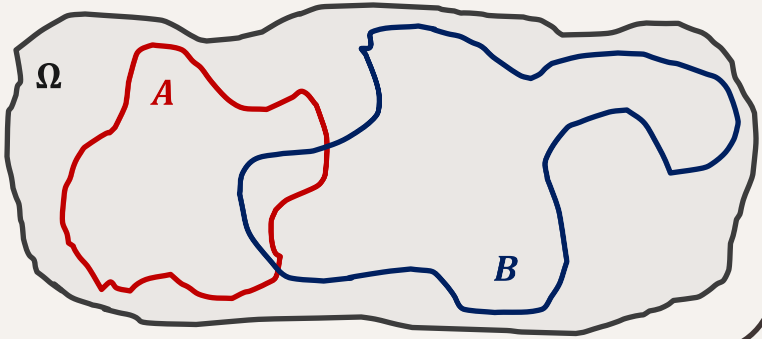
$$A = \{a_1, a_2, \dots, a_m\}$$
$$B = \{b_1, b_2, \dots, b_N\}$$

- The intersection of these sets is a set C which includes all elements of both sets:

$$C = A \cap B = \{c \in A \; and \; c \in B\}$$

  - $c$ is any element that belongs to $C$: $c \in C$

- In programming, it corresponds to **AND**



# Intersection

- Suppose we have 2 sets $A$ (red) and $B$ (blue)

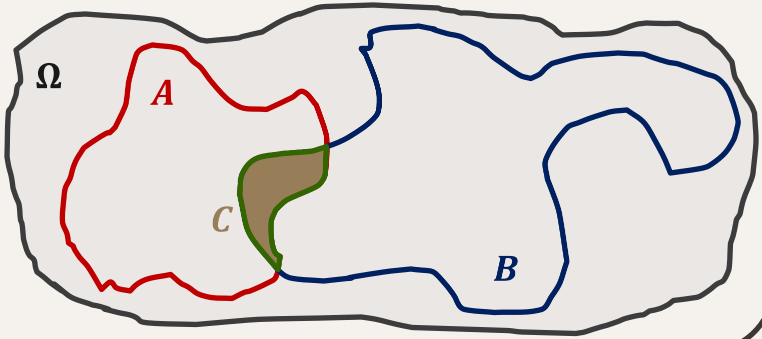
$$A = \{a_1, a_2, \dots, a_m\}$$
$$B = \{b_1, b_2, \dots, b_N\}$$

- The intersection of these sets is a set C which includes all elements of both sets:

$$C = A \cap B = \{c \in A \ and \ c \in B\}$$

  - $c$ is any element that belongs to $C$: $c \in C$

- In programming, it corresponds to **AND**



# Complement

- Let's say $A$ is an arbitrary set:

$$A = \{a_1, \dots, a_n\}$$

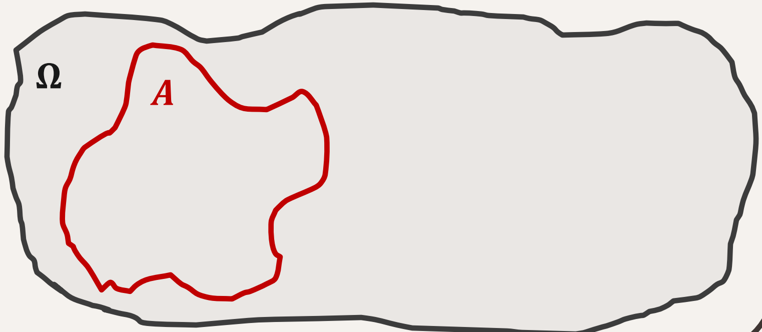
- The complement of $A$ is a new set, where:

$$\bar{A} = \{x \notin A\}$$

  - $x$ stands for any element that is in set $\bar{A}$;
  - It is also represented with $A^c$;

- $\Omega$ is a universal set;

- For a set and its complement, following property holds: $A + \bar{A} = \Omega$



# Complement

- Let's say $A$ is an arbitrary set:

$$A = \{a_1, \dots, a_n\}$$

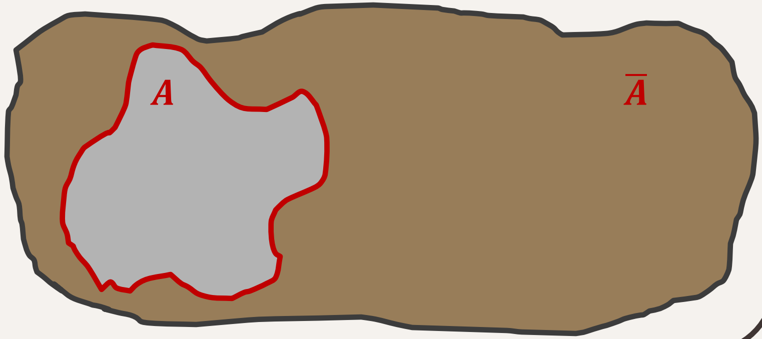
- The complement of $A$ is a new set, where:

$$\bar{A} = \{x \notin A\}$$

  - $x$ stands for any element that is in set $\bar{A}$;
  - It is also represented with $A^c$;

- $\Omega$ is a universal set;

- For a set and its complement, following property holds: $A + \bar{A} = \Omega$



# Subtraction

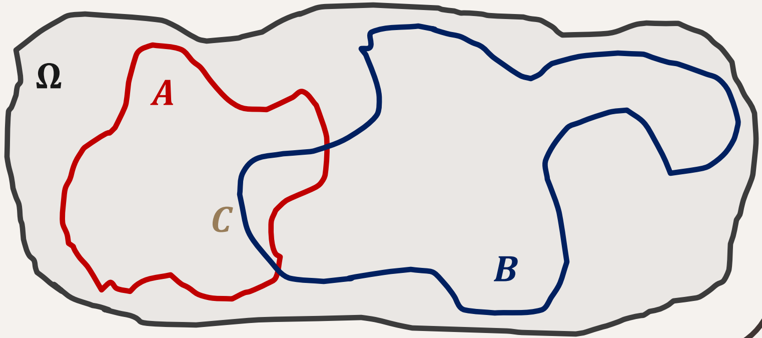
- Suppose we have 2 sets $A$ (red) and $B$ (blue)

$$A = \{a_1, a_2, \ldots, a_m\}$$
$$B = \{b_1, b_2, \ldots, b_N\}$$

- Subtracting $B$ from $A$ is a set $C$, which includes all elements of $B$ except for ones in $A$:

$$C = A - B = \{c \in A \; and \; c \notin B\}$$

  - $c$ is any element that belongs to $C$: $c \in C$



# Subtraction

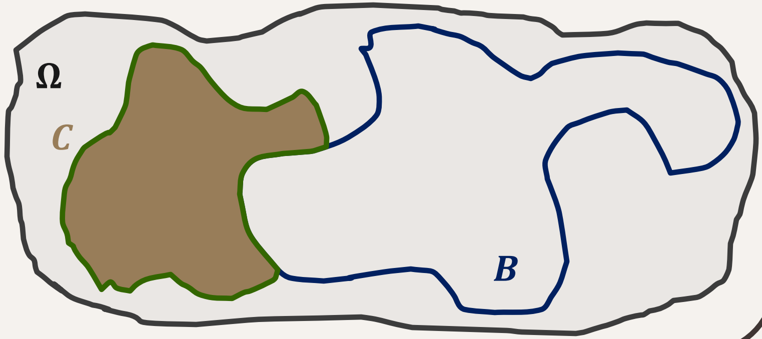
- Suppose we have 2 sets $A$ (red) and $B$ (blue)

$$A = \{a_1, a_2, \dots, a_m\}$$
$$B = \{b_1, b_2, \dots, b_N\}$$

- Subtracting $B$ from $A$ is a set $C$, which includes all elements of $B$ except for ones in $A$:

$$C = A - B = \{c \in A \; and \; c \notin B\}$$

  - $c$ is any element that belongs to $C$: $c \in C$



# Disjoint

Sets

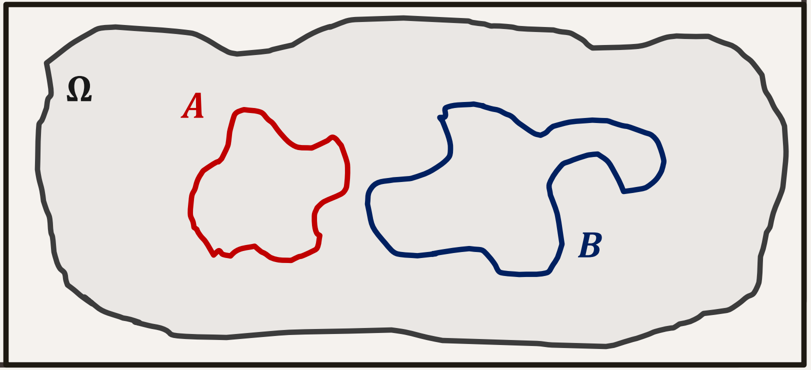
- When $A$ (blue) and $B$ (red) are sets, so that:

$$A = \{a_1, \dots, a_n\}$$
$$B = \{b_1, \dots, b_m\}$$

- A and B are disjoint sets when they do not share any element:

$$A \cap B = \emptyset$$

  - Their intersection is an empty set;



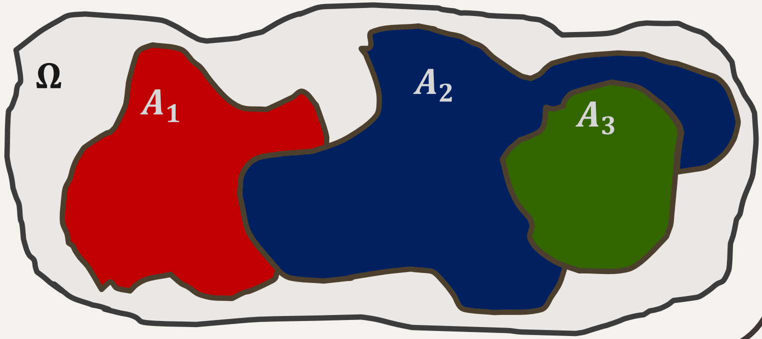
# Partition

Sets

- Let's say $A$ is an arbitrary set:

$$A = \{a_1, \dots, a_n\}$$

- $A_1, A_2, \dots$ are disjoint subsets of $A$;
  - None of them share any element:
    - $A_1 \cap A_2 \cap \cdots = \emptyset$
  - Union of them is $A$:
    - $A_1 \cup A_2 \cup \cdots = A$
- For instance: $A_1, A_2, A_3 \subset A$
- Spoiler Alert:
  - Law of Total Probability;
  - Bayes Theorem;



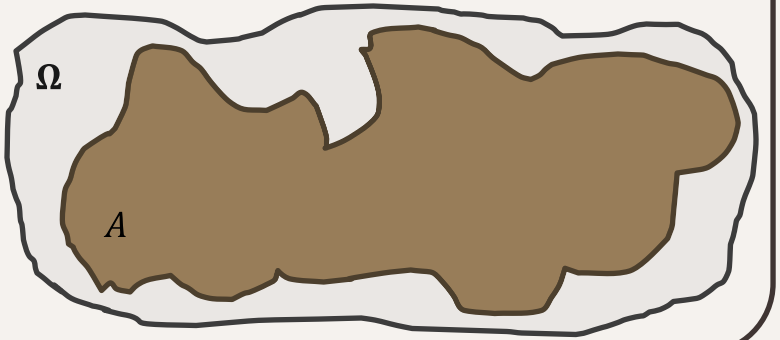
# Partition

Sets

- Let's say $A$ is an arbitrary set:

$$A = \{a_1, \dots, a_n\}$$

- $A_1, A_2, \dots$ are disjoint subsets of $A$;
  - None of them share any element:
    - $A_1 \cap A_2 \cap \cdots = \emptyset$
  - Union of them is $A$:
    - $A_1 \cup A_2 \cup \cdots = A$
- For instance: $A_1, A_2, A_3 \subset A$
- Spoiler Alert:
  - Law of Total Probability;
  - Bayes Theorem;



# Countability

- A set $A$ is countable:
  - If it is a finite set: $|A| < \infty$;
  - Or its elements have one to one correspondence with natural numbers (i.e., countably infinite): $\{0.1, 0.3, 0.7, 1.2, 23, \dots\}$
- A set is uncountable if it is not countable: $[a, b], [a, b)$ where $a < b$;
- Discrete variables' range is countable set;
- Continuous variables' range is uncountable set;

# Probability Theory

Probability

Scenarios

Marginal

Joint

Conditional

Independence

# Kolmogorov Axioms

- Assume we have a fair die to roll;
- Set of all possible values that die can end up: $\boldsymbol{S} = \{1, 2, 3, 4, 5, 6\}$
- **Probability of any outcome cannot be negative:** $\mathbf{0} \leq \boldsymbol{P}(\boldsymbol{x})$
- **Probability of any outcome will be one of those numbers:** $\boldsymbol{P}(\boldsymbol{S}) = \mathbf{1}$
- Another scenario:
  - Die will end up with any of those values in the sample space;
  - Any event does not share any information: they are disjoint;
  - **To compute several disjoint events' probability is just summing up:**

$$P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$$

# Notation

- **Sample Space ($\Omega$):**
  - The set of all possible outcomes of the experiment
- **Event Space ($\mathcal{A}$):**
  - The space of potential results of the experiment
  - Event Space is often the power set of Sample Space
- **Probability ($P$):**
  - For each event $A \in \mathcal{A}$, degree of belief for the occurrence of this very event;
- **Target Space ($\mathcal{T}$):**
  - i.e., States where unique cases are taken into account from sample space;
- **Random Variable:**
  - A variable that maps elements of sample space into target space:

$$X : \Omega \rightarrow \mathcal{T}$$

# Probability Models

- We will work with Sample Space for simplicity;

- Probability model is:
  - Discrete when sample space $\Omega$ is a countable set;
  - Continuous when sample space $\Omega$ is uncountable set (interval);

- Let's check such scenario:
  - Sample space will be: $\Omega = \{s_1, s_2, \dots\}$
  - An event $\mathcal{A}$ is subset of $\Omega$, so that $\mathcal{A} \subset \Omega$;

$$P(\mathcal{A}) = P\left(\bigcup_{s_j \in \mathcal{A}} \{s_j\}\right) = \sum_{s_j \in \mathcal{A}} P(s_j)$$

- A bit more specific (equally likely scenario when $\Omega$ has $N$ elements):

$$P(s_i) = \frac{1}{N}, \qquad i \in \{1, 2, \dots, N\};$$

  - What if $A$ is subset of $\Omega$ that includes $M$ possible outcomes from $\Omega$? What is $P(A) =$?

# Card Deck Environment

Scenarios

- There are 52 cards in the standard poker card deck;

$$j \in Symbols = \{Spade, \quad Heart, \quad Club, \quad Diamond\}$$

$$Numbers_j = \{2,\ 3,\ 4,\ 5,\ 6,\ 7,\ 8,\ 9,\ 10\};$$

$$Faces_j = \{J, Q, K\}$$

$$Specials_j = \{A\}$$

$$Type = \bigcup_{j \in Symbols} Type_j$$

- $Clubs = Numbers_{clubs} \cup Faces_{clubs} \cup Specials_{clubs} = \{1, 2, \dots, A\}$

# Dice Environment

Scenarios

- There are 2 fair dice with 6 sides;

$$S_1 = S_2 = \{1, 2, 3, 4, 5, 6\}$$

- Possible events:
  - Event of the first die will be rolled and get 3;
  - Event that the second die will get even numbers;
  - Event that sum of dice results is $\leq 7$, given that the first is 3;

# Marginal Probability

- Computing the possibility of the single event;
- We do not care the relation with other events;
- For instance, having 3 after the rolling the first die:
- It is independent event from other possibilities;
- Let's name this event as $A$:

$$P(A) = \frac{possibilities\ that\ the\ first\ die\ will\ be\ 3}{all\ possible\ outcomes} = \frac{1}{6}$$

# Marginal Probability

- Probability of two or more events happen together;
- Suppose we draw a card from the deck;
- We analyze two events ($A$ and $B$) occurrence at the same time;
  - $A$: The picked card is spade => There are 13 cards like that;
  - $B$: The picked card is  number => 9 of such cards are numbers;

$$P(card\ is\ spade\ \boldsymbol{and}\ number) = \frac{9}{52}$$

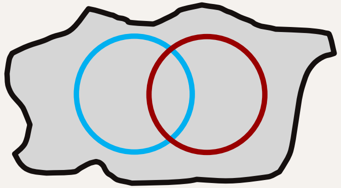
- What if we multiply marginal probabilities $P(A)$ and $P(B)$:

$$P(card\ is\ spade\ \boldsymbol{and}\ number) = P(A \cap B) = P(A)P(B) = \frac{36}{52}\frac{13}{52} = \frac{9}{52}$$

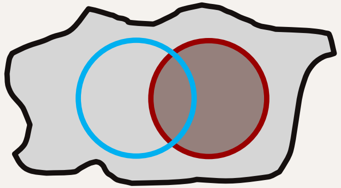
# Conditional Probability

- Definition:
  - There are two events: $A \subset S$ and $B \subset S$
  - Compute the probability of the scenario: Given that $B$ occurs, how likely $A$ will occur?

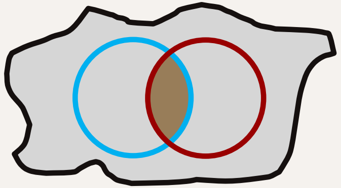
- Let's imagine what is going on here: (Divide and conquer)



There are 2 events:
$A$ and $B$



Given that $B$ occurs



How likely $A$ will occur?

# Conditional Probability

- Specific region that we investigate: $A \cap B$;
- Out of the area that we are certain about its occurrence: $B$
- Then mathematically:

$$P(A|B) = \frac{|A \cap B|}{|B|} = \frac{P(A \cap B)}{P(B)}, P(B) > 0$$

- Now think reversely: $B$ occurs given that $A$ occurs:

$$P(B|A) = \frac{|A \cap B|}{|A|} = \frac{P(A \cap B)}{P(A)}, P(A) > 0$$

**Question: What is the difference between causality and conditionality?**

# Independence

- Happening of one event does not impact the other;
- **Definition**: Events A and B are independent, if and only if:
$$P(A \cap B) = P(A)P(B)$$
- How does this impact conditional probability?
$$P(A|B)P(B) = P(A \cap B) = P(B|A)P(A)$$
- What if two events are dependent?
$$P(A \cap B) = P(A|B)P(B)$$

# Theories for ML

Total Probability

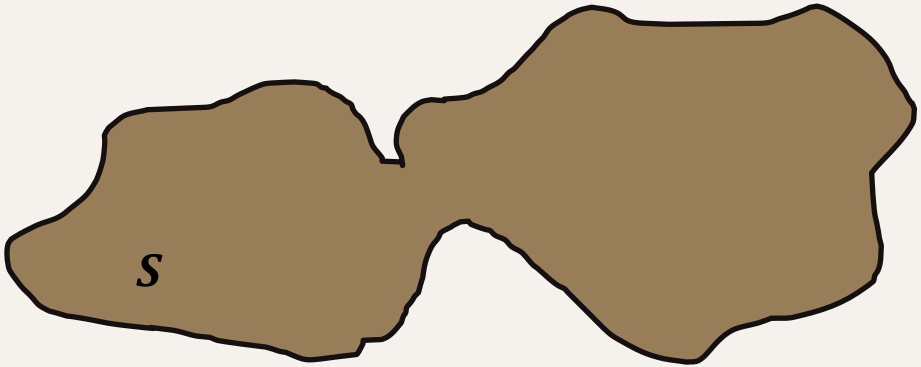
Bayes' Theorem

Conditional Independence

# Idea

## Law of Total Probability

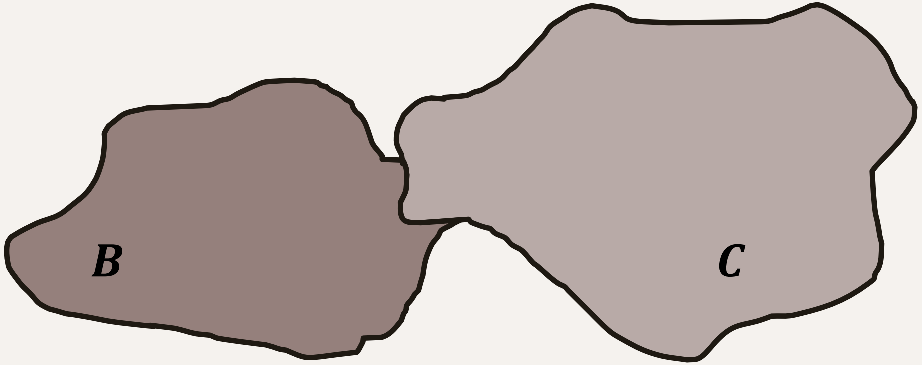
- ***Divide et impera*** (Phillip II Greek) or ***Divide et regnes*** (Napoleon)
- You have a problem to solve ($S$):



# Idea

Law of Total Probability

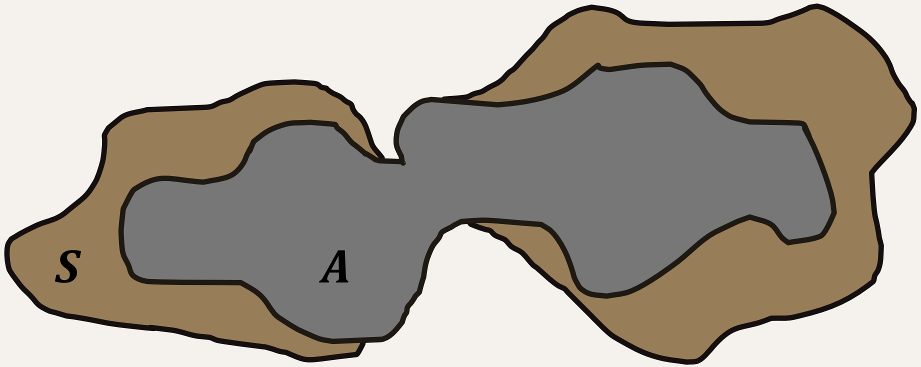
- ***Divide et impera*** (Phillip II Greek) or ***Divide et regnes*** (Napoleon)
- We know that solving $B$ and $C$ can solve $S$:



# Idea

## Law of Total Probability

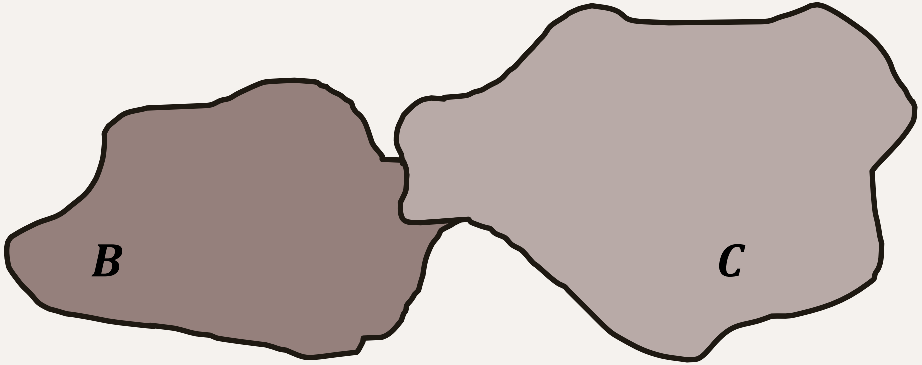
- ***Divide et impera*** (Phillip II Greek) or ***Divide et regnes*** (Napoleon)
- Task: Solve A, which cannot be solved directly:



# Idea

Law of Total Probability

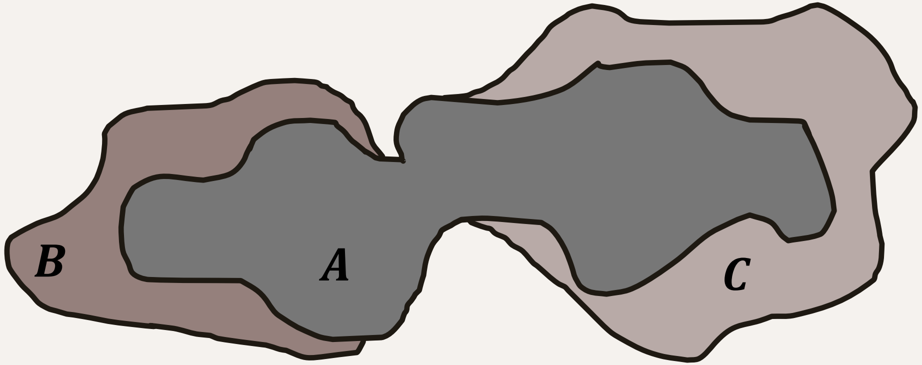
- ***Divide et impera*** (Phillip II Greek) or ***Divide et regnes*** (Napoleon)
- Can we solve it with $B$ and $C$?



# Idea

Law of Total Probability

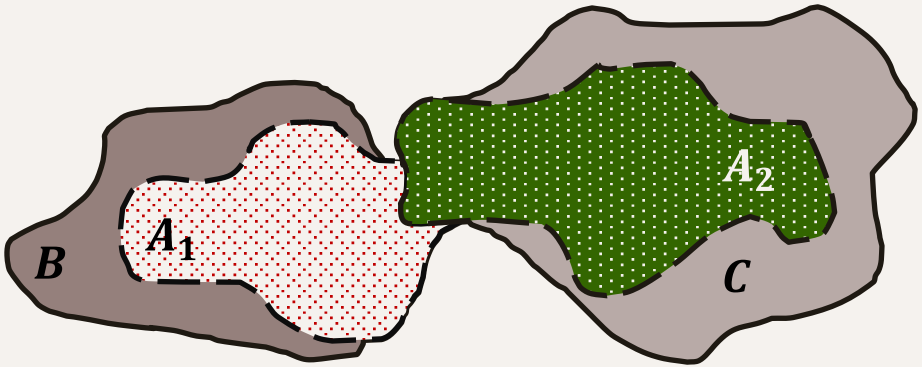
- ***Divide et impera*** (Phillip II Greek) or ***Divide et regnes*** (Napoleon)
- In fact, it is possible, since $A$ can also be partitioned by them:



# Idea

Law of Total Probability

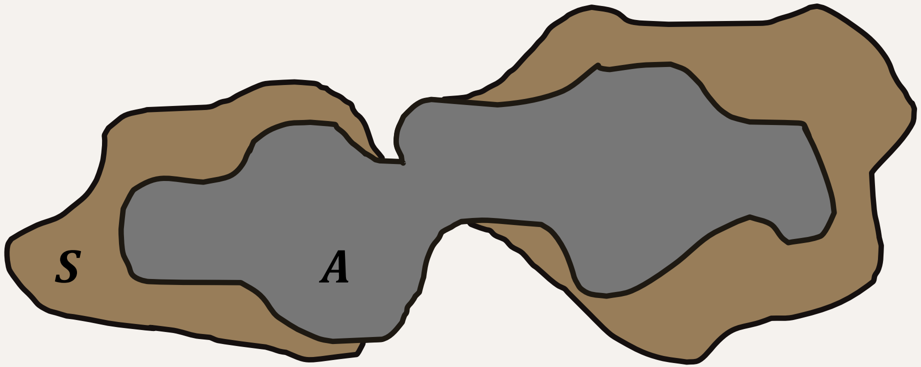
- ***Divide et impera*** (Phillip II Greek) or ***Divide et regnes*** (Napoleon)
- Voila: Solving $B$ and $C$ will implicitly solve $A_1$, $A_2$, respectively!



# Idea

Law of Total Probability

- ***Divide et impera*** (Phillip II Greek) or ***Divide et regnes*** (Napoleon)
- Then, we not only solved $S$ but also $A$;



# Mathematically

Law of Total Probability

- We know what are partitions of a set, right?
  - They don't share any (even very little) point;
  - When you sum them up, they make the set;

- For $S$, we have following partitions:

$$B \cap C = \emptyset, B \cup C = S$$

- For $A$, we have following partitions:

$$A_1 \cap A_2 = \emptyset, A_1 \cup A_2 = A$$

- Since we know how to solve $S$ by using $B$ and $C$, we can also solve $A$ by them:
  - Solving $S$:

$$P(S) = P(B) + P(C)$$

  - Solving $A$:

$$P(A \cap S) = P(A \cap B) + P(A \cap C) = P(A)$$

# Generalization

## Law of Total Probability

- Suppose the sample space $S$ has $n$ partitions $B = \{B_1, B_2, \dots, B_n\}$:

$$\bigcup_{i=1}^{n} B_i = S\ ; \bigcap_{i=1}^{n} B_i = \emptyset$$

- We know that $\mathrm{A} \subset S$, thus $A$ is also partitioned by the partitions of $S$. Then:

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i)$$

- Using the conditional probability:

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$$

# Brief Introduction

Bayes' Theorem

- The quantification of uncertainty based on experience;

- The knowledge that we developed through experience, without any other impacts, is called as **Prior Knowledge;**

- However, it is not the case always!
    - New events can modify the knowledge we have;
    - In other words, we update our knowledge through new learning steps;
    - New state of our knowledge is **Posterior Knowledge**

- To sum up:
  *Conditioning our* **Prior Knowledge** *based on new events and updating it with these new events bring us* **Posterior Knowledge***;*

# Scenario

## Bayes' Theorem

- Conditional Probability tell us:

$$P(A|B) = \frac{P(A,B)}{P(B)} \; or \; P(B|A) = \frac{P(A,B)}{P(A)}$$

- Let's use the second one to get better understanding;
- Let's generalize the equation with "Erasmus" example;
- To sum up, using new knowledge we update Prior Knowledge, to get Posterior Knowledge;
- Learn through experiences, apply them to forecast the next step with more informative way, rather than randomly guessing;

# Theorem

## Bayes' Theorem

- Bayes' Theorem (or Bayes' Rule, Bayes' Law)
  - For any two events A and B, where P(A) not zero, we have:
$$P(B_i|A) = \frac{P(A, B_i)}{P(A)}$$
  - If $B_1, B_2, \ldots$ forms a partition of sample Space $S$, and $A$ is any event with $P(A) \neq 0$:
$$P(B_i|A) = \frac{P(A, B_i)}{\sum_j P(A|B_j)P(B_j)} = \frac{P(A|B_i)P(B_i)}{\sum_j P(A|B_j)P(B_j)}$$

- What does Bayes want to tell us:
  - You have a belief (Prior Knowledge) (i.e., hypothesis);
  - You see a specific evidence;
  - Focusing on the evidence itself and your hypothesis will mislead you;
  - Rather focus on all possible evidences to update your belief (including all evidences that falsify your hypothesis)

# First Scenario

## Conditional Independence

- Since it is a bit complicated scenario, let's start with an example;

- Alice and Bob are expected in one event;
  - Event A: Alice will be late to the event;
  - Event B: Bob will be late to the event;
  - Let's define that these events are independent;

- Let's introduce new event:
  - Event C: They are coming from the same neighborhood;

- Then if Alice will be late, then Bob will be too and vice versa.

- Then given information make them conditionally dependent given the extra information C;

# Second Scenario

## Conditional Independence

- Now let’s continue the scenario:
- The meeting is over, and everyone leaves.
  - Bob is expected for a dinner at home;
  - Alice was invited for a dinner by her cousin;
  - Event A: Alice will arrive on time;
  - Event B: Bob will arrive on time;
  - Let’s define that these events are independent;
- Let’s introduce new event:
  - Event C: Thunderstorm hits, and traffic becomes awful in general;
- If Alice will be late, will Bob arrive in time?

# Mathematics

## Conditional Independence

- In case of introduced new event C does not help to deduce one's (A) outcome based on other's (B), then A and B are conditionally independent given C occurs.

- Formal Definition: Events A and B are conditionally independent given that C occurs, if and only if:

$$P(A,B|C) = P(A \cap B|C) = P(A|C)P(B|C)$$

- Another equation for this case:

$$P(A|B,C) \ = \ P(A|C)$$

- Similarly:

$$P(B|A,C) \ = \ P(B|C)$$

# Conclusion

Summary

Takeaways

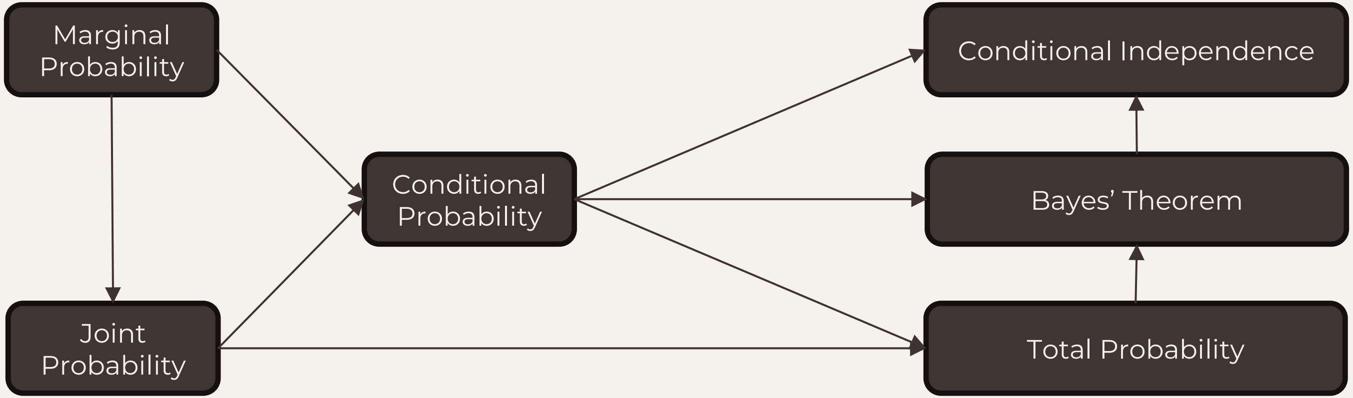
References

# Summary

- We have seen what are marginal, joint and conditional probability;
- When we discuss two events occur together, we compute joint;
- Conditioning event A to B, pushes us to investigate specific region where both occurs;
- There is relation between joint and conditional probabilities;
- The law of total probability, Bayes' Theorem and Conditional Independence will come to visit us, frequently;
- We established fundamental knowledge, what is next?

# Takeaways

- Why have we seen today's contents?



# References

- Further information to read:
  - Deisenroth, M. P., Faisal, A. A., & Ong, C. S. (2020). *Mathematics for machine learning*. Cambridge University Press.
    - Chapter 6, Section 1 only
    - Author's suggestion: Chapter 2 from (Walpole et al., 2011)
  - https://www.probabilitycourse.com
    - Chapter 1

# The End

Thanks for your attention!

Mahammad Namazov