Lecture 4: Bayes' Rule and Independency

Yunwen Lei

School of Computer Science, University of Birmingham

1 Bayes' Formula

In some cases, we have some prior information on how likely an event would happen and we are interested in how likely the observation of other events would affect this probability. Bayes' formula provides an useful approach to solve this problem.

Theorem 1 (Bayes' Formula)**.** *Let A and B be two events. If $\mathbb{P}(A) \neq 0$, then*

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c)}$$

The proof of Bayes' formula needs to use the definition of conditional probability and the Law of Total Probability.

Proof of Theorem 1. First, we have two formulations of $\mathbb{P}(A \cap B)$

$$\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A). \tag{1.1}$$

If $\mathbb{P}(A) \neq 0$, then it follows that

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)} \tag{1.2}$$

The probability of A can be computed through the *law of total probability*

$$\mathbb{P}(A) = \mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c)$$

The stated bound then follows by combining the above two equalities together. □

Remark 1. We now provide some explanations on Bayes' formula. In Bayes' formula, the event A can be interpreted as effect and the event B can be interpreted as cause. Bayes' formula tells us how to compute $\mathbb{P}(B|A)$, which is the probability of cause under the observation of an effect. That is, when we observe an effect we want to find which causes this to happen. To apply Bayes' formula, we require the information of $\mathbb{P}(B)$ and $\mathbb{P}(A|B), \mathbb{P}(A|B^c)$.

- $\mathbb{P}(B)$ is a prior probability. It is a guess of how likely the cause would occur without any observation.
- $\mathbb{P}(A|B), \mathbb{P}(A|B^c)$ are conditional probability of A (effect) given B or B^c (cause). Usually this information is relatively easy to obtain: it is easy to find how cause would affect the effect because this is a forward direction. As a comparison, it is not easy to directly get $\mathbb{P}(B|A)$ since this is a backward direction. This shows the power of the Bayes' formula in getting the probability of a cause under the information of an effect.

Example 1 (Diagnosis)**.** Suppose we want to find how likely a person has a disease after a test.

- Suppose 1 in 1000 persons has a certain disease
- A test detects the disease in 99% of diseased persons

- The test also "detects" the disease in 5% of healthy persons
- With what probability does a positive test diagnose the disease?

Solution: We first introduce several events in the language of probability.

$$D = \{``diseased''\}, \quad H = \{``healthy''\}, \quad + = \{``positive''\}$$

It is intuitive that the event D, H are cause, while the event $+$ is an effect. A healthy person is likely to cause a negative test, while a disease is likely to cause a positive test. We are given that

$$\mathbb{P}(D) = 0.001, \quad \mathbb{P}(+|D) = 0.99 \quad \mathbb{P}(+|H) = 0.05.$$

By Bayes' formula

$$\mathbb{P}(D|+) = \frac{\mathbb{P}(+|D) \cdot \mathbb{P}(D)}{\mathbb{P}(+|D)\mathbb{P}(D) + \mathbb{P}(+|H) \cdot \mathbb{P}(H)} = \frac{0.99 \cdot 0.001}{0.99 \cdot 0.001 + 0.05 \cdot 0.999} = 0.0194$$

This shows that after a positive test, the probability of getting the disease increases from 0.001 to 0.0194. Even with a positive test, it is still unlikely that the person has a disease. Therefore, the person should do more tests to get more convincing results.

Theorem 1 considers a special case where the cause can be either B or its complement. We can generalize Theorem 1 to a general case with more possible causes.

Theorem 2 (Bayes' Rule: General Case)**.** *If $B_1, B_2, B_3, \ldots$ form a partition of the sample space S, and A is any event with $P(A) \neq 0$, we have*

$$\mathbb{P}(B_j|A) = \frac{\mathbb{P}(A \cap B_j)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A|B_j)\mathbb{P}(B_j)}{\sum_i \mathbb{P}(A|B_i)\mathbb{P}(B_i)} \tag{1.3}$$

The result can be proved in a similar way and we omit the proof here. Here we give some explanations. We consider several possible cause $B_1, B_2, \ldots$ and one effect A. Upon the observation of this effect, we want to find how likely is each cause.

- We know $\mathbb{P}(A|B_i), i = 1, 2, 3, \ldots$, but we are interested in the probability $\mathbb{P}(B_j|A)$
- We are provided with prior probability $\mathbb{P}(B_i)$. This is a guess of each cause without any observation.
- We are also given likelihood (causal knowledge) $\mathbb{P}(A|B_i)$. As we explained before, this information is relative easy to obtain.
- Bayes' formula computes the posterior probability $\mathbb{P}(B_j|A)$. Based on an observation, we should update the probability of cause.

Example 2. Experiment on Human Memory We consider an experiment related to human memory on different items.

- Participants have to memorize a set of words (B_1), numbers (B_2) and pictures (B_3).
- These occur in the experiment with $\mathbb{P}(B_1) = 0.5, \mathbb{P}(B_2) = 0.4, \mathbb{P}(B_3) = 0.1$.
- Let $A = \{\text{correctly recalled}\}$
- Suppose $\mathbb{P}(A|B_1) = 0.4, \mathbb{P}(A|B_2) = 0.2, \mathbb{P}(A|B_3) = 0.1$.

Question: What is the probability that an item that is correctly recalled (A) is a picture (B_3)?

Solution: We use the Bayes' formula to solve this problem. In this example, the possible causes are B_1, B_2 and B_3, while the effect is A. We have the prior information on the probability of each cause. We also know how likely each cause would lead to this effect. We want to find the mostly like cause for this effect.

- By the law of total probability

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}(B_1)\mathbb{P}(A|B_1) + \mathbb{P}(B_2)\mathbb{P}(A|B_2) + \mathbb{P}(B_3)\mathbb{P}(A|B_3) \\ &= 0.5 \cdot 0.4 + 0.4 \cdot 0.2 + 0.1 \cdot 0.1 = 0.29\end{aligned}$$

- By Bayes' theorem

$$\mathbb{P}(B_3|A) = \frac{\mathbb{P}(B_3)\mathbb{P}(A|B_3)}{\mathbb{P}(A)} = \frac{0.1 \cdot 0.1}{0.29} = 0.0345.$$

2 Independency

2.1 Independency

As we discussed before, conditional probability is used to describe how an event would affect other event. In some special cases partial information on an experiment does not change the likelihood of an event. Here are some examples showing the dependency of different events.

- Sex of first child has nothing to do with sex of second (independent)
- What about eyecolor? (not independent) The underlying reason is that the eyecolor depends on genes of parent. If a child has a blue eye, then it is likely that the parent would have a blue eye. This shows that the second child is also likely to have a blue eye.
- Tosses of different coins have nothing to do with each other (independent)
- Cloudy and rainy days. When it rains, there must be clouds (not independent)

Definition 1 (Independence)**.** We say two events A and B are independent if (we use the notation $A \perp B$ or $A \perp\!\!\!\perp B$)

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B). \tag{2.1}$$

In this case (assuming $\mathbb{P}(B) \neq 0$)

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A)\mathbb{P}(B)}{\mathbb{P}(B)} = \mathbb{P}(A). \tag{2.2}$$

Remark 2. We give two different ways to check whether two events are independent: one based on the probability of intersection (Eq. (**??**)) and one based on the concept of conditional probability (Eq. (**??**)). From the view of intersection, it shows that the probability of intersection can be factored into the product of probability of each event. From the view of conditional probability, it shows that the information of B does not affect the probability of A, i.e., knowing B occurred doesn't change the probability of A

Below we give two examples regarding the rolling of two dices.

Example 3 (Rolling Two Dices)**.** We consider two events as follows.

- Let A be the event that the sum is 6

$$A = \big\{(1,5),(2,4),(3,3),(4,2),(5,1)\big\}$$

- Let B be the event that the first die shows 3

$$B = \big\{(3,1),(3,2),(3,3),(3,4),(3,5),(3,6)\big\}$$

- Then $A \cap B = \{(3,3)\}$. Since each outcome has the same likelihood we know

$$\frac{1}{36} = \mathbb{P}(A \cap B) \neq \mathbb{P}(A)\mathbb{P}(B) = \frac{5}{36}\frac{6}{36} = \frac{5}{6 * 36}$$

This shows that these two events are dependent.

Example 4 (Rolling Two Dices)**.** We consider two events as follows.

- Let A be the event that the sum is 7

$$A = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}$$

- Let B be the event that the first die shows 3

$$B = \{(3,1),(3,2),(3,3),(3,4),(3,5),(3,6)\}$$

- Then $A \cap B = \{(3,4)\}$ and therefore

$$\frac{1}{36} = \mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B) = \frac{6}{36}\frac{6}{36}$$

- These two events are independent

2.2 Pairwise Independency and Independency

If we consider more than two events, then we can have other definitions of independency. We can define the pairwise independency and (mutual) independency. Note that the (mutual) independency is stronger than pairwise independency. We often omit the word mutual.

Definition 2. Consider three events A, B, C.

- We say A, B, C are pairwise independent iff

$$\begin{aligned} \mathbb{P}(A \cap B) &= \mathbb{P}(A)\mathbb{P}(B) \\ \mathbb{P}(B \cap C) &= \mathbb{P}(B)\mathbb{P}(C) \\ \mathbb{P}(A \cap C) &= \mathbb{P}(A)\mathbb{P}(C). \end{aligned} \tag{2.3}$$

- We say A, B, C are (mutually) independent if Eq. (2.2) holds and

$$\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)$$

We can extend this definition to more events. We omit this discussion for brevity. Now we provide an example to show some events that are pairwise independent but are not independent. We still consider the experiment of rolling two dices.

Example 5 (Rolling Two Dices)**.** We define the following three events A, B, C.

- Let A be the event that the sum is 7

$$A = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\} \Longrightarrow \mathbb{P}(A) = \frac{1}{6}$$

- Let B be the event that the first die shows 3

$$B = \{(3,1),(3,2),(3,3),(3,4),(3,5),(3,6)\} \Longrightarrow \mathbb{P}(B) = \frac{1}{6}$$

- Let C be the event that the second die shows 4

$$C = \{(1,4),(2,4),(3,4),(4,4),(5,4),(6,4)\} \Longrightarrow \mathbb{P}(C) = \frac{1}{6}$$

Then $A \cap B = \{(3,4)\}, A \cap C = \{(3,4)\}, B \cap C = \{(3,4)\}$.

Pairwise independency. We first check that A, B, C are pairwise independent as follows

$$\frac{1}{36} = \mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B) = \frac{1}{6}\frac{1}{6}$$
$$\frac{1}{36} = \mathbb{P}(B \cap C) = \mathbb{P}(B)\mathbb{P}(C) = \frac{1}{6}\frac{1}{6}$$
$$\frac{1}{36} = \mathbb{P}(A \cap C) = \mathbb{P}(A)\mathbb{P}(C) = \frac{1}{6}\frac{1}{6}$$

We now show that A, B, C are **not** independent

$$\mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C) = \frac{1}{6*6*6}$$
$$A \cap B \cap C = \{(3,4)\} \Longrightarrow \mathbb{P}(A \cap B \cap C) = \frac{1}{36}.$$

It is then clear that $\mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C) \neq \mathbb{P}(A \cap B \cap C)$.

2.3 Application of Independency

Recall that in the last class, we introduce the product rule

$$\mathbb{P}(A_1 \cap \cdots \cap A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2|A_1)\cdots\mathbb{P}(A_n|A_1 \cap \cdots \cap A_{n-1}).$$

If $A_1, \ldots, A_n$ are independent, the above product rule can be simplified as follows

$$\mathbb{P}(A_1 \cap \cdots \cap A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2)\cdots\mathbb{P}(A_n). \tag{2.4}$$

This is useful to compute the probability of intersection of events or its complement.

Example 6. Consider n independent experiments, each of which succeeds with probability $p \in [0,1]$.
Question: What is the probability that every single experiment succeeds?

- This is relatively easy. We can just apply the simplified product rule in Eq. (2.3) and get that the probability is $p^n \to 0$ as $n \to \infty$.

Question: What is the probability that at least one experiment succeeds?

- The basic idea is to look at the complement. The complement is that each experiment fails, which is an intersection of several events. We can apply Eq. (2.3) to get

$$\mathbb{P}(\{\text{each fails}\}) = (1-p)^n$$

- Then we can use the property of probability to show that

$$\mathbb{P}(\{\text{at least succeeds}\}) = 1 - (1-p)^n \to 1 \text{ as } n \to \infty$$

The above example shows that the complement can be easy to compute sometimes. This trick is called "getting rid of ORs". Suppose we are finding the probability of OR of events, which is not easy. We can turn it into an AND. The underlying reason is that AND refers to the intersection of several events. The probability of intersection can be easy to compute if these events are independent:

$$\mathbb{P}(\underbrace{A_1 \cup A_2 \cup \cdots \cup A_n}_{\text{at least one } n \text{ events happen}}) = 1 - \mathbb{P}(\underbrace{A_1^c \cap A_2^c \cap \cdots A_n^c}_{\text{none of these } A_i \text{ happen}}) = 1 - \mathbb{P}(A_1^c)\mathbb{P}(A_2^c)\cdots\mathbb{P}(A_n^c) \tag{2.5}$$

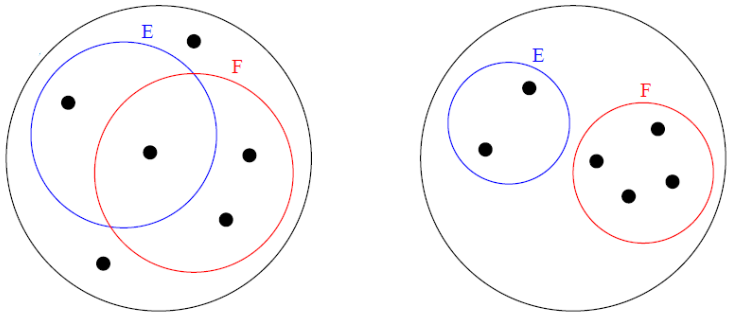
Example 7. A multiple choice test consists of six questions. Each question has four choices for answers, only one of which is correct. A student guesses on all six questions. What is the probability that he gets at least one answer correct?
Solution:

- Let A_i be the event that the answer to the i-th question is correct. It is clear $\mathbb{P}(A_i) = 1/4$
- We are interested in $\mathbb{P}(A_1 \cup A_2 \cup A_3 \cup A_4 \cup A_5 \cup A_6)$
- According to Eq. (2.4) we know

$$
\begin{aligned}
\mathbb{P}(A_1 \cup A_2 \cup A_3 \cup A_4 \cup A_5 \cup A_6) &= 1 - \mathbb{P}(A_1^c \cap A_2^c \cap A_3^c \cap A_4^c \cap A_5^c \cap A_6) \\
&= 1 - 0.75^6 = 1 - 00178 = 0.822.
\end{aligned}
$$

Before ending this section, we give some remarks.

Remark 3 (Independence and Disjointness)**.** The first remark is on the difference between independence and disjointness.



Independent, but not disjoint. Disjoint, but not independent.

(The six outcomes in Ω are assumed to have equal probability.)

- If E and F are independent then $\mathbb{P}(E \cap F) = \mathbb{P}(E)\mathbb{P}(F)$. Let us consider the left figure. The sample space has 6 outcomes, E has 2 outcomes, F has 3 outcomes and $E \cap F$ has one outcome. Therefore

$$
\text{Left :} \qquad \mathbb{P}(E) = 2/6 \quad \mathbb{P}(F) = 3/6 \quad \mathbb{P}(E \cap F) = 1/6.
$$

  This shows the independency between E and F. However, E and F are not disjoint.

- If E and F are disjoint then $\mathbb{P}(E \cap F) = \mathbb{P}(\emptyset) = 0$. Let us consider the right figure. Note E has 2 outcomes, F has 4 outcomes and $E \cap F$ has zero outcome.

$$
\text{Right :} \qquad \mathbb{P}(E) = 2/6 \quad \mathbb{P}(F) = 4/6 \quad \mathbb{P}(E \cap F) = 0
$$

  Therefore, $\mathbb{P}(E \cap F) \neq \mathbb{P}(E)\mathbb{P}(F)$ and E, F are not independent. However, these two events are disjoint.

Remark 4 (Independence Extends to Complements)**.** The independency between two events carry over to their complements. That is

$$
\begin{aligned}
& A \text{ and } B \text{ are independent } \iff A^c \text{ and } B \text{ are independent} \\
\iff & A \text{ and } B^c \text{ are independent } \iff A^c \text{ and } B^c \text{ are independent}
\end{aligned}
$$

Intuition: independence means information about A does not give you any information about B. Therefore, information about A (A^c) does not give you any information about B (B^c).

3 Conditional Independency

In the previous section, we use probability to define independence. We can extend this idea: using conditional probability to define conditional independency.

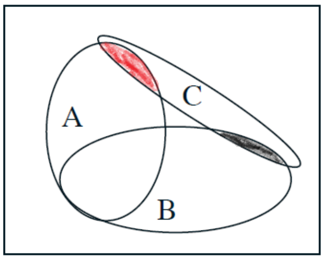
Definition 3 (Conditional Independence)**.** We say A and B are conditionally independent given C iff (we write $A \perp\!\!\!\perp B|C$)

$$\mathbb{P}(A \cap B|C) = \mathbb{P}(A|C)\mathbb{P}(B|C) \Longleftrightarrow \mathbb{P}(A|B, C) = \mathbb{P}(A|C).$$

According to the definition, conditional independence, given C, is defined as independence under probability law $\mathbb{P}(\cdot|C)$. If C is the sample space, then this conditional independence recovers our original definition of independence. This shows that conditional independence is an extension of independence. We can take different conditions and this leads to different concepts of conditional independence.

The motivation to introduce conditional independence is to notice that take conditions may affect the independence. Below is an example.

- Assume A and B are independent



$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$$

- If we are told that C occurred, are A and B independent?

  According to the definition, we know that A and B are not independent given C.

$$\begin{aligned}\mathbb{P}(A \cap B|C) &= 0\\ &\neq \mathbb{P}(A|C)\mathbb{P}(B|C)\end{aligned}$$

Example 8 (Conditioning May Affect Independence)**.** We give here an example to show several independent events can be not conditionally independent.

Let us throw a die

- $A = \{\text{die outcome is even}\} = \{2, 4, 6\}$
- $B = \{\text{die outcome is} \leq 4\} = \{1, 2, 3, 4\}$
- $C = \{\text{die outcome} > 1\} = \{2, 3, 4, 5, 6\}$

$$\mathbb{P}(A|C) = \frac{|A \cap C|}{|C|} = \frac{3}{5}$$

$$\mathbb{P}(B|C) = \frac{|B \cap C|}{|C|} = \frac{3}{5}$$

$$\mathbb{P}(A \cap B \cap C) = \frac{|\{2, 4\}|}{6} = \frac{1}{3}$$

- First, we know A and B are independent

$$\frac{2}{6} = \mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B) = \frac{1}{2}\frac{4}{6}$$

- Second, A and B are not conditionally independent given C

$$\begin{aligned}\mathbb{P}(A \cap B|C) = \mathbb{P}(A \cap B \cap C)/\mathbb{P}(C) &= \frac{2/6}{5/6} = 2/5\\ \neq \mathbb{P}(A|C)\mathbb{P}(B|C) &= \frac{3}{5}\frac{3}{5} = \frac{9}{25}\end{aligned}$$

Example 9 (Example of Conditional Independence)**.** We give here an example to show several conditionally independent events can be not independent.

- Kevin separately phones two students, Alice and Bob.
- To each, he tells the same number in $\{1, 2\}$
- Due to the noise in the phone, Alice and Bob each imperfectly (and independently) draw a conclusion about what number Kevin said

$$\begin{aligned} A_1 &= \{\text{Alice think he heard number 1}\} & A_2 &= \{\text{Alice think he heard number 2}\}\\ B_1 &= \{\text{Bob think he heard number 1}\} & B_2 &= \{\text{Bob think he heard number 2}\}\\ K_1 &= \{\text{Kevin says number 1}\} & K_2 &= \{\text{Kevin says number 2}\}\end{aligned}$$

- Are A_1 and B_1 probabilistically independent?
  - No: we'd expect (e.g.) $\mathbb{P}(A_1|B_1) > \mathbb{P}(A_1)$. If B_1 happens, then it is likely that K_1 happens. Therefore, it is likely that A_1 happens. Therefore, the probability of A_1 happening increases with the information of B_1.
- Why are A_1 and B_1 conditionally independent under K_1?
  - Because if we know the number that Kevin actually said, the two variables are no longer correlated. The only possible way B_1 can affect A_1 is through K_1 or K_2. As we already know K_1 has happened, B_1 can no longer affect A_1.
  - e.g., $\mathbb{P}(A_1|B_1, K_1) = \mathbb{P}(A_1|K_1)$, $\mathbb{P}(A_1|B_2, K_2) = \mathbb{P}(A_1|K_2)$
  - $A_i \perp B_j | K_r$, $i, j, r \in \{1, 2\}$

Remark 5 (Product rule)**.** An application of conditional independency is to simplify the product rule.

- If $A \perp\!\!\!\perp B|C$, then
$$\mathbb{P}(A|B, C) = \mathbb{P}(A|C) \tag{3.1}$$
- Recall the product rule
$$\mathbb{P}(A_1 \cap \cdots \cap A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2|A_1)\cdots\mathbb{P}(A_n|A_1 \cap \cdots \cap A_{n-1}) \tag{3.2}$$
- We can simplify the product rule if we have conditional independency
- For example, if A_i is conditionally independent of $A_j, j < i - 1$ given A_{i-1}, Eq. (3.1) implies
$$\mathbb{P}(A_i|A_1 \cap \cdots \cap A_{i-1}) = \mathbb{P}(A_i|A_{i-1}).$$
  and therefore
$$\mathbb{P}(A_1 \cap \cdots \cap A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2|A_1)\mathbb{P}(A_3|A_2)\cdots\mathbb{P}(A_n|A_{n-1}).$$
  This is much more simplified as compared to Eq. (3.2) as we only need to consider the condition of one event in the product rule.