CIS6930/4930

Sample Space, Event, Probability, Conditional Probability, Independence, Bayes' Formula

Jan. 14, 2021

Prof. Ye Xia

Sample Space and Event

- A probability space is a triple $(\Omega, \mathcal{F}, P)$.
- Ω: a sample space representing the set of all possible outcomes of an experiment. Example: A die thrown once.

$$\Omega = \{1, 2, 3, 4, 5, 6\}.$$

- An event A is a subset of Ω, $A \subseteq \Omega$. Examples of events:
  (a) the outcome is the number 2: $A = \{2\}$
  (b) the outcome is even: $B = \{2, 4, 6\}$.

  When an experiment is conducted, there is an outcome $\omega \in \Omega$ as a result. ω is not an event in itself. For any event A, if $\omega \in A$, we say event A happened/occurred. An experiment may result in the occurrence of multiple events.
  Suppose the outcome is $\omega = 2$. Then, both events A and B occurred.

$\mathcal{F}$: the σ-field

- $\mathcal{F}$ is a collection of subsets of Ω, representing a set of events that we are interested in. Each $A \in \mathcal{F}$ is a subset of Ω, i.e., $A \subseteq \Omega$.
- Implication: We may not be interested in or have information about all the subsets of Ω.
- $\mathcal{F}$ must satisfy certain properties, which make it a σ-field.

Definition: A collection $\mathcal{F}$ of subsets of Ω is called a σ**-field** if it satisfies the following conditions:
(a) $\emptyset \in \mathcal{F}$.
(b) (countable union): If $A_1, A_2, \ldots \in \mathcal{F}$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{F}$.
(c) (complement) If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$.

These properties imply that $\Omega \in \mathcal{F}$, and $\mathcal{F}$ is closed for countable intersections. That is, if $A_1, A_2, \ldots \in \mathcal{F}$, then $\cap_{i=1}^{\infty} A_i \in \mathcal{F}$.

What Are We Doing?

- The pair $(\Omega, \mathcal{F})$ is called a **measurable space**. One can assign a measure μ to such a space. Probability assignment is a measure.
- A **measure** is a nonnegative countably additive set function; that is, a function $\mu : \mathcal{F} \to \mathbb{R}$ with
  (a) $\mu(A) \geq 0$ for all $A \in \mathcal{F}$
  (b) $\mu(\emptyset) = 0$.
  (c) if $A_1, A_2, \ldots$ is a countable sequence of disjoint sets, then $\mu(\cup_i A_i) = \sum_i \mu(A_i)$.
- Intuitively, $\mu(A)$ is the 'size' or 'weight' of the set A.
- If $\mu(\Omega) = 1$, we call μ a **probability measure**. We usually use the notation P for a probability measure.

- The reason to incorporate σ-field into probability theory is that we want to think about probability as a measure. We can then use all the mathematical results and ways of thinking from measure theory.

Comments

- $(\Omega, \mathcal{F}, P)$ is called a **probability space**.
- Probabilities are assigned to events, in fact, events in $\mathcal{F}$; not to outcomes.
- When thinking about probabilities, we need to know the sample space and the σ-field. Not all subsets of Ω are assigned probabilities.
- An experiment produces an outcome. Before the experiment, we wish to anticipate the outcome with the help of the probabilities. However, the probabilities are not necessarily specified as how likely each outcome (more precisely, the singleton event that contains the outcome) is, although that is a possibility. In general, the probabilities are specified as how likely different events in $\mathcal{F}$ will occur. It may happen that none of the events in $\mathcal{F}$ is a singleton subset containing exactly one outcome.

- $\mathcal{F}$ and the probability assignments P together tell the chances that the experiment outcome will lie in each of the event in $\mathcal{F}$, or equivalently, how likely each event in $\mathcal{F}$ occurs.
- Very often, we will condition on some other σ-field $\mathcal{G}$ on the same sample space Ω, e.g., $P(B|\mathcal{G})$. In that context, it is helpful to think about σ-field $\mathcal{G}$ as an apparatus that helps us to collect information about the outcome of an experiment. Each $A \in \mathcal{G}$ is like a detector. Suppose the outcome is ω, which we don't really know. When we say 'conditioning on $\mathcal{G}$', we mean that we know whether $w \in A$ for each $A \in \mathcal{G}$, i.e., whether each event in $\mathcal{G}$ happened or not. That gives us some information about the outcome.
- It is often this 'conditional' view of the σ-field that illustrates why we like to think about a σ-field as an information set.
- Examples of σ-field:
  (a) $\mathcal{F} = \{\emptyset, \Omega\}$; known as the smallest σ-field or trivial σ-field

Why trivial: For any probability assignment P, we must have $P(\emptyset) = 0$ and $P(\Omega) = 1$. For any outcome ω of an experiment, $\omega \notin \emptyset$ and $\omega \in \Omega$. For any P, such $\mathcal{F}$ and P tell no information about the experimental outcome.

When conditioning on $\mathcal{F}$, all the outcomes are indistinguishable from each other and we learn nothing about the unobserved ω.

(b) For the single-toss die example: Let $\mathcal{F} =$ all subsets of Ω, which is σ-field. It is the largest or the most detailed σ-field on this sample space.

Suppose the outcome of the experiment is $w = 1$, which we don't know. Consider conditioning on $\mathcal{F}$. Since $\{1\} \in \mathcal{F}$, we will know the event $\{1\}$ happened. This really tells us the outcome is exactly 1. This applies to each other outcome as well. When conditioning on $\mathcal{F}$, we have very precise information about the outcome of the experiment, regardless of which outcome shows up.

(c) If $A \subseteq \Omega$, then $\mathcal{F} = \{\emptyset, A, A^c, \Omega\}$ is σ-field.

For the die example, suppose $A = \{1\}$. Then, $A^c = \{2, 3, 4, 5, 6\}$.

A probability assignment P might be $P(A) = 1/6$, which implies $P(A^c) = 5/6$. In such a case, we don't know what the probability is for the die to come up 2.

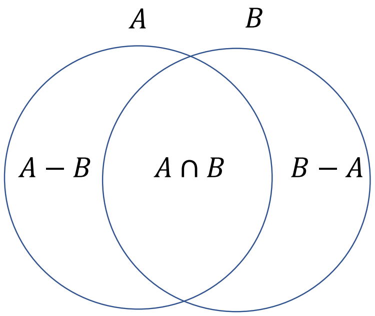
Also, suppose in an experiment, we observe A^c happened. The outcome can be any number from the set A^c. In other words, when conditioning on it, $\mathcal{F}$ cannot provide enough information to distinguish outcomes $2, 3, 4, 5$ and 6. On the other hand, the collection of all the subsets of $\{1, 2, 3, 4, 5, 6\}$ is a more detailed σ-field, and it can distinguish the outcomes when we condition on it.

(d) If $\mathcal{F}_1$ and $\mathcal{F}_2$ are two σ-fields on Ω and $\mathcal{F}_1 \subset \mathcal{F}_2$, then we say $\mathcal{F}_2$ is larger or more detailed, and $\mathcal{F}_1$ is coarser. $\mathcal{F}_2$ allows us to collect more information about the experimental outcome.

- In conventional probability, σ-field is often not mentioned. Usually the implicit assumption is $\mathcal{F} =$ the collection of all the subsets of Ω.

Probability Calculation - Trick 1

Split an event into disjoint ones and add the probabilities.



Example: Toss two fair coins. What is the probability that either the first coin or the second coin is heads?

- Sample space: $\Omega = \{HH, HT, TH, TT\}$. Each outcome, when viewed as an event, has probability $1/4$.
- The answer to the question is obviously $3/4$. But, we will do it in a harder way.

- Event of coin 1 being heads: $A = \{HH, HT\}$; $P(A) = 1/2$.
- Event of coin 2 being heads: $B = \{HH, TH\}$; $P(B) = 1/2$.
- Event of either coin is heads: $A \cup B = A \cup (B - A)$, where $B - A = B \cap A^c$ is the set difference.
- A and $(B - A)$ are disjoint events. Therefore, $P(A \cup B) = P(A) + P(B - A) = 1/2 + P(B - A)$.
- $B = (B - A) \cup (A \cap B)$, and the latter two events are disjoint. Therefore, $P(B) = P(B - A) + P(A \cap B)$. We have $P(B - A) = 1/2 - 1/4 = 1/4$.
- Finally, $P(A \cup B) = 1/2 + 1/4 = 3/4$.
- It may appear that we are doing something unnecessarily complicated, since we can easily see $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. How do we see that? We think about computing the areas of regions. When we add the areas

of regions A and B, the area of $A \cap B$ shows up twice.

- Indeed, probability assignment is like ‘area’ (in fact, area is a measure – the Lesbegue measure on $\mathbb{R}^2$). You can add the areas of disjoint regions when computing the area of the total region.

  $P(A) = P(A - B) + P(A \cap B)$; $P(B) = P(B - A) + P(A \cap B)$; hence, $P(A) + P(B) = P(A \cup B) + P(A \cap B)$.

- But, for more difficult examples, the trick can be useful because (i) it is a divide-and-conquer strategy and (ii) it allows careful book-keeping to avoid mistakes.

- Key question: How do we split an event into disjoint ones? The usual trick: **conditioning on something**

Conditional Probability

Given two events A and B (more rigorously, $A, B \in \mathcal{F}$) with $P(B) > 0$, the **conditional probability** of A occurs given B occurs is defined to be

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

- Conditional probability is a probability measure $P(\cdot|B)$. In particular, it is normalized so that $P(\Omega|B) = 1$.
- Check $(\Omega, B \cap \mathcal{F}, P(\cdot|B))$ is a probability space.
- Frequentist view: We look a large number of experiments in which B occurs. What is the fraction of them where A also occurs? That fraction is equal to $P(A|B)$.
- Given B occurs, A occurs if and only if $A \cap B$ occurs. Therefore, $P(A|B)$ must be proportional to $P(A \cap B)$, i.e.,

$P(A|B) = \alpha P(A \cap B)$ for some α. If we want $P(\Omega|B)$ to be a probability, then, $1 = P(\Omega|B) = \alpha P(\Omega \cap B) = \alpha P(B)$. We must have $\alpha = 1/P(B)$.

Example: A family has two children. What is the probability that both are boys given that at least one is a boy?

Sample space $\Omega = \{BB, BG, GB, GG\}$.
Event that both children are boys $E = \{BB\}$.
Event that at least one child is a boy $F = \{BB, BG, GB\}$. Then,

$$P(E|F) = \frac{P(E \cap F)}{P(F)} = 1/3.$$

Two Interpretations of Probability

Frequentist interpretation:

- $P(A)$ is the fraction of times (i.e., proportion) that A occurs if you repeat the underlying experiment a very large number of times.
- $P(A|B)$ is the fraction of times that A occurs out of all the experimental results where B occurs.
- The thinking of 'fraction of times' is very useful when pondering difficult probability problems or when building intuitions about mathematical results.

Bayesian interpretation:

- $P(A)$ is the degree of belief that A will occur.
- If X is the evidence you collect (e.g., training data), $P(A|X)$ is the updated degree of belief that A will occur after the evidence X is

observed.

- $P(A|X)$ is called the **posterior** probability/belief; $P(A)$ is called the **prior** probability/belief.
- The Bayesian interpretation is quite prominent when applying estimation/prediction to various real-world complex applications, e.g., economics, politics.

Example: Consider the next game for your favorite sports team. Suppose you thoroughly analyze the situations of your team and the opposing team and you say the chance of your team winning is 70%. What does that mean?

It is highly improbable that the exact match-up under the exact conditions can be repeated many times in the future. But, we still evaluate the magnitude of chances. Bayesian interpretation will say the assessment of 70% is your belief.

The Monty Hall Problem

Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1. Then, the host opens one of the other two doors that has a goat. Say that is door 3. He then says to you, "Do you want to switch to door 2?" Is it to your advantage to switch your choice?

You should compare the probabilities of events A and B.

A: the event that you won't switch and you get the car

B: the event that you switch and you get the car

Tricks: (i) conditional on something (usually a random variable); (ii) split the event of interest into disjoint events based on what you will condition on.

Natural choice: the outcome of the first step. Different outcomes (when viewed as events) help to partition the event of interest into disjoint ones.

C: the event that behind the first door you picked is a car.

$P(A) = P(A \cap C) + P(A \cap C^c) = P(A|C)P(C) + P(A|C^c)P(C^c) = 1 \cdot 1/3 + 0 \cdot 2/3 = 1/3$.

$P(B) = P(B \cap C) + P(B \cap C^c) = P(B|C)P(C) + P(B|C^c)P(C^c) = 0 \cdot 1/3 + 1 \cdot 2/3 = 2/3$.

Answer: You should switch!

What is the sample space? One answer is that there are two experiments and two sample spaces. The first one is associated with the not-switching strategy; the second is associated with the switching strategy.

For the first sample space, consider what the experiment is. It consists of you picking one door in the first step. The outcome is either C or G. $\Omega_1 = \{C, G\}$. The probability assignment is $P_1(\{C\}) = 1/3$; $P_1(\{G\}) = 2/3$. Therefore, the probability you get the car is $1/3$.

For the second sample space, your experiment is: You pick one door in the first step and later you switch. The set of all the possible outcomes is

$\Omega_2 = \{CG, GC, GG\}$. And,
$P_2(\{CG\}) = 1/3; P_2(\{GC\}) = 2/3; P_2(\{GG\}) = 0$. Therefore, the probability you get the car is equal to $P_2(\{GC\}) = 2/3$.

This is strange! Can we really compare two probabilities on two different sample spaces? It is not a problem if we are frequentist. In the not-switching strategy, on average, 1 out of 3 times you get the car; in the switching strategy, on average, 2 out of 3 times you get the car. The switching strategy is superior.

Sometimes, it is not easy to figure out what the sample space is. It is beneficial to think about it, and when things get complicated, you do need to know it.

Independence of Events

Events A and B are called **independent** if $P(A \cap B) = P(A)P(B)$.

A family of events $\{A_i : i \in I\}$ is independent if $P(\bigcap_{j \in J} A_j) = \prod_{j \in J} P(A_j)$ for all **finite** subsets J of I. (This is not to be confused with pairwise independence.)

Suppose $P(B) > 0$. Since $P(A|B) = P(A \cap B)/P(B)$, A and B are independent iff $P(A|B) = P(A)$. Similarly, if $P(A) > 0$, A and B are independent iff $P(B|A) = P(B)$.

Intuitively, the events A_i, $i \in I$, are independent if knowledge of the occurrence of any of these events has no effect on the probability of any other event.

Suppose $P(C) > 0$. Two events A and B are **conditionally independent given** C if $P(A \cap B|C) = P(A|C)P(B|C)$.

Bayes' Formula/Theorem

Version I: Given two events A and B with $P(A) > 0$ and $P(B) > 0$,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

Given our definition of $P(A|B)$, this is not so much a theorem but straightly comes from the definition, as

$$P(A|B)P(B) = P(A \cap B) = P(B|A)P(A).$$

Version II: Suppose $\{A_1, A_2, \ldots, A_n\}$ is a partition of Ω (i.e., they are disjoint and $\cup_{i=1}^{n} A_i = \Omega$). Suppose $P(B) > 0$ and $P(A_j) > 0$ for each

j. Then,

$$\begin{aligned}
P(A_j|B) =& \frac{P(A_j \cap B)}{P(B)} \\
=& \frac{P(B|A_j)P(A_j)}{P(B)} \\
=& \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^n P(B \cap A_i)} \\
=& \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^n P(B|A_i)P(A_i)}. \qquad (1)
\end{aligned}$$

Note how $P(B)$ is calculated by splitting B into disjoint events $\{B \cap A_i\}$.

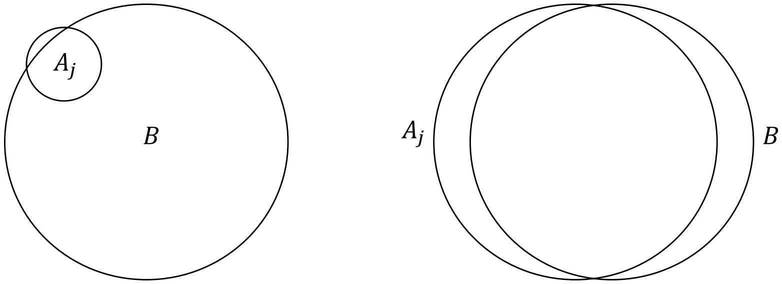
Some authors call (1) Bayes' formula.

$A_1, A_2, \ldots, A_n$ are alternatives. (1) helps us to calculate the chances of different alternatives, given the evidence B.

Suppose we observe that when A_j occurs, B almost always occurs. Suppose we observe B one day. There is the usual tendency to say it must be due to that A_j is happening.

This reasoning is problematic and often leads to wrong conclusions.

Why? We see that $P(A_j|B)$ can be very small even if $P(B|A_j)$ is very large (nearly equal to 1). If $P(A_j) \ll P(B)$ so that $P(B|A_j)P(A_j) = P(B \cap A_j) \leq P(A_j) \ll P(B)$ (left figure), then $P(A_j|B)$ is small.



$P(B)$ may be relatively large compared with $P(B|A_j)P(A_j) = P(B \cap A_j)$ if the other alternatives in the sum in (1) have enough contributions. In practice, it is often easy to see whether $P(B)$ is relatively large or not.

The right figure shows the opposite situation, where $P(B \cap A_j)$ is close to $P(B)$. In that case, $P(A_j|B)$ is large.

To summarize, under the condition that $P(B|A_j) \approx 1$, the reverse inference $B \to A_j$ with a high probability is invalid if $P(A_j)$ is substantially smaller than $P(B)$. **There are other reasons that B happens.** But, if $P(A_j)$ and $P(B)$ are comparable, the reverse inference is fine.

Example: PCR Test for COVID-19

The PCR test is highly reliable. Consider conducting the PCR test on an individual.

Let A be the event of the individual indeed has been infected.

Let B be the event that the PCR testing shows a positive result.

The conditional probability $P(B|A)$ is called **sensitivity** of the test.

High sensitivity means that the test can pick up infected individuals with high probability, and therefore, low false negative rate, which is $1 - P(B|A)$.

We also need $P(B^c|A^c)$, called **specificity**, which is the conditional probability that given the individual is not infected, the test result is negative. High specificity means low false positive rate.

Suppose $P(B|A) = 0.99$ and $P(B^c|A^c) = 0.9$, which indicate a very good test.

Now, consider a particular test on a particular individual. Suppose the test result is positive. Are we pretty sure that the individual has been infected?

At around February 2020, the prevalence of the infection is very low. Suppose only 0.2% of the individuals in the general population have been infected. There were calls for widespread testing to find and isolate the infected individuals.

Taking an arbitrary individual, the prior is $P(A) = 0.002$. Then, the posterior

$$
\begin{aligned}
P(A|B) =& \frac{P(B|A)P(A)}{P(B)} \\
=& \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)} \\
=& \frac{0.99 \times 0.002}{0.99 \times 0.002 + (1 - 0.9) \times 0.998} \\
=& 0.019.
\end{aligned}
$$

Note that $P(B)$ is roughly equal to the positivity rate of tests, which can be observed once widespread testing starts. $P(B)$ is not small since the 10% false positive rate easily contributes a lot to $P(B)$. This is a situation where $P(A) \ll P(B)$.

Intuitively, the problem lies in that the false positives swamp the test results.

On the other hand, suppose $P(B)$ is also very small. This can happen when the specificity is very high. Suppose $P(B^c|A^c) = 0.999$. Since there are very few false positives, a positive test most likely indicates true infection. Calculation shows $P(A|B) = 0.665$.

Conclusion: You need both high sensitivity and sufficiently high specificity. Equivalently, you need both low false negative rate and low false positive rate.

Testing in low prevalence scenarios can still be justified under additional conditions. If the individual to be tested is not chosen arbitrarily from the population, i.e., chosen uniformly at random, then the posterior probability can be very different. Examples:

- testing individuals with COVID-19-like symptoms
- testing passengers from a cruise ship with a known outbreak

In both cases, the prior $P(A)$ may be substantially larger than 0.002.

Widespread testing in low prevalence scenarios (hence, lower prior) can make sense under additional conditions.

- Repeated testing may help to rule out some false positives.
- In some country, individuals with positive test results are set aside as suspected cases and they will be observed for a period of time. If they later show COVID-like symptoms, their infection will be confirmed.

Overall Lesson: Knowing A implies B frequently is not enough for us to make the reverse inference. Once B is observed, we cannot jump to the conclusion that it must be because of A. You also need that the alternatives don't contribute much to B.

Doomsday Argument

There are two urns A and B. A contains 10 balls numbered 1 through 10. B contains 1 million balls, numbered 1 through one million. You don't know which is which.

Suppose you choose one urn randomly (w. p. $1/2$) and pick a ball from it. Suppose the number on the ball is 7. Which urn have you picked?

$$\begin{aligned} P(A|7) =& \frac{P(7|A)P(A)}{P(7|A)P(A) + P(7|B)P(B)} \\ =& \frac{0.1 \times 0.5}{0.1 \times 0.5 + 0.000001 \times 0.5} \approx 1. \end{aligned}$$

With very high probability, you picked urn A.

Now, consider two possible numbers of humans whoever will ever live: 100 billion or 100 trillion. Suppose you happen to be the 60th billon person. Using a similar calculation, you will find that 100 billion is much more likely scenario.

Assumptions: (i) You are a uniformly random sample from a population, given you have picked the population. (ii) The prior is $1/2$ for each possibility.

Possible Issues: (i) You are not a uniformly random sample. Then, what are you? (ii) There might be reasons to support a different prior. If so, your answer may be different. (iii) Why consider only two population sizes?

What do you think about this argument? Is there something wrong with this argument? Philosophers have a lot to say about it. Check it out online.

What Constitute Good Evidence?

Consider hypothesis/claim A and its converse A^c. Suppose $P(A)$ is small so that A is an extraordinary claim. Let B be the evidence. Then, the posterior

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

Good evidence needs to be sufficiently discriminate to separate the two alternatives. It is not enough that $P(B|A)$ is close to 1.

If $P(B|A^c)$ is also close to 1, then $P(A|B) \approx P(A)$: This is a case where $P(A) \ll P(B)$.

If $P(B|A^c)$ is sufficiently close to 0, then $P(A|B)$ may be close to 1: This is a case where $P(A) \approx P(B)$.

Next, does extraordinary claim requires extraordinary evidence? It depends on what we mean by ‘extraordinary evidence’. If we mean it by $P(B|A) \approx 1$, then not necessarily. Suppose $P(B|A) = 0.5$ but $P(B|A^c) = 0$. Then, $P(A|B) = 1$. From

$$P(A|B) = \frac{1}{1 + \frac{P(B|A^c)}{P(B|A)} \frac{P(A^c)}{P(A)}},$$

$P(B|A)$ matters less than $P(B|A^c)$, as long as $P(B|A)$ is not too small. An extraordinary evidence B should have the property that $P(B|A^c)$ is close to zero; in other words, there is almost no alternative explanation for B.

On a related note, if $P(A) = 0$, then $P(A|B)$ is always equal to 0: No amount of evidence can make A acceptable.

To be a good Bayesian, it is crucial to keep an open mind about any hypothesis. For a wacky idea, it is right that the prior should be very close to 0; but it should not be equal to 0.