How to Understand and Visualize Bayes Theorem

Authors: Wei Ren, Wei Ren

Date: 2025-09-20T19:56:34+00:00

Abstract

This paper discusses how to visualize conditional probability by a geometric graph for better understanding, and how to easily understand related derivations or applications such as combinatorial probability, posterior probability, likelihood, Bayes theorem, Markov Chain, and naive Bayes classification. We intend to facilitate learners to understand Bayes theorem and related concepts in an intuitive, straightforward, uniformly visualized, easily understood, eye-catched, user-friendly, self-contained, and one-stop manner.

Full Text

Preamble

How to Understand and Visualize Bayes Theorem

Wei Ren
School of Computer Science, China University of Geosciences, Wuhan, China
ORCID: 0000-0001-8590-1737

Abstract

This paper discusses how to visualize conditional probability through geometric graphs for better understanding, and how to easily comprehend related derivations or applications such as combinatorial probability, posterior probability, likelihood, Bayes theorem, Markov Chain, and naive Bayes classification.

We intend to facilitate learners' understanding of Bayes theorem and related concepts in an intuitive, straightforward, uniformly visualized, easily understood, eye-catching, user-friendly, self-contained, and one-stop manner.

Keywords: Bayes Theorem, Conditional Probability, Geometric Representation, Prior Probability, Posterior Probability

1. Introduction

In current well-known textbooks for college-level introductory courses on probability or statistics (e.g., [**?**, **?**]), widely-read popular science books on Bayes theory (e.g., [**?**]), or even scientific papers, it is rarely discussed how to visualize Bayes theory, its underlying fundamental concepts (e.g., conditional probability, joint probability, independence), and its major applications (e.g., false positive, precision, inference, maximum likelihood estimation, Markov chain, naive Bayes classification). Some popular science websites [**?**] may provide some attempts, but none of them present systematic, rigorous, formal, and extensive statements, proofs, and discussions.

In this paper, we study how to understand Bayes theorem thoroughly, easily, and intuitively, with the help of visualizing Bayes theorem together with underlying critical concepts and related application cases through a proposed unified geometric graph. We also explain how observations may change the probability of a hypothesis using this graph, and visualize the computational processes for Bayes probability through the graph, which remarkably facilitates better understanding and easier application of Bayes theorem.

Notation 1.1.
(1) $\Pr[E]$: the probability that event E occurs.
(2) $\square ABCD$: the rectangle with 4 vertices named A, B, C, and D.
(3) S_{ABCD}: the area of rectangle $\square ABCD$.
(4) $|AB|$: the length of line segment AB.
(5) Ω: full probability domain, which has $\Pr[\Omega] = 1$.
(6) $E_1 \cup E_2$, $E_1 \cap E_2$, $\overline{E}$: E_1 or E_2 occurs, E_1 and E_2 occur, and E does not occur, respectively. Note that we may use $\Pr[E_1, E_2]$ to replace $\Pr[E_1 \cap E_2]$ for simplicity.

2.1. Hypothesis Event and Prior Probability

Suppose the probability that an event A occurs is x, i.e., $\Pr[A] = x$, where $x \in [0, 1]$. As the basic version, we suppose there exist only two events (e.g., A and B) in the hypothesis domain (denoted as Ω). That is, $\Omega = A \cup B$.

Remark 2.1.
(1) Two concepts must be distinguished: mutual exclusivity and independence. Events A and B are mutually exclusive if and only if $\Pr[A \cap B] = 0$. Events A and B are independent if and only if $\Pr[A \cap B] = \Pr[A] \times \Pr[B]$.
(2) If two events are mutually exclusive, then $\Pr[A \cap B] = 0$. If two events are independent, then $\Pr[A \cap B] > 0$ and $\Pr[A \cap B] = \Pr[A] \times \Pr[B]$. If two events are neither independent nor mutually exclusive, then $\Pr[A \cap B] > 0$ and $\Pr[A \cap B] \neq \Pr[A] \times \Pr[B]$.
(3) If two events are mutually exclusive and both probabilities are non-zero, then the two events are not independent. If two events are independent and both probabilities are non-zero, then the two events are not mutually exclusive.
(4) If two events are mutually exclusive, then $\Pr[A \cup B] = \Pr[A] + \Pr[B]$.

(5) Generally, A_i, A_j $(i, j \in [1, n])$ are independent if and only if $\Pr[A_i, A_j] = \Pr[A_i] \times \Pr[A_j]$. More generally, $\{A_1, A_2, ..., A_n\}$ are independent if and only if for all $A_{i1}, A_{i2}, ..., A_{im} \subseteq \{A_1, A_2, ..., A_n\}$, we have $\Pr[A_{i1}, A_{i2}, ..., A_{im}] = \Pr[A_{i1}] \times \Pr[A_{i2}] \times ... \times \Pr[A_{im}]$.
(6) Generally, $A_1, A_2, ..., A_n$ are mutually exclusive if and only if for all $i, j \in [1, n]$, $\Pr[A_i, A_j] = 0$.
(7) Generally, A_i, A_j are mutually exclusive and A_i $(i = 1, 2, ..., n)$ are partitions of the full domain if and only if for all $i, j \in [1, n]$, $\Pr[A_i, A_j] = 0$, and $\sum_{i=1}^{n} \Pr[A_i] = 1$.
(8) If $A_1, A_2, ..., A_n$ are not mutually exclusive, then $\Pr[A_1 \cup A_2 \cup ... \cup A_n] = \Pr[A_1]+...+\Pr[A_n]-\Pr[A_1, A_2]-\Pr[A_1, A_3]-...-\Pr[A_{n-1}, A_n]+\Pr[A_1, A_2, A_3]+\Pr[A_1, A_2, A_4] + ... + \Pr[A_{n-2}, A_{n-1}, A_n] + (-1)^{n+1} \Pr[A_1, A_2, ..., A_n]$.

Geometric Representation. We use a geometric graph to show probabilities for better understanding (see Fig. 1 [Figure 1: see original paper]). $\Pr[A] = S_{A_1A_2A_3A_4}$, where $S_{A_1A_2A_3A_4}$ is the area of rectangle $\square A_1A_2A_3A_4$, as the area of the full domain Ω is 1. $\Pr[\overline{A}] = 1-S_{A_1A_2A_3A_4}$, which is indeed the area in the domain outside rectangle $\square A_1A_2A_3A_4$ (namely, the complement). $\Pr[\Omega] = 1$ can be represented by the area of the full domain, which is a 1×1 square.

[Figure 1: see original paper]

Exclusive probability can be easily observed by the fact that there is no overlap (i.e., intersection) between exclusive events. In contrast, there must be overlaps between independent events. Interestingly, independent events can also be represented by a geometric graph similarly, which will be discussed later (in Section 2.3).

2.2. Observation Event

$\Pr[A]$ may sometimes be called prior probability or hypothesis probability. It could be a probability computed from statistics over current samples, previous observations, empirical estimation, or from subjective conjecture. We will derive and explain the posterior probability after some observations, or the impact of observations on the probability of the hypothesis.

As the basic version, we suppose there exists only one observation event, called O. Suppose $\Pr[O] = y$, where $y \in [0, 1]$. Then $\Pr[\overline{O}] = 1 - \Pr[O] = 1 - y$.

Geometric Representation. The geometric representation is shown in Fig. 2 [Figure 2: see original paper]. $\Pr[O] = S_{O_1O_2O_3O_4}$, where $S_{O_1O_2O_3O_4}$ is the area of rectangle $\square O_1O_2O_3O_4$. $\Pr[\overline{O}] = 1-S_{O_1O_2O_3O_4}$, which is the complementary area in the domain outside rectangle $\square O_1O_2O_3O_4$.

Note that observation events may happen multiple times, and each observation event may result in one impact on the hypothesis probability. We will discuss this later (in Section 4.1).

[Figure 2: see original paper]

2.3. Joint Probability

Joint probability is the probability that all events occur, e.g., $\Pr[A, B]$ is the probability that both event A and event B occur.

The widely accepted graph for showing joint probabilities is Fig. 3 [Figure 3: see original paper], which is a diagrammatic drawing. It can easily show set relations, but it is not able to quantitatively show the probabilities. That is, quantitative relations among the areas for $\Pr[A]$, $\Pr[O]$, and $\Pr[A, O]$ cannot be envisioned easily. An alternative graph should thus be proposed, but strangely, such a graph is rarely found. We hereby propose a new method to show them, see Fig. 4 [Figure 4: see original paper].

[Figure 3: see original paper]

Geometric Representation. We still use rectangles like Fig. 1 and Fig. 2, but we need to show the joint probabilities in Fig. 4. The geometric explanation is as follows:

$\Pr[\Omega] = S_{O_1O_2P_5A_4} = |O_2P_5| \times |A_4P_5| = 1 \times 1 = 1.$

$\Pr[A] = S_{O_1A_2A_3A_4} = 0.4 \times 1 = 0.4;$

$\Pr[O] = S_{O_1O_2O_3O_4} = 1 \times 0.6 = 0.6;$

$\Pr[A, O] = S_{O_1A_2A_5O_4} = |O_1A_2| \times |O_1O_4| = 0.4 \times 0.6 = 0.24;$

Thus, $\Pr[A, O] = \Pr[A] \times \Pr[O]$.

Besides, $\Pr[\overline{A}, O] = S_{A_2O_2O_3A_5} = 0.6 \times 0.6 = 0.36.$

$\Pr[A, \overline{O}] = S_{O_4A_5A_3A_4} = 0.4 \times 0.4 = 0.16.$

$\Pr[\overline{A}, \overline{O}] = S_{A_5O_3P_5A_3} = 0.6 \times 0.4 = 0.24.$

Note that this is a general visualization method for independent probabilities. Any intersection of two independent probabilities equals the area of the intersection of two mutually perpendicular rectangles.

Remark 2.2.
(1) If all events are mutually exclusive, then all rectangles should be placed in parallel and there should be no overlap between them.
(2) If we represent two independent events in a parallel manner (e.g., two rectangles in parallel columns or rows), then the intersection (overlap) between two rectangles is difficult to show quantitatively by the area of the intersection (overlapped) rectangle (e.g., with double slash shadows). Therefore, we use two perpendicular rectangles to represent two independent events.
(3) This method can represent multiple independent events using multiple perpendicular rectangles, e.g., one rectangle that is perpendicular to multiple rectangles.
(4) The sum of the areas from two perpendicular rectangles (e.g., rectangles for $\Pr[A]$ and $\Pr[O]$) may be larger than 1, because in this case there must be an

overlap area for these two rectangles. For example, suppose $\Pr[A] = 0.6$ and $\Pr[O] = 0.6$. Then $\Pr[A]+\Pr[O]-\Pr[A\cap O] = 0.6+0.6-0.36 = 0.84 = \Pr[A\cup O]$. $\Pr[\overline{A \cup O}] = \Pr[\overline{A}, \overline{O}] = 1 - 0.84 = 0.16$.
(5) In contrast, the sum of the areas from two perpendicular rectangles may be smaller than 1, because there could be other events in the domain. For example, $\Pr[A] = 0.4$ and $\Pr[O] = 0.4$ could be possible. The key point is that $\Pr[A] + \Pr[O] - \Pr[A \cap O] = 0.4 + 0.4 - 0.16 = 0.64 = \Pr[A \cup O]$. $\Pr[\overline{A \cup O}] = \Pr[\overline{A}, \overline{O}] = 1 - 0.64 = 0.36$.
(6) We can observe that the rectangle for A and the rectangle for O are perpendicular; the rectangle for $\overline{A}$ and the rectangle for O are perpendicular; the rectangle for A and $\overline{O}$ is perpendicular. By simple computation, it confirms that $\Pr[A, O] = \Pr[A] \times \Pr[O]$, $\Pr[\overline{A}, O] = \Pr[\overline{A}] \times \Pr[O]$, $\Pr[A, \overline{O}] = \Pr[A] \times \Pr[\overline{O}]$. These observations will be formally proved later (see Theorem 2.3).
(7) If there are more than two events, the independence of all events requires that any subset of all events must be independent. Note that here subsets should be iteratively computed (i.e., the requirement is iterative for all subsets). Taking three events as an example, namely A_1, A_2, A_3, they are independent if and only if $\Pr[A_1, A_2] = \Pr[A_1] \times \Pr[A_2]$, $\Pr[A_2, A_3] = \Pr[A_2] \times \Pr[A_3]$, $\Pr[A_1, A_3] = \Pr[A_1] \times \Pr[A_3]$, and $\Pr[A_1, A_2, A_3] = \Pr[A_1] \times \Pr[A_2] \times \Pr[A_3]$.
(8) Naturally, it will be interesting to explore how to represent the relation (especially in the above geometric graph consisting of rectangles quantitatively) between two events that are NOT independent. The natural question is whether the above graph can still work or not, which will be discussed later (e.g., in Section 3.2).

Theorem 2.3. If $\Pr[A, O] = \Pr[A] \times \Pr[O]$, then
(1) $\Pr[\overline{A}, O] = \Pr[\overline{A}] \times \Pr[O]$;
(2) $\Pr[A, \overline{O}] = \Pr[A] \times \Pr[\overline{O}]$;
(3) $\Pr[\overline{A}, \overline{O}] = \Pr[\overline{A}] \times \Pr[\overline{O}]$.

Proof.
(1) $\Pr[\overline{A}, O] = \Pr[O] - \Pr[A, O] = \Pr[O] - \Pr[A] \times \Pr[O] = \Pr[O] \times (1 - \Pr[A]) = \Pr[\overline{A}] \times \Pr[O]$.
(2) $\Pr[A, \overline{O}] = \Pr[A] - \Pr[A, O] = \Pr[A] - \Pr[A] \times \Pr[O] = \Pr[A] \times (1 - \Pr[O]) = \Pr[A] \times \Pr[\overline{O}]$.
(3) $\Pr[\overline{A}, \overline{O}] = \Pr[\overline{A}] - \Pr[\overline{A}, O] = \Pr[\overline{A}] - \Pr[\overline{A}] \times \Pr[O] = \Pr[\overline{A}](1 - \Pr[O]) = (1 - \Pr[A]) \times \Pr[\overline{O}] = \Pr[\overline{A}] \times \Pr[\overline{O}]$. □

The above theorem reveals that the independence of A and O implies the independence of the other three pairs: $\overline{A}$ and O, A and $\overline{O}$, and $\overline{A}$ and $\overline{O}$.

2.4. Conditional Probability

Conditional probability concerns the probability of one event occurring given that another event has occurred. If event O occurs, then A occurs with probability denoted as $\Pr[A|O]$, and $\Pr[A|O] = \Pr[A, O] / \Pr[O]$.

Remark 2.4.

(1) From the above equation, $\Pr[O] > 0$ is thus assumed once $\Pr[A|O]$ is defined.
(2) $\Pr[A|O]$ is indeed a ratio—the probability that both A and O occur divided by the probability that O occurs.
(3) It is worth noting that the rationale behind the above division computation is that the conditional probability must be "normalized into" a new domain (namely, O) that is different from the original full domain (namely, Ω) related to event O.
(4) In the above equation, $\Pr[A|O]$ is a ratio (related to a new domain, namely, O); $\Pr[A, O]$, $\Pr[A]$, and $\Pr[O]$ are probabilities in the same original domain, namely, Ω. That is, when comparing $\Pr[A]$ and $\Pr[A|O]$, one should be aware that the domain has changed (yet, $\Pr[A|O] \times \Pr[O] = \Pr[A, O]$ is in the domain Ω). Usually, $\Pr[A]$ is called prior probability, which may be "imagined" or "conjectured," but $\Pr[A|O]$ can be looked at as an amended probability for re-estimating the probability of A after event O occurs, which is the so-called posterior probability. It is worth noting that the selection of O should be subtle—we will explain this later (e.g., in Section 2.7).

Concretely, we may state conditional probability in an observation-and-hypothesis form: If (observation) event O occurs, then (hypothesis) event A occurs with probability $\Pr[A|O] \in [0, 1]$. For example, let A be the event that a disease occurs. $\Pr[A]$ is the probability of the disease occurring. Let O be the event that the testing result for the disease is positive. $\Pr[O]$ is the probability that O occurs. $\Pr[A|O]$ is the probability that A occurs when O occurs. That is, when the testing result is positive, the probability of being infected with the disease.

As $\Pr[A|O] = \Pr[A, O]/\Pr[O]$, for visualizing it, $\Pr[A|O]$ is indeed a "ratio" instead of a "probability"; it is a portion of $\Pr[A, O]$ over $\Pr[O]$. Nonetheless, it is not a probability over the original full domain (i.e., Ω); instead, it is a portion over a new basis (i.e., $\Pr[O]$). Note that we hereby intentionally and explicitly use the notation "ratio" to denote conditional probability to distinguish it from the notation "probability."

Geometric Representation. The geometric representations are shown in Fig. 5 [Figure 5: see original paper].

$\Pr[A, O] = S_{O_1A_2A_5O_4}$, which is the area of rectangle $\square O_1A_2A_5O_4$. $\Pr[A, O]$ is a probability over the full domain, denoted as an area in the geometric graph.

$\Pr[A|O] = \Pr[A, O]/\Pr[O] = S_{O_1A_2A_5O_4}/S_{O_1O_2O_3O_4}$, where $S_{O_1O_2O_3O_4}$ is the area of rectangle $\square O_1O_2O_3O_4$ that equals the probability that the observation event occurs.

Note that $\Pr[A|O]$ is NOT an area; instead, it is a ratio of two areas (rectangles)—the area of the rectangle for $\Pr[A, O]$ over the area of the rectangle for $\Pr[O]$. It cannot be shown by an area of a rectangle in the graph; it can only be shown as a ratio of two areas from two corresponding rectangles.

Fortunately, the ratio of two areas can be viewed as the ratio between two lines

from two rectangles (e.g., $|O_1A_2|$ over $|O_1O_2|$). As $|O_1O_2| = 1$, this ratio can be viewed as $|O_1A_2|$. Similarly, $\Pr[O|A]$ can be viewed as $|O_1O_4|$; $\Pr[\overline{A}|O]$ can be viewed as $|A_2O_2|$; $\Pr[\overline{O}|A]$ can be viewed as $|O_4A_4|$.

[Figure 5: see original paper]

Remark 2.5.
(1) $\Pr[\overline{A}|O] = 1 - \Pr[A|O]$. Since $\Pr[A|O]$ is a ratio, $\Pr[\overline{A}|O]$ is a complementary proportion of ratio $\Pr[A|O]$ to 1.
(2) $\Pr[\overline{A}|O] = \Pr[\overline{A}, O]/\Pr[O] = S_{A_2O_2O_3A_5}/S_{O_1O_2O_3O_4}$.
(3) Similarly, $\Pr[\overline{O}|A] = 1 - \Pr[O|A]$.
$\Pr[\overline{O}|A] = \Pr[\overline{O}, A]/\Pr[A] = S_{O_4A_5A_3A_4}/S_{O_1A_2A_3A_4}$.

2.5. Special Independence and Inequality

In this section, we discuss some special cases.

Proposition 2.6. If $\Pr[A] = 1$, then any event O is independent of A.

Proof. Since $\Pr[A] = 1$, then $A \cap O = O$. Thus, $\Pr[A, O] = \Pr[O] = 1 \times \Pr[O] = \Pr[A] \times \Pr[O]$. □

Generally, we have the following proposition.

Proposition 2.7. Any event is independent with the event with probability 1 (or the event with full domain).

Proof. Straightforward. It is an alternative statement of Proposition 2.6. □

Proposition 2.8. If $A \subset O$ (i.e., if O occurs, then A occurs), $\Pr[A] \neq 0$, and $\Pr[O] \neq 1$, then any event O is not independent of event A.

Proof. Since $A \subset O$, $\Pr[A, O] = \Pr[A]$. $\Pr[A, O] = \Pr[A] \neq \Pr[A] \times \Pr[O]$, as $\Pr[A] \neq 0$ and $\Pr[O] \neq 1$. □

The next proposition is indeed a corollary of Proposition 2.8.

Proposition 2.9. Suppose $\Pr[A \cap O] \neq 0$. Event $A \cap O$ (namely, A and O both occur) and event O (or A) are not independent if $\Pr[O] \neq 1$ (or $\Pr[A] \neq 1$).

Proof. $\Pr[A, O, O] = \Pr[A, O]$. If $\Pr[A, O] \times \Pr[O] = \Pr[A, O]$, then $\Pr[O] = 1$, which is a contradiction. Similarly, $\Pr[A] = 1$ for the case of A. □

Proposition 2.10. Suppose $\Pr[A] \neq 0$ and $\Pr[O] \neq 0$. Then $\Pr[A|O] = 0$ if and only if $\Pr[O|A] = 0$.

Proof. $\Pr[A|O] = 0 \Leftrightarrow \Pr[A, O] = 0 \Leftrightarrow \Pr[O|A] = \Pr[A, O]/\Pr[A] = 0$. □

We can also analyze the inequalities between two probabilities.

Remark 2.11.
(1) As $0 \leq \Pr[A, O] \leq \Pr[O]$, then $0 \leq \Pr[A, O]/\Pr[O] = \Pr[A|O] \leq 1$. If and only if $\Pr[A, O] = 0$, then $\Pr[A|O] = 0$; if and only if $\Pr[A, O] = \Pr[O]$, then

$\Pr[A|O] = 1$.
(2) $\Pr[A|O]$ has no inequality relation with $\Pr[O]$. In other words, $\Pr[A|O]$ could be a ratio of any value in $[0, 1]$; $\Pr[O]$ could be a probability of any value in $(0, 1]$.
(3) Similarly, as $0 \leq \Pr[A, O] \leq \Pr[A]$, then $0 \leq \Pr[A, O]/\Pr[A] = \Pr[O|A] \leq 1$. $\Pr[A|O]$ has no inequality relation with $\Pr[A]$. That is, $\Pr[A|O]$ could be any value in $[0, 1]$; $\Pr[A]$ could be any value in $[0, 1]$.

The next propositions discuss the lower bound, upper bound, and properties of $\Pr[A|O]$.

Proposition 2.12. $\Pr[A|O] = 0$ and $\Pr[O|A] = 0$ if and only if $\Pr[A, O] = 0$.

Proof. $\Pr[A, O] = 0 \Leftrightarrow \Pr[A|O] = \Pr[A, O]/\Pr[O] = 0/\Pr[O] = 0$. The case for $\Pr[O|A] = 0$ can be proved similarly. □

Proposition 2.13. $\Pr[A|O] = 1$ if and only if $\Pr[A, O] = \Pr[O]$. $\Pr[O|A] = 1$ if and only if $\Pr[A, O] = \Pr[A]$.

Proof. $\Pr[A, O] = \Pr[O] \Leftrightarrow \Pr[A|O] = \Pr[A, O]/\Pr[O] = \Pr[O]/\Pr[O] = 1$. The case for $\Pr[O|A] = 1$ can be proved similarly. □

Proposition 2.14. If A and O are independent, then $\Pr[A|O] = \Pr[A]$ and $\Pr[O|A] = \Pr[O]$.

Proof. If A and O are independent, then $\Pr[A, O] = \Pr[A] \times \Pr[O]$. Thus, $\Pr[A|O] = (\Pr[A] \times \Pr[O])/\Pr[O] = \Pr[A]$. The case for $\Pr[O|A] = \Pr[O]$ can be proved similarly. □

Remark 2.15.
(1) Proposition 2.12 states that if and only if A and O are mutually exclusive, then $\Pr[A|O] = 0$ and $\Pr[O|A] = 0$.
(2) Proposition 2.13 states that if O occurs then A will occur, we have $\Pr[A|O] = 1$; if A occurs then O will occur, we have $\Pr[O|A] = 1$.
(3) Proposition 2.14 states that if A and O are independent, then A occurs no matter whether O occurs or not; O occurs no matter whether A occurs or not.

2.6. Other Joint or Conditional Probabilities and Their Independence

In the geometric graph representation, conditional probabilities (i.e., $\Pr[s|t]$, $s, t \in \{A, O, \overline{A}, \overline{O}\}$, $s \neq t$) are ratios of two areas—one is the area of the rectangle for joint probability $\Pr[s, t]$, the other is the area of the rectangle for $\Pr[t]$.

Interestingly, given $\Pr[A] = x$, $\Pr[O] = y$, and $\Pr[A|O] = \alpha$, the other related joint and conditional probabilities can be computed. The following remarks state the computational process (we still use Fig. 5 to explain).

Remark 2.16.
(1) $\Pr[\overline{A}|O]$. $\Pr[\overline{A}|O] = 1 - \Pr[A|O] = 1 - \alpha$. Geometrically, it is a ratio

of two areas—the area of rectangle $\square A_2O_2O_3A_5$ over the area of rectangle $\square O_1O_2O_3O_4$.
(2) $\Pr[\overline{A}, O]$. $\Pr[\overline{A}, O] = \Pr[\overline{A}|O] \times \Pr[O] = (1-\alpha) \times y = y - \alpha \times y$. Geometrically, $\Pr[\overline{A}, O] = S_{A_2O_2O_3A_5}$, which is the area of rectangle $\square A_2O_2O_3A_5$. In other words, it is the complementary area in rectangle $\square O_1O_2O_3O_4$ but outside rectangle $\square O_1A_2A_5O_4$. $\Pr[\overline{A}, O]$ is a probability over the full domain, denoted as an area in the geometric graph.
(3) $\Pr[O|A]$. $\Pr[O|A] = \Pr[O, A]/\Pr[A] = \Pr[A|O] \times \Pr[O]/\Pr[A] = \alpha \times y/x = \frac{\alpha y}{x}$. It is a ratio of two areas—the area of rectangle $\square O_1A_2A_5O_4$ over the area of rectangle $\square O_1A_2A_3A_4$. The difference between $\Pr[O|A]$ and $\Pr[A|O]$ is that the same area (i.e., $\square O_1A_2A_5O_4$) is over different areas (i.e., $\square O_1O_2O_3O_4$, $\square O_1A_2A_3A_4$, respectively).
(4) $\Pr[\overline{O}|A]$. $\Pr[\overline{O}|A] = 1 - \Pr[O|A] = 1 - \alpha \times y/x = 1 - \frac{\alpha y}{x}$. It is a ratio of two areas—the area of rectangle $\square O_4A_5A_3A_4$ over the area of rectangle $\square O_1A_2A_3A_4$.
(5) $\Pr[A, \overline{O}]$. $\Pr[A, \overline{O}] = \Pr[\overline{O}|A] \times \Pr[A] = (1 - \Pr[O|A]) \times \Pr[A] = (1 - \alpha \times y/x) \times x = x - \alpha \times y = S_{O_1A_2A_3A_4} - S_{O_1A_2A_5O_4}$. Note that it is an area of rectangle $\square O_4A_5A_3A_4$.
(6) $\Pr[A|\overline{O}]$. $\Pr[A|\overline{O}] = \Pr[A, \overline{O}]/\Pr[\overline{O}] = (x - \alpha \times y)/(1-y) = \frac{x-\alpha y}{1-y}$. Note that it is a ratio of two areas—the area of rectangle $\square O_4A_5A_3A_4$ over the area of rectangle $\square O_4O_3P_5A_4$.
(7) $\Pr[\overline{A}, \overline{O}]$. $\Pr[\overline{A}, \overline{O}] = \Pr[\overline{A}|\overline{O}] \times \Pr[\overline{O}] = (1 - \Pr[A|\overline{O}]) \times (1-y) = (1 - \frac{x-\alpha y}{1-y}) \times (1-y) = (1-y) - (x - \alpha \times y) = S_{O_4O_3P_5A_4} - S_{O_4A_5A_3A_4}$. Indeed, it is the area of rectangle $\square A_5O_3P_5A_3$.
(8) $\Pr[\overline{A}|\overline{O}]$. $\Pr[\overline{A}|\overline{O}] = \Pr[\overline{A}, \overline{O}]/\Pr[\overline{O}] = ((1-y) - (x - \alpha \times y))/(1-y) = 1 - (x - \alpha \times y)/(1-y) = 1 - \frac{x-\alpha y}{1-y} = S_{A_5O_3P_5A_3}/S_{O_4O_3P_5A_4}$. Indeed, it is the ratio of the area of rectangle $\square A_5O_3P_5A_3$ over the area of rectangle $\square O_4O_3P_5A_4$.

Or, $\Pr[\overline{A}|\overline{O}] = 1 - \Pr[A|\overline{O}] = 1 - \frac{x-\alpha y}{1-y}$.

Geometric Representation. The geometric representation is shown in Fig. 6 [Figure 6: see original paper]. The joint probabilities are shown in the graph and represented by x, y, and α. The conditional probabilities can be computed by calculating corresponding ratios.

[Figure 6: see original paper]

Lemma 2.17. Given $\Pr[A]$, $\Pr[O]$, and $\Pr[A|O]$, four joint probabilities—namely, $\Pr[A, O]$, $\Pr[\overline{A}, O]$, $\Pr[A, \overline{O}]$, $\Pr[\overline{A}, \overline{O}]$—and the other seven conditional probabilities—namely, $\Pr[\overline{A}|O]$, $\Pr[A|\overline{O}]$, $\Pr[\overline{A}|\overline{O}]$, $\Pr[O|A]$, $\Pr[\overline{O}|A]$, $\Pr[O|\overline{A}]$, $\Pr[\overline{O}|\overline{A}]$—can be computed.

Proof. Straightforward (recall Remark 2.16). □

Lemma 2.18. Generally, given two probabilities and their conditional probability, then m_1 (i.e., the number) joint probabilities can be computed, where $m_1 = C(2,1) \times C(2,1) = 2 \times 2 = 4$. The other related $m_2 - 1$ (i.e., the number) conditional probabilities can be computed, where $m_2 = C(4,1) \times C(2,1) =$

$4 \times 2 = 8$.

Proof. Straightforward. $m_1 = 4$ is the combinatorial number equal to selecting one from s and $\overline{s}$, where $s \in \{A, O\}$, and then selecting one from t and $\overline{t}$ where $t \neq s$, $t \in \{A, O\}$. $m_2 = 8$ is the combinatorial number equal to selecting one from A, $\overline{A}$, O, and $\overline{O}$, and selecting one from two. Namely, $\Pr[A|O]$, $\Pr[\overline{A}|O]$, $\Pr[O|A]$, $\Pr[\overline{O}|A]$, $\Pr[A|\overline{O}]$, $\Pr[\overline{A}|\overline{O}]$, $\Pr[O|\overline{A}]$, $\Pr[\overline{O}|\overline{A}]$. □

Lemma 2.19. Given $\Pr[A|O]$, $\Pr[O|A]$, $\Pr[A|\overline{O}]$, and $\Pr[O|\overline{A}]$, the other 4 conditional probabilities can be computed (indeed, they are complementary to these 4 probabilities correspondingly).

Proof. Straightforward. $\Pr[\overline{A}|O] = 1 - \Pr[A|O]$, $\Pr[\overline{O}|A] = 1 - \Pr[O|A]$, $\Pr[\overline{A}|\overline{O}] = 1 - \Pr[A|\overline{O}]$, $\Pr[\overline{O}|\overline{A}] = 1 - \Pr[O|\overline{A}]$. □

Since one conditional probability may deduce the others, we are interested in whether 4 joint probabilities and 8 conditional probabilities are independent, or the LEAST number of them needed to determine all (i.e., similar to the concept "rank" in linear algebra).

Lemma 2.20. Given $\Pr[A]$, $\Pr[O]$, and $r \in \{\Pr[O|A], \Pr[A|O], \Pr[O|\overline{A}]\}$, then the others (i.e., 4 joint probabilities and the other 7 conditional probabilities) can be computed.

Proof. Let $\Pr[A] = x$, $\Pr[O] = y$. If $\Pr[A|O]$ is available, then the problem reduces to Lemma 2.17 and the proof is done. Next, we prove that $\Pr[A|O]$ can be made available as follows:

(1) If $r = \Pr[O|A]$, then $\Pr[A, O] = r \times x$. $\Pr[A|O] = \Pr[A, O]/\Pr[O] = r \times x/y$. Thus, $\Pr[A|O]$ is available.

(2) If $r = \Pr[A|\overline{O}]$, then $\Pr[A, \overline{O}] = r \times (1 - \Pr[O]) = r \times (1 - y)$. $\Pr[\overline{A}, O] = \Pr[A] - \Pr[A, \overline{O}] = (1 - \Pr[A]) - r \times (1 - \Pr[O]) = (1 - x) - r \times (1 - y)$. $\Pr[A|O] = \Pr[A, O]/\Pr[O] = \frac{(1-x)-r\times(1-y)}{1-y}$. $\Pr[\overline{A}|O] = 1 - \Pr[A|O] = 1 - \frac{(1-x)-r\times(1-y)}{1-y}$. Thus, $\Pr[A|O]$ is available.

(3) If $r = \Pr[O|\overline{A}]$, then $\Pr[O, \overline{A}] = r \times (1-x)$. $\Pr[\overline{A}, O] = (1-x) - r \times (1-x) = (1-x) \times (1-r)$. $\Pr[A|O] = 1 - \Pr[\overline{A}|O] = 1 - \frac{(1-x)\times(1-r)}{1-y}$. Thus, $\Pr[A|O]$ is available. □

Lemma 2.21. Given $\Pr[A|O] = r$, $\Pr[O|A] = s$, and $\Pr[A|\overline{O}] = t$, then all the other conditional probabilities and joint probabilities can be computed.

Proof. Let $\Pr[A] = x$, $\Pr[O] = y$, and $\Pr[A|O] = r$, $\Pr[O|A] = s$, $\Pr[A|\overline{O}] = t$. Then $\Pr[A, O] = r \times y$, $\Pr[O|A] = s = r \times y/x$, thus $x = r/s \times y$. $t = \Pr[A|\overline{O}] = 1 - \Pr[\overline{A}|\overline{O}] = 1 - \Pr[\overline{A}, \overline{O}]/(1-y) = 1 - (\Pr[A] - \Pr[A, O])/(1-y) = 1 - \frac{x - r\times y}{1-y}$.

Hence, x and y can be represented by r, s, and t. More specifically, $t = 1 - \frac{x-r\times y}{1-y} \Leftrightarrow (1-t)(1-y) = x - r \times y \Leftrightarrow 1 - t - y + t \times y = r/s \times y - r \times y \therefore x =$

$r/s \times y \Leftrightarrow y = \frac{r/s-r-t+1}{\cdots}$. Thus, $\Pr[A]$ and $\Pr[O]$ can both be represented by r, s, t, and together with $\Pr[A|O] = r$, the proof is done due to Lemma 2.17. □

Lemma 2.22. Given $r, s, t \in \{\Pr[A|O], \Pr[O|A], \Pr[A|\overline{O}], \Pr[O|\overline{A}]\}$, then all the other conditional probabilities and joint probabilities can be computed.

Proof. Let $\phi_1 = \Pr[A|O]$, $\phi_2 = \Pr[O|A]$, $\phi_3 = \Pr[A|\overline{O}]$, $\phi_4 = \Pr[O|\overline{A}]$. Let $\Pr[A] = x_1$, $\Pr[O] = x_2$, $\Pr[A, O] = x_3$, and $\Pr[\overline{A}, O] = x_4$. Thus,

$$\begin{cases} \phi_1 = \Pr[A, O]/\Pr[O] = x_3/x_2 \\ \phi_2 = \Pr[A, O]/\Pr[A] = x_3/x_1 \\ \phi_3 = \Pr[\overline{A}, O]/\Pr[\overline{O}] = x_4/(1-x_2) \\ \phi_4 = \Pr[\overline{A}, O]/\Pr[\overline{A}] = x_4/(1-x_1) \end{cases}$$

Note that $\Pr[A] - \Pr[A, O] + \Pr[\overline{A}, O] = \Pr[A, O] + \Pr[\overline{A}, O] = \Pr[O] \Rightarrow x_1 - x_3 + x_4 = 1 - x_2 \Rightarrow x_1 + x_2 - x_3 + x_4 = 1$. Thus, only three of $\{x_1, x_2, x_3, x_4\}$ are independent.

If and only if any three of $\{\phi_1, \phi_2, \phi_3, \phi_4\}$ are given, three of $\{x_1, x_2, x_3, x_4\}$ can be determined, and then the remaining one can be computed by $x_1 + x_2 - x_3 + x_4 = 1$. □

Lemma 2.23. Suppose $\Pr[A] = x_1$, $\Pr[O] = x_2$, $\Pr[A, O] = x_3$, and $\Pr[\overline{A}, O] = x_4$. If and only if any three of $\{x_1, x_2, x_3, x_4\}$ are given, then ϕ_j $(j = 1, 2, 3, 4)$ can be determined.

Proof. If and only if any three of $\{x_1, x_2, x_3, x_4\}$ are given, then the remaining one can be computed by $x_1 + x_2 - x_3 + x_4 = 1$, and ϕ_j $(j = 1, 2, 3, 4)$ can be determined by the equations above. □

Theorem 2.24. Given any three of $\{\Pr[A], \Pr[O], \Pr[A|O], \Pr[O|A], \Pr[A|\overline{O}], \Pr[O|\overline{A}]\}$, then all conditional probabilities and all joint probabilities can be computed.

Proof. If the three items are all in $\{\phi_1, \phi_2, \phi_3, \phi_4\}$, then this case is proved in Lemma 2.22. Next, we only need to prove that three items in $\{\phi_1, \phi_2, \phi_3, \phi_4\}$ can still be computed even in the other cases.

If two in $\{\phi_1, \phi_2, \phi_3, \phi_4\}$ and one in $\{x_1, x_2\}$ are given, then it is sufficient to solve one remaining ϕ_i that is not given due to the equations above. The reason is that there are two unknowns in $\{x_1, x_2, x_3, x_4\}$, which can be solved by two equations in $\{\phi_1, \phi_2, \phi_3, \phi_4\}$.

If one in $\{\phi_1, \phi_2, \phi_3, \phi_4\}$ and two in $\{x_1, x_2\}$ are given, then it is sufficient to solve two remaining ones in $\{\phi_1, \phi_2, \phi_3, \phi_4\}$ that are not given due to the equations above. The reason is that two in $\{x_1, x_2\}$ are given, so there is only one unknown in $\{x_1, x_2, x_3, x_4\}$, which can be solved by the given ϕ_i. □

Theorem 2.25. Given any three of $\{\Pr[A], \Pr[O], \Pr[A, O], \Pr[\overline{A}, O], \Pr[A|O], \Pr[O|A], \Pr[A|\overline{O}], \Pr[O|\overline{A}]\}$, except for the cases $\{\phi_1, x_2, x_3\}$, $\{\phi_2, x_1, x_3\}$, $\{\phi_3, x_2, x_4\}$, $\{\phi_4, x_1, x_4\}$, then

all remaining conditional probabilities, all remaining joint probabilities, and remaining probabilities (e.g., $\Pr[A]$, $\Pr[O]$, if not given) can be computed.

Proof. Straightforward due to Lemma 2.22 and Lemma 2.23. Due to the dependence (i.e., $x_1+x_2-x_3+x_4=1$), there are only three independent variables. Due to the equations above, if and only if three items in $\{\phi_1, \phi_2, \phi_3, \phi_4, x_1, x_2, x_3, x_4\}$ are available, then three variables can be determined. Certainly, repeated items should be removed, e.g., $\{\phi_1, x_2, x_3\}$, $\{\phi_2, x_1, x_3\}$, $\{\phi_3, x_2, x_4\}$, $\{\phi_4, x_1, x_4\}$, because in these cases the three items are not independent and thus they degenerate to two items. □

Indeed, we can provide another proof for Theorem 2.25 without relying on any lemmas as follows:

Proof. Let $\phi_1 = \Pr[A|O]$, $\phi_2 = \Pr[O|A]$, $\phi_3 = \Pr[A|\overline{O}]$, $\phi_4 = \Pr[O|\overline{A}]$. Let $\Pr[A] = x_1$, $\Pr[O] = x_2$, $\Pr[A, O] = x_3$, and $\Pr[\overline{A}, O] = x_4$.

Thus,

$$\begin{cases} \phi_1 = x_3/x_2 \\ \phi_2 = x_3/x_1 \\ \phi_3 = x_4/(1-x_2) \\ \phi_4 = x_4/(1-x_1) \\ x_1 + x_2 - x_3 + x_4 = 1 \end{cases}$$

Given any three of $\{\phi_1, \phi_2, \phi_3, \phi_4, x_1, x_2, x_3, x_4\}$, we have 5 conditional equations to determine 5 remaining variables, except for the 4 cases in which the three items are in the same conditional equation. (The reason is that if three items are in the same conditional equation, e.g., the first four, then the conditional equations will reduce to 4.) □

Geometric Representation. The geometric representation for three independent variables among the eight (i.e., $\{\phi_1, \phi_2, \phi_3, \phi_4, x_1, x_2, x_3, x_4\}$) is shown in Fig. 7 [Figure 7: see original paper]. Usually, $\Pr[A] = x_1$, $\Pr[O] = x_2$, and one conditional probability are given, and then all the other conditional probabilities and joint probabilities can be computed.

[Figure 7: see original paper]

2.7. Other Notations for Joint Probabilities and Conditional Probabilities

A typical context is that O is an observation for verifying a certain hypothesis A. We thus have the following remarks on the special notions for joint probabilities in this context, especially in machine learning contexts.

Remark 2.26.
(1) $\Pr[A, O]$ is the so-called true positive. $\Pr[\overline{A}, O]$ is the so-called false positive.

$\Pr[A, \overline{O}]$ is the so-called false negative. $\Pr[\overline{A}, \overline{O}]$ is the so-called true negative.
(2) $\Pr[A|O] = \Pr[A, O]/\Pr[O] = \Pr[A, O]/(\Pr[A, O] + \Pr[\overline{A}, O])$ is the so-called "precision," which indicates the proportion that observation O can discover event A—the number of A discovered by observation O over the total number of O. Roughly speaking, it implies the "efficiency" of the observation in terms of the proportion of "valid" observations that can confirm the hypothesis over all observations, e.g., the proportion of positive testing results that imply or can detect infected persons over all positive testing results. Simply speaking, suppose there are 100 persons with positive results in observations, but only 80 persons are indeed infected; then the "precision" is 80%.
(3) $\Pr[O|A] = \Pr[O, A]/\Pr[A] = \Pr[O, A]/(\Pr[O, A] + \Pr[\overline{O}, A])$ is the so-called "recall," which indicates the proportion that event A can be discovered by observation O—the number of A discovered by observation O over the total number of A. Roughly speaking, it implies the "effectiveness" of the observation in terms of the proportion of "valid" observations that can confirm the hypothesis over all hypotheses, e.g., the proportion of positive test results that imply or can detect infected persons over all infected persons. Simply speaking, suppose there are 100 persons who are indeed infected; if 70 persons have positive test results in observations, then the "recall" is 70%.
(4) $(\Pr[A, O]+\Pr[\overline{A}, \overline{O}])/(\Pr[A, O]+\Pr[\overline{A}, O]+\Pr[A, \overline{O}]+\Pr[\overline{A}, \overline{O}]) = \Pr[A, O]+\Pr[\overline{A}, \overline{O}]$ is the so-called "accuracy." For example, the trustworthiness of whether a person is infected whose test strip is positive, plus the trustworthiness of whether a person is not infected whose test result is negative. It implies the trustworthiness (or "value") of one observation. Besides, "accuracy" is a probability over the full domain, yet "precision" and "recall" are ratios (conditional probabilities).
(5) $F_1 = 2/(\frac{1}{\Pr[O|A]} + \frac{1}{\Pr[A|O]})$ is a combinational evaluation of both $\Pr[O|A]$ and $\Pr[A|O]$.
(6) It is worth noting that in the contexts of machine learning, $\Pr[\overline{A}|O]$ and $\Pr[\overline{O}|A]$ are usually ignored. Indeed, $\Pr[\overline{A}|O]$ can be looked at as the "precision" of negative results. $\Pr[\overline{O}|A]$ can be looked at as the "recall" of negative results. Since we always concern the evaluation of positive results, these two may be ignored.
(7) If $\Pr[A|O] < \Pr[\overline{A}|O]$, then the negative result is more valuable in the observation than the positive result. Usually, we call the observing result that presents $\Pr[A|O] > \Pr[\overline{A}|O]$ the positive result and denote it as event O occurring.

The following propositions explore the possible relations (e.g., independence) between the above metrics.

Lemma 2.27. $\Pr[O|A]/\Pr[A|O] = \Pr[O]/\Pr[A]$.

Proof. $\Pr[O|A]/\Pr[A|O] = \frac{\Pr[O,A]/\Pr[A]}{\Pr[A,O]/\Pr[O]} = \Pr[O]/\Pr[A]$. □

Geometric Representation. The geometric representation for $\Pr[O]/\Pr[A]$ is a constant (see Fig. 8 [Figure 8: see original paper]).

[Figure 8: see original paper]

Lemma 2.28. If $\Pr[O]/\Pr[A] = C$ is fixed, then $\Pr[O|A] = C \times \Pr[A|O]$, i.e., $\Pr[O|A] \propto \Pr[A|O]$.

Proof. $\Pr[O|A]/\Pr[A|O] = \Pr[O]/\Pr[A] = C$. Thus, $\Pr[O|A] = C \times \Pr[A|O]$. □

Lemma 2.29. If $\Pr[O]/\Pr[A] = C$ is fixed, then $F_1 = \frac{2}{1+1/C} \times \Pr[A|O]$, i.e., $F_1 \propto \Pr[A|O]$.

Proof. $\Pr[O|A]/\Pr[A|O] = \Pr[O]/\Pr[A] = C$. Let $\Pr[O|A] = s$, $\Pr[A|O] = t$. Thus, $s/t = C$. $F_1 = 1/((1/s + 1/t)/2) = 2st/(s + t) = 2t/(1 + t/s) = 2t/(1 + 1/C) = \frac{2}{1+1/C} \times \Pr[A|O]$. Thus, $F_1 \propto \Pr[A|O]$. □

Theorem 2.30. The precision, recall, and F_1 will be dependent if $\Pr[O]/\Pr[A]$ is fixed (so that the three metrics can be reduced to ONE).

Proof. Straightforward. It is due to Lemma 2.27, Lemma 2.28, and Lemma 2.29. □

Remark 2.31.
(1) If $\Pr[O]/\Pr[A] = C$ is fixed, then $F_1 = \frac{2C}{C+1} \times \Pr[O|A]$, by Lemma 2.28 and Lemma 2.29.
(2) Recall that $\Pr[O|A]/\Pr[A|O] = \Pr[O]/\Pr[A]$. If $\Pr[O]/\Pr[A] \gg 1$, e.g., 10, and $\Pr[O|A] = \Pr[A|O] \times \Pr[O]/\Pr[A] = 10$. As $\Pr[O|A] \in (0, 1)$, then $\Pr[A|O] \in (0, 0.1)$.
(3) If $\Pr[O|A] \in [0.9, 1)$ and $\Pr[A|O] \in [0.9, 1)$, then $\Pr[O]/\Pr[A] \in (0.9/1, 1/0.9) \approx (0.9, 1.11)$. Usually, $\Pr[A]$ is an estimation; $\Pr[O]$ should thus take more samples as input to compute a better value. If precision and recall both perform sufficiently well, i.e., $\Pr[O|A], \Pr[A|O] \in [0.9, 1)$, then $\Pr[O]$ should approach $\Pr[A]$, i.e., $\Pr[O] \in (0.9 \times \Pr[A], \min(1.11 \times \Pr[A], 1))$.
(4) $\Pr[O]/\Pr[A]$ is usually not fixed when evaluating distinct observations denoted as O on the identical hypothesis denoted as A, and $\Pr[A]$ is estimated well, i.e., the variation of $\Pr[A]$ is in a small range. Otherwise, $\Pr[A]$ varies in a large range (due to rough estimation), and then different O may have the same $\Pr[O]/\Pr[A]$. In this situation, three metrics are not required.
(5) If $\Pr[O]/\Pr[A] = c \gg 1$, e.g., 10, and varies in a small range, e.g., $0.090 \sim 0.095$, then $\Pr[O|A] = c \times \Pr[A|O]$ will vary in a large range, e.g., $0.9 \sim 0.95$.

Corollary 2.32. If $\Pr[O_2] > \Pr[O_1]$ and $\Pr[A|O_2] > \Pr[A|O_1]$, then $\Pr[O_2|A] > \Pr[O_1|A]$ and $F_{1O_2} > F_{1O_1}$.

Proof. Recall that $C = \Pr[O]/\Pr[A]$, $\Pr[O|A] = C \times \Pr[A|O]$, and $F_1 = \frac{2}{1+1/C} \times \Pr[A|O]$. If $\Pr[O]$ increases and $\Pr[A]$ remains unchanged, then C increases. Thus, if $\Pr[A|O]$ increases, then $\Pr[O|A]$ will increase and F_1 will increase. □

Corollary 2.33. If $\Pr[O_2] < \Pr[O_1]$ and $\Pr[A|O_2] < \Pr[A|O_1]$, then $\Pr[O_2|A] < \Pr[O_1|A]$ and $F_{1O_2} < F_{1O_1}$.

Proof. Recall that $C = \Pr[O]/\Pr[A]$, $\Pr[O|A] = C \times \Pr[A|O]$, and $F_1 = \frac{2}{1+1/C} \times \Pr[A|O]$. If $\Pr[O]$ decreases and $\Pr[A]$ remains unchanged, then C decreases. Thus, if $\Pr[A|O]$ decreases, then $\Pr[O|A]$ will decrease and F_1 will decrease. □

Precision, recall, and F_1 are usually used for evaluating which observation method (event O) has better performance (in machine learning context). However, it is worth noting that we discover that the widely accepted three metrics may be related and may not be required.

Theorem 2.34. If $\Pr[O]$ increases and precision increases, then recall and F_1 both increase. If $\Pr[O]$ decreases and precision decreases, then recall and F_1 both decrease. Thus, the comparison of different observations (event O) does not need three metrics.

Proof. It is due to Corollary 2.32 and Corollary 2.33. □

If $\Pr[O]$ increases but precision decreases, or if $\Pr[O]$ decreases but precision increases, then recall and F_1 may increase or decrease, in which case the comparison of different observations does require three metrics.

Next, we are interested in—to what degree the increase of $\Pr[O]$ and to what degree the decrease of precision may still not cause recall and F_1 to decrease. That is, if $\Pr[O]$ increases to a sufficiently large value, even if precision decreases, recall and F_1 may still increase. Correspondingly, the vice versa case can be obtained similarly.

Corollary 2.35. Suppose $\Pr[O_2] > \Pr[O_1]$ and $\Pr[A|O_2] < \Pr[A|O_1]$.
(1) $\Pr[O_2|A] > \Pr[O_1|A]$ if and only if $\Pr[A, O_2] > \Pr[A, O_1]$;
(2) $F_{1O_2} > F_{1O_1}$ if and only if $\frac{\Pr[A]+\Pr[O_1]}{\Pr[A]+\Pr[O_2]} > \frac{\Pr[A,O_1]}{\Pr[A,O_2]}$.

Proof. Since $\Pr[O_2] > \Pr[O_1]$ and $\Pr[A|O_2] < \Pr[A|O_1]$, the comparison for recall is pending because $\Pr[A, O_2]$ and $\Pr[A, O_1]$ is pending.

(1) $\Pr[O_2|A] > \Pr[O_1|A] \Leftrightarrow \Pr[O_2] \times \Pr[A|O_2]/\Pr[A] > \Pr[O_1] \times \Pr[A|O_1]/\Pr[A] \Leftrightarrow \Pr[O_2] \times \Pr[A|O_2] > \Pr[O_1] \times \Pr[A|O_1] \Leftrightarrow \Pr[A, O_2] > \Pr[A, O_1]$.

(2) $F_{1O_2} > F_{1O_1} \Leftrightarrow \frac{2}{1+\Pr[A]/\Pr[O_2]} \times \Pr[A|O_2] > \frac{2}{1+\Pr[A]/\Pr[O_1]} \times \Pr[A|O_1] \Leftrightarrow (1 + \Pr[A]/\Pr[O_2]) \times \Pr[A|O_2] > (1 + \Pr[A]/\Pr[O_1]) \times \Pr[A|O_1] \Leftrightarrow \Pr[A|O_2] + \Pr[A] \times \Pr[A|O_2]/\Pr[O_2] > \Pr[A|O_1] + \Pr[A] \times \Pr[A|O_1]/\Pr[O_1] \Leftrightarrow \Pr[A|O_2] - \Pr[A|O_1] > \Pr[A] \times (\Pr[A|O_1]/\Pr[O_1] - \Pr[A|O_2]/\Pr[O_2]) \Leftrightarrow \Pr[A] \times (\Pr[A|O_2] \times \Pr[O_2] - \Pr[A|O_1] \times \Pr[O_1]) > (\Pr[A|O_1] - \Pr[A|O_2]) \times \Pr[O_1] \times \Pr[O_2] \Leftrightarrow \Pr[A] \times (\Pr[A, O_2] - \Pr[A, O_1]) > \Pr[A, O_1] \times \Pr[O_2] - \Pr[A, O_2] \times \Pr[O_1] \Leftrightarrow \Pr[A, O_2] \times (\Pr[A] + \Pr[O_1]) > \Pr[A, O_1] \times (\Pr[A] + \Pr[O_2]) \Leftrightarrow \frac{\Pr[A,O_2]}{\Pr[A]+\Pr[O_2]} > \frac{\Pr[A,O_1]}{\Pr[A]+\Pr[O_1]}$. □

Remark 2.36.
(1) In the above proof (1), we intentionally choose more steps to show the influence of conditional probabilities.
(2) In proof (2), $\Pr[A, O_2] > \Pr[A, O_1] \times \frac{\Pr[A]+\Pr[O_2]}{\Pr[A]+\Pr[O_1]}$.
(3) If and only if $\Pr[A, O_1] < \Pr[A, O_2] < \Pr[A, O_1] \times \frac{\Pr[A]+\Pr[O_2]}{\Pr[A]+\Pr[O_1]}$, then recall increases and F_1 decreases.
(4) If and only if $\Pr[A, O_2] > \Pr[A, O_1] \times \frac{\Pr[A]+\Pr[O_2]}{\Pr[A]+\Pr[O_1]}$, then recall and F_1 both increase.

Next, we explore the influence on recall and F_1 due to the changing degree of $\Pr[O]$ and precision.

Corollary 2.37. Suppose $\Pr[O_2] = \Pr[O_1] + \delta_1$ and $\Pr[A|O_2] = \Pr[A|O_1] - \delta_2$, where $\delta_1, \delta_2 \in (0, 1)$.
(1) $\Pr[O_2|A] > \Pr[O_1|A]$ if and only if $\delta_1 \times \Pr[A|O_1] - \delta_2 \times \Pr[O_1] > \delta_1 \times \delta_2$;
(2) $F_{1O_2} > F_{1O_1}$ if and only if $\Pr[A] > \frac{\Pr[O_1]^2+\delta_1\times\Pr[O_1]}{\delta_1\times\Pr[A|O_1]-\delta_2\times\Pr[O_1]-\delta_1\times\delta_2}$.

Proof.
(1) $\Pr[O_2|A] > \Pr[O_1|A] \Leftrightarrow \Pr[O_2] \times \Pr[A|O_2]/\Pr[A] > \Pr[O_1] \times \Pr[A|O_1]/\Pr[A] \Leftrightarrow (\Pr[O_1] + \delta_1) \times (\Pr[A|O_1] - \delta_2) > \Pr[O_1] \times \Pr[A|O_1] \Leftrightarrow \delta_1 \times \Pr[A|O_1] - \delta_2 \times \Pr[O_1] - \delta_1 \times \delta_2 > 0 \Leftrightarrow \delta_1 \times \Pr[A|O_1] - \delta_2 \times \Pr[O_1] > \delta_1 \times \delta_2$.

(2) $F_{1O_2} > F_{1O_1} \Leftrightarrow \Pr[A] \times \frac{\Pr[A|O_2]\times\Pr[O_2]-\Pr[A|O_1]\times\Pr[O_1]}{\Pr[O_1]\times\Pr[O_2]} > \Pr[A|O_1] - \Pr[A|O_2]$ (by Corollary 2.35(2)) $\Leftrightarrow \Pr[A] \times \frac{(\Pr[A|O_1]-\delta_2)\times(\Pr[O_1]+\delta_1)-\Pr[A|O_1]\times\Pr[O_1]}{\Pr[O_1]\times(\Pr[O_1]+\delta_1)} > \delta_2 \Leftrightarrow \Pr[A] \times \frac{\delta_1\times\Pr[A|O_1]-\delta_2\times\Pr[O_1]-\delta_1\times\delta_2}{\Pr[O_1]^2+\delta_1\times\Pr[O_1]} > \delta_2$. □

The following propositions state the dependence between precision, recall, and accuracy.

Proposition 2.38. $\Pr[A|O]$ is independent of $\Pr[O|A]$ if and only if $\Pr[O]/\Pr[A]$ is not fixed.

Proof. Straightforward. Recall Remark 2.16: given $\Pr[A] = x$, $\Pr[O] = y$, and $\Pr[A|O] = \alpha$, we have $\Pr[O|A] = \alpha \times y/x$. $(\alpha, \alpha \times y/x)$ are independent if and only if y/x is not fixed. □

Proposition 2.39. $\Pr[A, O] + \Pr[\overline{A}, \overline{O}]$ is independent of $\Pr[O|A]$ and $\Pr[A|O]$.

Proof. Straightforward. Recall Remark 2.16: given $\Pr[A] = x$, $\Pr[O] = y$, and $\Pr[A|O] = \alpha$, we have $\Pr[A, O] + \Pr[\overline{A}, \overline{O}] = (\alpha \times y) + ((1 - y) - (x - \alpha \times y)) = 2\alpha \times y + 1 - y - x$.

$\Pr[O|A] = \alpha\times y/x$. $(\alpha, 2\alpha\times y+1-y-x)$ is independent; $(\alpha\times y/x, 2\alpha\times y+1-y-x)$ is independent. □

Next, we depict three comparative graphs for easily understanding the above three metrics—precision, recall, and accuracy—in a geometric graph.

Geometric Representation. Fig. 9 [Figure 9: see original paper] shows which $\Pr[A|O]$ is larger in a geometric manner.

[Figure 9: see original paper]

Geometric Representation. Fig. 10 [Figure 10: see original paper] shows which $\Pr[O|A]$ is larger in a geometric way.

[Figure 10: see original paper]

Geometric Representation. Fig. 11 [Figure 11: see original paper] shows which $\Pr[A, O] + \Pr[\overline{A}, \overline{O}]$ is larger in a geometric way.

[Figure 11: see original paper]

3. Advanced Topics

3.1. (In)dependence between Observation and Hypothesis

Recall Proposition 2.14: if A and O are independent, then $\Pr[A|O] = \Pr[A]$. In other words, the observation event O does not change the probability of A. It can also be regarded as there being no influence on the original hypothesis upon observation.

Proposition 3.1. If $\Pr[A, O] > \Pr[A] \times \Pr[O]$, then A and O are not independent, and $\Pr[A|O] > \Pr[A]$.

Proof. Straightforward, due to Proposition 2.14. $\Pr[A|O] = \Pr[A, O]/\Pr[O] > \Pr[A] \times \Pr[O]/\Pr[O] = \Pr[A]$. □

It can be understood that the observation of O increases the probability of A in the original hypothesis. For example, the prior probability is estimated as $\Pr[A]$, yet the later observation increases the probability of A (e.g., supporting hypothesis A). That is the impact of observation on the original hypothesis, or the probability of the original hypothesis increases after the observation.

Proposition 3.2. If $\Pr[A, O] < \Pr[A] \times \Pr[O]$, then A and O are not independent, and $\Pr[A|O] < \Pr[A]$.

Proof. Straightforward, due to Proposition 2.14. $\Pr[A|O] = \Pr[A, O]/\Pr[O] < \Pr[A] \times \Pr[O]/\Pr[O] = \Pr[A]$. □

It can be understood that the observation of O decreases the probability of A in the original hypothesis. That is the impact of observation on the original hypothesis, or the decreasing amendment of the original hypothesis in terms of probability after the observation.

Actually, if $\Pr[A, O] > \Pr[A] \times \Pr[O]$, we also have $\Pr[O|A] > \Pr[O]$; if $\Pr[A, O] < \Pr[A] \times \Pr[O]$, we also have $\Pr[O|A] < \Pr[O]$. However, in the context, $\Pr[A|O]$ is more natural to understand than $\Pr[O|A]$ because the (later) observation O influences the probability of the (earlier) hypothesis A.

Proposition 3.3. If A and O are not independent, $\Pr[A|O]$ cannot be computed from $\Pr[A]$ and $\Pr[O]$.

Proof. It is a corollary of Proposition 2.14. $\Pr[A|O] = \Pr[A,O]/\Pr[O]$, but $\Pr[A,O]$ is unknown and cannot be computed from $\Pr[A]$ and $\Pr[O]$ because A and O are not independent. □

Geometric Representation. We use Fig. 12 [Figure 12: see original paper] to show the dependence by changing the area of $O_1A_2A_5O_4$, which equals $\Pr[A,O]$. If A and O are not independent, then $\Pr[A,O] \neq \Pr[A]\times\Pr[O] = 0.4\times0.6 = 0.24$.

Note that in this version of the graph, we intentionally keep the total area for $\Pr[O]$ in the left graph unchanged (i.e., not moving the horizontal line at O_3) for simplicity. Similarly, we intentionally keep the total area for $\Pr[A]$ in the right graph unchanged (i.e., not moving the vertical line at A_3 left). That is, when the area of $\Pr[A,O]$ is increased, the area for $\Pr[O,\overline{A}]$ or $\Pr[\overline{A},O]$ is not decreased for simplicity. (In contrast, the amended version is shown in Fig. 13 [Figure 13: see original paper] or Fig. 14 [Figure 14: see original paper] later.)

[Figure 12: see original paper]

To show the relation between $\Pr[A|O]$ and $\Pr[A]$, we have $\Pr[A|O] = \frac{\Pr[A]\times\Pr[O|A]}{\Pr[O]}$, which is discussed in the following remark.

Remark 3.4.
(1) $\Pr[A]$ is sometimes called prior probability. $\Pr[O]$ is sometimes called marginal likelihood. $\Pr[O|A]$ is sometimes called likelihood.
(2) $\Pr[A|O]$ is sometimes called posterior probability. If the likelihood $\Pr[O|A]/\Pr[O]$ is larger than 1, then $\Pr[A|O] = \Pr[A]\times\Pr[O|A]/\Pr[O] > \Pr[A]\times1 = \Pr[A]$.
(3) Interestingly, the above can be observed in Fig. 12. In the right half of this figure, initially, $\Pr[O|A] = S_{O_1A_2A_5O_4}/S_{O_1A_2A_3A_4} = 0.6$. If rectangle $\square O_1A_2A_5O_4$ becomes a little larger, then $\Pr[O|A] > 0.6$. Meanwhile, $\Pr[A|O] = |O_1A_2| > \Pr[A] = 0.4$.
(4) If $\Pr[A|O] > \Pr[A]$, then $\Pr[O|A]/\Pr[O] > 1$, which implies $\Pr[O|A] > \Pr[O]$. Simply speaking, if and only if O can "support" A, then A can "support" O. That is, $\Pr[A|O] > \Pr[A] \Leftrightarrow \Pr[O|A] > \Pr[O]$.
(5) It can be shown in Fig. 12 by the left and right graphs. The right graph is explained in (2). In the left graph, initially, $\Pr[A|O] = S_{O_1A_2A_5O_4}/S_{O_1O_2O_3O_4} = 0.4$. If rectangle $\square O_1A_2A_5O_4$ becomes a little larger, then $\Pr[A|O] > 0.4$. Meanwhile, $\Pr[O|A] = |O_1O_4| > \Pr[O] = 0.6$.
(6) $\Pr[O|A]/\Pr[O] > 1 \Leftrightarrow \Pr[O|A] > \Pr[O] \Leftrightarrow \Pr[O,A] > \Pr[O]\times\Pr[A]$, which again obtains Proposition 3.1.

3.2. Changing $\Pr[A,O]$ with Fixed $\Pr[A]$ and $\Pr[O]$

We further use a geometric graph to show the un-linkage between $\Pr[A|O]$ and $\Pr[O]$ in Fig. 13. That is, we can keep the area for $\Pr[O]$ (namely, rectangle $\square O_1O_2O_3O_4$) and the area for $\Pr[A]$ (namely, rectangle $\square O_1A_2A_3A_4$) unchanged, but change the area for $\Pr[A,O]$ (namely, rectangle $\square O_1A_2A_5O_4$).

The areas for $\Pr[A]$ and $\Pr[O]$ remain immutable, but the area for $\Pr[A,O]$ is

decreased from 0.24 to 0 in Case II. Or, we can increase $\Pr[A, O]$ from 0.24 to 0.4, i.e., $\Pr[A]$, in Case III. In Fig. 13, we keep the rectangle for $\Pr[A]$ unchanged and change the area for $\Pr[O]$ by moving the horizontal line (i.e., B_4O_3) up or down.

Geometric Representation. The extra benefit of this graph is that we can represent $\Pr[O|A]$ as a length of a line directly, i.e., $\Pr[O|A] = |A_2A_5|/|A_2A_3| = |A_2A_5|/1 = |A_2A_5| = |O_1O_4|$.

More specifically, in Case II, $\Pr[A, O]$ is changed to $0.24 - \Delta$ (i.e., the area of rectangle $\square O_1A_2A_5O_4$). $\Pr[A]$ already remains unchanged. To keep $\Pr[O]$ unchanged, the area for $\square A_2O_2O_3B_4$ is changed to $0.6-(0.24-\Delta) = 0.36+\Delta$. Thus, the area for $\Pr[O]$ still remains 0.6.

Similarly, in Case III, $\Pr[A, O]$ is changed to $0.24+\Delta$. To keep $\Pr[O]$ unchanged, the area for $\square A_2O_2O_3B_4$ is changed to $0.36-\Delta$. Thus, the area for representing $\Pr[O]$ is still 0.6.

Remark 3.5.
(1) In Case I, $\Pr[A, O] = \Pr[A] \times \Pr[O]$, which is the area of rectangle $\square O_1A_2A_5O_4$. $\Pr[O|A] = \Pr[A] \times \Pr[O]/\Pr[A] = \Pr[O]$. It can also be observed that $\Pr[O|A] = |A_2A_5|/|A_2A_3| = |A_2A_5|/1 = |A_2A_5| = 0.6$, which equals $\Pr[O] = 0.6$.
(2) In Case II, $\Pr[O|A] < \Pr[O]$, $\Pr[O|A] = |A_2A_5|/|A_2A_3| = |A_2A_5| < 0.6$.
(3) In Case III, $\Pr[O|A] > \Pr[O]$, $\Pr[O|A] = |A_2A_5|/|A_2A_3| = |A_2A_5| > 0.6$.

In contrast to Fig. 13, in Fig. 14 we simply keep the rectangle for $\Pr[O]$ unchanged and change the area for $\Pr[A]$ by moving the vertical line A_3B_4 left or right.

Geometric Representation. Similarly, an extra benefit in Fig. 14 is that $\Pr[A|O]$ can be represented by the length of a line directly, as $\Pr[A|O] = |O_1A_2|/|O_1O_2| = |O_1A_2|/1 = |O_1A_2| = |O_4A_5|$.

More specifically, in Case II, $\Pr[A, O]$ is changed to $0.24 - \Delta$ (i.e., the area of $\square O_1A_2A_5O_4$). $\Pr[O]$ already remains unchanged. To keep $\Pr[A]$ unchanged, the area for $\square O_4B_4A_3A_4$ is changed to $0.4 - (0.24 - \Delta) = 0.16 + \Delta$. Thus, the total area for $\Pr[A]$ is still 0.4. Similarly, in Case III, $\Pr[A, O]$ is changed to $0.24 + \Delta$. To maintain $\Pr[A]$ unchanged, the area for $\square O_4B_4A_3A_4$ is changed to $0.4 - (0.24 + \Delta) = 0.16 - \Delta$. Thus, the total area for $\Pr[A]$ is still 0.4.

Proposition 3.6. Suppose $\Pr[A] = x$, $\Pr[O] = y$, and $\Pr[A, O] = z$. Then $\min(z/x, z/y) \leq \Pr[O|A], \Pr[A|O] \leq \max(z/x, z/y) \leq 1$.

Proof. If A and O are independent, then $\Pr[A|O] = \Pr[A, O]/\Pr[O] = (x \times y)/y = x$. Similarly, $\Pr[O|A] = y$.

If A and O are not independent, $0 \leq z \leq \min(x, y)$. Thus, $\Pr[O|A] = z/x$, $\Pr[A|O] = z/y$. $\min(z/x, z/y) \leq \Pr[O|A], \Pr[A|O] \leq \max(z/x, z/y)$. As $z \leq \min(x, y)$, $\max(z/x, z/y) \leq \max(\min(x, y)/x, \min(x, y)/y) = 1$.

More specifically, if $x < y$, then $0 \le z \le x$, $z/y \le \Pr[O|A], \Pr[A|O] \le z/x$; if $x > y$, then $0 \le z \le y$, $z/x \le \Pr[O|A], \Pr[A|O] \le z/y$. □

3.3. How to Solve Calculating Questions by Filling Partitions in the Graph

An interesting or somewhat weird phenomenon is that although Bayes theorem has been learned, a large number of learners may still face considerable difficulties in calculations on conditional probabilities (e.g., in examinations). In this section, we thus propose a method (resembling an algorithm) for computing conditional probabilities using our aforementioned graph.

Without the graph, the calculation may not be easily understood or operated. Currently, such a step-by-step method is rarely proposed, or nonetheless a method like a binary tree method is rarely used. We hereby propose an easily operated and step-by-step approach by filling four partitions in our geometric graph to solve any calculating questions for conditional probabilities.

Here we take a typical calculating question as an example. A disease infects a person with probability $\Pr[A]$, which is sometimes called prior probability or hypothesis probability. A testing method is conducted to predict whether a person is infected with the disease. Usually, the most concerning quantities (i.e., the requested answers of most questions) are $\Pr[A|O]$ and $\Pr[A|\overline{O}]$, which can evaluate the trustworthiness (reliability) of the testing results (whether positive or negative). $\Pr[A|O]$ is the so-called precision (recall Remark 2.26), and $\Pr[A|\overline{O}]$ indicates the ratio—the number of infected persons with negative testing results over the number of persons with negative testing results.

It is worth noting that to calculate the above two conditional probabilities, two conditions must be given in the questions—either $\Pr[O|A]$ or $\Pr[\overline{O}|A]$ (usually the former), and either $\Pr[O|\overline{A}]$ or $\Pr[\overline{O}|\overline{A}]$ (usually the latter). The reason is that it is easy to compute $\Pr[O|A]$ by calculating the number of positive testing results after conducting tests on infected persons over the number of infected persons. Similarly, it is easy to compute $\Pr[\overline{O}|\overline{A}]$ by calculating the number of negative testing results after conducting tests on uninfected persons over the number of uninfected persons.

Therefore and in summary, in most calculating questions, $\Pr[A]$, $\Pr[O|A]$, and either $\Pr[O|\overline{A}]$ or $\Pr[\overline{O}|\overline{A}]$ must be given as known conditions; $\Pr[A|O]$ and $\Pr[A|\overline{O}]$ are the requested answers.

We have four partitions in the graph to represent $\Pr[O, A]$, $\Pr[O, \overline{A}]$, $\Pr[\overline{O}, A]$, and $\Pr[\overline{O}, \overline{A}]$ (see Fig. 15 [Figure 15: see original paper]). Usually, the left-upper partition can be calculated (likewise the rectangle is "filled") according to one given condition in the question such as $\Pr[O|A]$, and then the left-lower partition can be calculated (or "filled"). The right-upper partition can be calculated (or "filled") if the other given condition in the calculating question is $\Pr[O|\overline{A}]$, and then the right-lower partition can be calculated (or "filled"). Or, the right-

lower partition can be calculated (or "filled") if the other given condition in the calculating question is $\Pr[\overline{O}|\overline{A}]$, and then the right-upper partition can be calculated (or "filled"). Thereafter, all partitions are available.

In summary, we propose four steps as follows:

1. Given $\Pr[A] = a$ and $\Pr[O|A] = b$, compute $\Pr[O, A] = \Pr[A] \times \Pr[O|A] = a \times b$, and "fill" the left-upper partition in the graph.
2. Compute $\Pr[\overline{O}, A] = \Pr[A] \times \Pr[\overline{O}|A] = a \times (1-b)$, and "fill" the left-lower partition in the graph.
3. Given $\Pr[\overline{A}] = 1 - a$ and $\Pr[O|\overline{A}] = c$, compute $\Pr[O, \overline{A}] = \Pr[\overline{A}] \times \Pr[O|\overline{A}] = (1 - a) \times c$, and "fill" the right-upper partition in the graph. Compute $\Pr[\overline{O}, \overline{A}] = \Pr[\overline{A}] \times \Pr[\overline{O}|\overline{A}] = (1 - a) \times (1 - c)$, and "fill" the right-lower partition in the graph.
4. Therefore, the two results are available. That is, compute $\Pr[A|O] = \Pr[A, O]/\Pr[O] = a \times b/(a \times b + (1 - a) \times c)$, and $\Pr[A|\overline{O}] = \Pr[A, \overline{O}]/\Pr[\overline{O}] = a \times (1 - b)/(a \times (1 - b) + (1 - a) \times (1 - c))$.

If $\Pr[\overline{O}|\overline{A}] = c'$ instead of $\Pr[O|\overline{A}] = c$ is given, then the above Steps 3 and 4 will be replaced by the following two steps:

1. Given $\Pr[\overline{A}] = 1 - a$ and $\Pr[\overline{O}|\overline{A}] = c'$, compute $\Pr[\overline{O}, \overline{A}] = \Pr[\overline{A}] \times \Pr[\overline{O}|\overline{A}] = (1 - a) \times c'$, and "fill" the right-lower partition in the graph. Compute $\Pr[O, \overline{A}] = \Pr[\overline{A}] \times \Pr[O|\overline{A}] = (1 - a) \times (1 - c')$, and "fill" the right-upper partition in the graph.
2. Therefore, the two results are available. That is, compute $\Pr[A|O] = \Pr[A, O]/\Pr[O] = a \times b/(a \times b + (1 - a) \times (1 - c'))$, and $\Pr[A|\overline{O}] = \Pr[A, \overline{O}]/\Pr[\overline{O}] = a \times (1 - b)/(a \times (1 - b) + (1 - a) \times c')$. (Certainly, $c = 1 - c'$, so just replacing c with $1 - c'$ in Steps 3 and 4 can also obtain the results.)

Geometric Representation. The geometric representation is shown in Fig. 15, which illustrates how to calculate $\Pr[A|O]$ and $\Pr[A|\overline{O}]$ given $\Pr[A]$, $\Pr[O|A]$, and $\Pr[O|\overline{A}]$. That is, after given $\Pr[A]$, two conditions must be given in the questions—one for computing either partition in the left two (e.g., $\Pr[O|A]$) and the other for computing either partition in the right two (e.g., $\Pr[O|\overline{A}]$). Thus, the left two partitions can be calculated. (It is different from Fig. 6, which represents the computational relations between conditional probabilities.)

[Figure 15: see original paper]

Theorem 3.7. Given $\Pr[A] = a$, $\Pr[O|A] = b$, and $\Pr[O|\overline{A}] = c$, we have $\Pr[A|O] = a \times b/(a \times b + (1 - a) \times c)$ and $\Pr[A|\overline{O}] = a \times (1 - b)/(a \times (1 - b) + (1 - a) \times (1 - c))$.

Proof. Straightforward. $\Pr[A|O] = \Pr[A, O]/\Pr[O] = \Pr[A] \times \Pr[O|A]/(\Pr[A] \times \Pr[O|A] + \Pr[\overline{A}] \times \Pr[O|\overline{A}]) = a \times b/(a \times b + (1 - a) \times c)$. $\Pr[A|\overline{O}] =$

$\Pr[A,\overline{O}]/\Pr[\overline{O}] = \Pr[A] \times \Pr[\overline{O}|A]/(\Pr[A] \times \Pr[\overline{O}|A] + \Pr[\overline{A}] \times \Pr[\overline{O}|\overline{A}]) = a \times (1-b)/(a \times (1-b) + (1-a) \times (1-c))$. □

Corollary 3.8. $\Pr[A|O] \propto 1/c$, $\Pr[A|O] \propto b$, and $\Pr[A|O] \propto a$.

Proof. Straightforward. It is due to $\Pr[A|O] = a \times b/(a \times b + (1-a) \times c)$. That is, if a and b are fixed, $\Pr[A|O]$ increases when c decreases; if a and c are fixed, $\Pr[A|O]$ increases when b increases; if b and c are fixed, $\Pr[A|O]$ increases when a increases. □

Corollary 3.9. $\Pr[A|\overline{O}] \propto c$, $\Pr[A|\overline{O}] \propto 1/b$, and $\Pr[A|\overline{O}] \propto a$.

Proof. Straightforward. It is due to $\Pr[A|\overline{O}] = a \times (1-b)/(a \times (1-b) + (1-a) \times (1-c))$. That is, if a and b are fixed, $\Pr[A|\overline{O}]$ increases when c increases; if a and c are fixed, $\Pr[A|\overline{O}]$ decreases when b increases; if b and c are fixed, $\Pr[A|\overline{O}]$ increases when a increases. □

Theorem 3.10. Given $\Pr[A] = a$, $\Pr[O|A] = b$, and $\Pr[\overline{O}|\overline{A}] = c'$, we have $\Pr[A|O] = a \times b/(a \times b + (1-a) \times (1-c'))$ and $\Pr[A|\overline{O}] = a \times (1-b)/(a \times (1-b) + (1-a) \times c')$.

Proof. Straightforward. $\Pr[A|O] = \Pr[A,O]/\Pr[O] = \Pr[A] \times \Pr[O|A]/(\Pr[A] \times \Pr[O|A] + \Pr[\overline{A}] \times \Pr[O|\overline{A}]) = a \times b/(a \times b + (1-a) \times (1-c'))$. $\Pr[A|\overline{O}] = \Pr[A,\overline{O}]/\Pr[\overline{O}] = \Pr[A] \times \Pr[\overline{O}|A]/(\Pr[A] \times \Pr[\overline{O}|A] + \Pr[\overline{A}] \times \Pr[\overline{O}|\overline{A}]) = a \times (1-b)/(a \times (1-b) + (1-a) \times c')$. □

Corollary 3.11. $\Pr[A|O] \propto c'$, $\Pr[A|O] \propto b$, and $\Pr[A|O] \propto a$.

Proof. Straightforward. It is due to $\Pr[A|O] = a \times b/(a \times b + (1-a) \times (1-c'))$. That is, if a and b are fixed, $\Pr[A|O]$ increases when c' increases; if a and c' are fixed, $\Pr[A|O]$ increases when b increases; if b and c' are fixed, $\Pr[A|O]$ increases when a increases. □

Corollary 3.12. $\Pr[A|\overline{O}] \propto 1/c'$, $\Pr[A|\overline{O}] \propto 1/b$, and $\Pr[A|\overline{O}] \propto a$.

Proof. Straightforward. It is due to $\Pr[A|\overline{O}] = a \times (1-b)/(a \times (1-b) + (1-a) \times c')$. That is, if a and b are fixed, $\Pr[A|\overline{O}]$ decreases when c' increases; if a and c' are fixed, $\Pr[A|\overline{O}]$ decreases when b increases; if b and c' are fixed, $\Pr[A|\overline{O}]$ increases when a increases. □

3.4. Multiple Hypothesis

In this section, we explore situations involving multiple hypotheses. The subtlety lies in the relations between hypotheses. If any two hypothesis events are mutually exclusive, then the computation will be easy. Otherwise, the computation MUST consider the intersection between hypotheses, roughly speaking, the "overlapping" area between them. Furthermore, if hypothesis events are not all included in observations, then the computation MUST consider the intersection between hypotheses and observations. Indeed, these statements are not emphasized in many books (e.g., [**?**]).

Lemma 3.13. $\Pr[A, O] + \Pr[B, O] = \Pr[A \cup B, O] + \Pr[A, B, O]$.

Proof. $\Pr[A \cup B, O] = \Pr[(A \cup B) \cap O] = \Pr[(A \cap O) \cup (B \cap O)] = \Pr[A \cap O] + \Pr[B \cap O] - \Pr[(A \cap O) \cap (B \cap O)] = \Pr[A, O] + \Pr[B, O] - \Pr[A, B, O]$. (Recall that $\Pr[A \cap B]$ is simply denoted as $\Pr[A, B]$.) □

Theorem 3.14. Suppose A_1 and A_2 are two hypothesis events, and O is an observation event. Then $\Pr[A_1|O] + \Pr[A_2|O] = \Pr[(A_1 \cup A_2)|O] + \Pr[(A_1 \cap A_2)|O]$.

Proof. $\Pr[A_1|O] + \Pr[A_2|O] = \Pr[A_1, O]/\Pr[O] + \Pr[A_2, O]/\Pr[O] = (\Pr[A_1, O] + \Pr[A_2, O])/\Pr[O] = (\Pr[A_1 \cup A_2, O] + \Pr[A_1 \cap A_2, O])/\Pr[O] = \Pr[(A_1 \cup A_2)|O] + \Pr[(A_1 \cap A_2)|O]$. □

Corollary 3.15. Suppose A_1 and A_2 are two hypothesis events, and O is an observation event. If $\Pr[A_1 \cup A_2] = 1$, then $\Pr[A_1|O] + \Pr[A_2|O] = 1 + \Pr[(A_1 \cap A_2)|O]$.

Proof. If $\Pr[A_1 \cup A_2] = 1$, then $\Pr[(A_1 \cup A_2)|O] = \Pr[\Omega|O]/\Pr[O] = \Pr[O]/\Pr[O] = 1$. Thus, $\Pr[A_1|O] + \Pr[A_2|O] = 1 + \Pr[(A_1 \cap A_2)|O]$. □

Corollary 3.16. Suppose A_1 and A_2 are two hypothesis events, and O is an observation event. If $\Pr[A_1 \cap A_2] = 0$, then $\Pr[A_1|O] + \Pr[A_2|O] = \Pr[(A_1 \cup A_2)|O]$.

Proof. If $\Pr[A_1 \cap A_2] = 0$, then $\Pr[(A_1 \cap A_2)|O] = \Pr[A_1 \cap A_2 \cap O]/\Pr[O] = 0$. Thus, $\Pr[A_1|O] + \Pr[A_2|O] = \Pr[(A_1 \cup A_2)|O]$. □

Corollary 3.17. Suppose A_1 and A_2 are two hypothesis events, and O is an observation event. If $\Pr[A_1 \cup A_2] = 1$ and $\Pr[A_1 \cap A_2] = 0$, then $\Pr[A_1|O] + \Pr[A_2|O] = 1$.

Proof. $\Pr[A_1|O] + \Pr[A_2|O] = 1 + \Pr[(A_1 \cap A_2)|O] = 1 + 0 = 1$. □

Theorem 3.18. If A_1, A_2, O are independent, then
(1) $\Pr[A_1, A_2|O] = \Pr[A_1, A_2]$;
(2) $\Pr[A_1, A_2|O] = \Pr[A_1|O] \times \Pr[A_2|O]$.

Proof.
(1) $\Pr[A_1, A_2|O] = \Pr[A_1, A_2, O]/\Pr[O] = \Pr[A_1] \times \Pr[A_2] \times \Pr[O]/\Pr[O] = \Pr[A_1] \times \Pr[A_2] = \Pr[A_1, A_2]$.
(2) $\Pr[A_1] \times \Pr[A_2] = (\Pr[A_1] \times \Pr[O]/\Pr[O]) \times (\Pr[A_2] \times \Pr[O]/\Pr[O]) = (\Pr[A_1, O]/\Pr[O]) \times (\Pr[A_2, O]/\Pr[O]) = \Pr[A_1|O] \times \Pr[A_2|O]$. □

Corollary 3.19. $\Pr[A_1, A_2|O] = \Pr[A_1|O] \times \Pr[A_2|O]$ if and only if $\Pr[A_2|A_1, O] = \Pr[A_2|O]$ and $\Pr[A_1|A_2, O] = \Pr[A_1|O]$.

Proof. $\Pr[A_1, A_2, O]/\Pr[O] = \Pr[A_1, A_2|O] = \Pr[A_1|O] \times \Pr[A_2|O] = \Pr[A_1, O]/\Pr[O] \times \Pr[A_2, O]/\Pr[O] \Leftrightarrow \Pr[A_1, A_2, O] = \Pr[A_1, O] \times \Pr[A_2, O]/\Pr[O] \Leftrightarrow \Pr[A_1, A_2, O]/\Pr[A_1, O] = \Pr[A_2, O]/\Pr[O] \Leftrightarrow \Pr[A_2|A_1, O] = \Pr[A_2|O] \Leftrightarrow \Pr[A_1, A_2, O]/\Pr[A_2, O] = \Pr[A_1, O]/\Pr[O] \Leftrightarrow \Pr[A_1|A_2, O] = \Pr[A_1|O]$. □

Remark 3.20.
(1) If A_1 and A_2 are mutually exclusive, then A_1, A_2 can be looked at as a single combinational event and an observation may change its probability, due to Corollary 3.16.
(2) Theorem 3.18 is the fundamental theory for the simple version of Naive Bayes Classification (see Remark 4.13).
(3) Corollary 3.19 means that when O occurs and no matter whether A_1 occurs or not, the probability of A_2 occurring is the same. Also, it means that when O occurs and no matter whether A_2 occurs or not, the probability of A_1 occurring is the same. A and B are thus independent over the observation space O.

Next, we explore a more general case in terms of two hypotheses and one observation.

Theorem 3.21. $\Pr[A_1, A_2|O] = \frac{\Pr[A_1 \cup A_2 \cup O] - (\Pr[A_1] + \Pr[A_2] + \Pr[O] - \Pr[A_1, A_2] - \Pr[A_1, O] - \Pr[A_2, O])}{\Pr[O]}$.

Proof. Recall that $\Pr[A_1 \cup A_2 \cup O] = (\Pr[A_1] + \Pr[A_2] + \Pr[O] - \Pr[A_1, A_2] - \Pr[A_1, O] - \Pr[A_2, O]) + \Pr[A_1, A_2, O]$.

Thus, $\Pr[A_1, A_2|O] = \Pr[A_1, A_2, O]/\Pr[O] = \frac{\Pr[A_1 \cup A_2 \cup O] - (\Pr[A_1] + \Pr[A_2] + \Pr[O] - \Pr[A_1, A_2] - \Pr[A_1, O] - \Pr[A_2, O])}{\Pr[O]}$. □

Corollary 3.22. Suppose $\Pr[A_1 \cup A_2 \cup O] = 1$. If A_1, O and A_2, O are both independent, but A_1, A_2, O are not independent, then $\Pr[A_1, A_2|O] = \frac{(\Pr[O]-1)(1-\Pr[A_1]-\Pr[A_2])+\Pr[A_1, A_2]}{\Pr[O]}$.

Proof. $\Pr[A_1, A_2|O] = \frac{\Pr[A_1 \cup A_2 \cup O] - (\Pr[A_1] + \Pr[A_2] + \Pr[O] - \Pr[A_1, A_2] - \Pr[A_1, O] - \Pr[A_2, O])}{\Pr[O]} = \frac{1 - (\Pr[A_1] + \Pr[A_2] + \Pr[O] - \Pr[A_1, A_2] - \Pr[A_1] \times \Pr[O] - \Pr[A_2] \times \Pr[O])}{\Pr[O]} = \frac{1 - \Pr[A_1] - \Pr[A_2] + \Pr[A_1, A_2]}{\Pr[O]} + \Pr[A_1] + \Pr[A_2] - 1 = \frac{(\Pr[O]-1)(1-\Pr[A_1]-\Pr[A_2])+\Pr[A_1, A_2]}{\Pr[O]}$. □

Corollary 3.23. Suppose $\Pr[A_1 \cup A_2 \cup O] = \alpha$. If A_1, O and A_2, O are both independent, but A_1, A_2, O are not independent, then $\Pr[A_1, A_2|O] = \frac{\alpha - \Pr[A_1] - \Pr[A_2] + \Pr[A_1, A_2]}{\Pr[O]} + \Pr[A_1] + \Pr[A_2] - 1$.

Proof. $\Pr[A_1, A_2|O] = \frac{\Pr[A_1 \cup A_2 \cup O] - (\Pr[A_1] + \Pr[A_2] + \Pr[O] - \Pr[A_1, A_2] - \Pr[A_1, O] - \Pr[A_2, O])}{\Pr[O]} = \frac{\alpha - (\Pr[A_1] + \Pr[A_2] + \Pr[O] - \Pr[A_1, A_2] - \Pr[A_1] \times \Pr[O] - \Pr[A_2] \times \Pr[O])}{\Pr[O]} = \frac{\alpha - \Pr[A_1] - \Pr[A_2] + \Pr[A_1, A_2]}{\Pr[O]} + \Pr[A_1] + \Pr[A_2] - 1$. □

Next, we discuss another general case as preparation.

Lemma 3.24. $\Pr[A, O] + \Pr[B, O] + \Pr[C, O] - \Pr[A, B, O] - \Pr[A, C, O] - \Pr[B, C, O] + \Pr[A, B, C, O] = \Pr[A \cup B \cup C, O]$.

Proof. $\Pr[A \cup B \cup C, O] = \Pr[(A \cup B \cup C) \cap O] = \Pr[(A \cap O) \cup (B \cap O) \cup (C \cap O)] = \Pr[A, O] + \Pr[B, O] + \Pr[C, O] - \Pr[A, B, O] - \Pr[A, C, O] - \Pr[B, C, O] + \Pr[A, B, C, O]$. □

Next, we discuss the most general case.

Lemma 3.25. $\sum_{i=1}^{n} \Pr[A_i, O] - \sum_{1\le i<j\le n} \Pr[A_i, A_j, O] + \sum_{1\le i<j<k\le n} \Pr[A_i, A_j, A_k, O] + \cdots + (-1)^{n+1} \Pr[A_1, A_2, ..., A_n, O] = \Pr[\bigcup_{i=1}^{n} A_i, O]$.

Proof. Recall Eq. 1: $\sum_{i=1}^{n} \Pr[A_i] - \sum_{1\le i<j\le n} \Pr[A_i, A_j] + \sum_{1\le i<j<k\le n} \Pr[A_i, A_j, A_k] + \cdots + (-1)^{n+1} \Pr[A_1, A_2, ..., A_n] = \Pr[\bigcup_{i=1}^{n} A_i]$. Thus, we have $\Pr[\bigcup_{i=1}^{n} A_i, O] = \Pr[(\bigcup_{i=1}^{n} A_i) \cap O] = \Pr[\bigcup_{i=1}^{n} (A_i \cap O)] = \sum_{i=1}^{n} \Pr[A_i, O] - \sum_{1\le i<j\le n} \Pr[A_i, A_j, O] + \sum_{1\le i<j<k\le n} \Pr[A_i, A_j, A_k, O] + \cdots + (-1)^{n+1} \Pr[A_1, A_2, ..., A_n, O]$. □

Theorem 3.26. Suppose A_i $(i = 1, ..., n)$ are n events. Then $\sum_{i=1}^{n} \Pr[A_i|O] - \sum_{1\le i<j\le n} \Pr[A_i, A_j|O] + \sum_{1\le i<j<k\le n} \Pr[A_i, A_j, A_k|O] + \cdots + (-1)^{n+1} \Pr[A_1, A_2, ..., A_n|O] = \Pr[\bigcup_{i=1}^{n} A_i|O]$.

Proof. Due to Lemma 3.25, $\sum_{i=1}^{n} \Pr[A_i, O] - \sum_{1\le i<j\le n} \Pr[A_i, A_j, O] + \sum_{1\le i<j<k\le n} \Pr[A_i, A_j, A_k, O] + \cdots + (-1)^{n+1} \Pr[A_1, A_2, ..., A_n, O] = \Pr[\bigcup_{i=1}^{n} A_i, O]$. Dividing all terms by $\Pr[O]$ yields the conclusion due to $\Pr[A_i, ..., A_j, O] / \Pr[O] = \Pr[A_i, ..., A_j|O]$ $(i, j \in [1, n],\ i \le j)$, and $\Pr[\bigcup_{i=1}^{n} A_i, O] / \Pr[O] = \Pr[\bigcup_{i=1}^{n} A_i|O]$. □

Corollary 3.27. Suppose A_i $(i = 1, ..., n)$ are n events, and $\bigcap_{i=1}^{n} A_i = \emptyset$. Then $\sum_{i=1}^{n} \Pr[A_i|O] = \Pr[\bigcup_{i=1}^{n} A_i|O]$.

Proof. $\Pr[A_i, ..., A_j|O] = \Pr[A_i, ..., A_j, O] / \Pr[O] = 0$ $(\forall i, j \in [1, n])$. Thus, by Theorem 3.26, $\sum_{i=1}^{n} \Pr[A_i|O] = \Pr[\bigcup_{i=1}^{n} A_i|O]$. □

Corollary 3.28. Suppose A_i $(i = 1, ..., n)$ are n events, and $\bigcup_{i=1}^{n} A_i = \Omega$. Then $\sum_{i=1}^{n} \Pr[A_i|O] = 1$.

Proof. By Corollary 3.27, $\sum_{i=1}^{n} \Pr[A_i|O] = \Pr[\bigcup_{i=1}^{n} A_i|O] = \Pr[\Omega|O] = \Pr[\Omega, O] / \Pr[O] = \Pr[O] / \Pr[O] = 1$. □

Lemma 3.29. Given any function $f(A_1, ..., A_n)$ taking as input $A_1, A_2, ..., A_n$ with operators in terms of $\cap$, $\cup$, and $\div$, if $f(A_1, ..., A_n) \cap O = O$, then $\Pr[f(A_1, ..., A_n)|O] = 1$.

Proof. Straightforward. $\Pr[f(A_1, ..., A_n)|O] = \Pr[f(A_1, ..., A_n), O] / \Pr[O] = \Pr[O] / \Pr[O] = 1$. □

Lemma 3.30. Let $S = \{A_1, A_2, ..., A_n\}$. Given any function taking as input $A_{i1}, ..., A_{ip} \in S$, $A_{j1}, ..., A_{jq} \in S$, with operators in terms of $\cap$, $\cup$, $\div$, which is denoted as $f_1(A_{i1}, A_{i2}, ..., A_{ip})$, and $f_2(A_{i1}, A_{i2}, ..., A_{iq})$. Suppose $f_1(A_{i1}, A_{i2}, ..., A_{ip}) \subseteq f_2(A_{i1}, A_{i2}, ..., A_{iq})$. If $f_2(A_{i1}, A_{i2}, ..., A_{iq}) \subseteq O$, then $\Pr[f_1(A_{i1}, A_{i2}, ..., A_{ip})|O] \le \Pr[f_2(A_{i1}, A_{i2}, ..., A_{iq})|O]$.

Proof. Straightforward. $f_1(A_{i1}, A_{i2}, ..., A_{ip}) \subseteq f_2(A_{i1}, A_{i2}, ..., A_{iq})$ and $f_2(A_{i1}, A_{i2}, ..., A_{iq}) \subseteq O \Leftrightarrow f_1(A_{i1}, A_{i2}, ..., A_{ip}) \cap O = f_1(A_{i1}, A_{i2}, ..., A_{ip}) \subseteq f_2(A_{i1}, A_{i2}, ..., A_{iq}) \cap O = f_2(A_{i1}, A_{i2}, ..., A_{iq}) \Leftrightarrow \Pr[f_1(A_{i1}, A_{i2}, ..., A_{ip}), O] \le \Pr[f_2(A_{i1}, A_{i2}, ..., A_{iq}), O] \Leftrightarrow \Pr[f_1(A_{i1}, A_{i2}, ..., A_{ip})|O] \le \Pr[f_2(A_{i1}, A_{i2}, ..., A_{iq})|O]$. □

Theorem 3.31. Given any function $f_1(\cdot)$ that takes as input $A_1, A_2, ..., A_n$, $A_i \subseteq \Omega$ $(i = 1, 2, ..., n)$ with operators in terms of $\cap$, $\cup$, and $\bar{\cdot}$, and any function $f_2(\cdot)$ that takes as input $B_1, B_2, ..., B_m$, $B_j \subseteq \Omega$ $(j = 1, 2, ..., m)$ with operators in terms of $\cap$, $\cup$, and $\bar{\cdot}$. If $f_1(\cdot) \subseteq f_2(\cdot) \subseteq O \subseteq \Omega$, then $\Pr[f_1(\cdot)|O] \leq \Pr[f_2(\cdot)|O]$.

Proof. Straightforward. Since $f_1(\cdot) \subseteq f_2(\cdot) \subseteq O$, $\Pr[f_1(\cdot)|O] = \Pr[f_1(\cdot), O]/\Pr[O] = \Pr[f_1(\cdot)]/\Pr[O] \leq \Pr[f_2(\cdot)]/\Pr[O] = \Pr[f_2(\cdot), O]/\Pr[O] = \Pr[f_2(\cdot)|O]$. □

Corollary 3.32. Any equality or inequality function taking as input probabilities with operators "+" or "−" (e.g., $\Pr[A] + \Pr[B] \leq \Pr[C]$) still holds in the conditional probability space (i.e., O) (e.g., $\Pr[A|O] + \Pr[B|O] \leq \Pr[C|O]$), if all event spaces are subsets of or equal to the conditional event space (e.g., $A, B, C \subseteq O$).

Proof. Straightforward. All event spaces are subsets of or equal to the conditional event space, thus all probabilities (e.g., $\Pr[A]$) over the original space divided by $\Pr[O]$ equal the conditional probabilities over the conditional event space $\Pr[A|O]$, as $\Pr[A]/\Pr[O] = \Pr[A, O]/\Pr[O] = \Pr[A|O]$. □

Corollary 3.33. Any formulas consisting of plus, minus, and probabilities (in the original space) are still guaranteed in the conditional probability space (i.e., replacing the probabilities with the conditional probabilities), if all event spaces are subsets of or equal to the conditional event space.

Proof. The reason is that the intersection (or joint) of all event spaces and the conditional event space is still the event spaces, so all probabilities in the formula divided by the probability of the conditional event space result in the corresponding conditional probabilities. □

3.5. Bayes Theorem

As a basic version, we suppose there exist only two events (i.e., A and B) in the observation domain.

Note that $\Pr[O] = \Pr[A, O] + \Pr[B, O]$ if and only if $\Pr[A \cup B] = 1$ and $\Pr[A \cap B] = 0$. Strictly speaking, $\Pr[O] = \Pr[A, O] + \Pr[B, O]$ if and only if $\Pr[O \cap (A \cup B)] = 0$ and $\Pr[O \cap (A \cap B)] = 0$. In other words, the intersection of O with A and with B constitutes the entire O.

$\Pr[A|O] = \frac{\Pr[O|A] \times \Pr[A]}{\Pr[O|A] \times \Pr[A] + \Pr[O|B] \times \Pr[B]}$.

Geometric Representation. Fig. 16 [Figure 16: see original paper] shows Bayes theorem in a geometric way.

For example, $\Pr[A_1|O] = \Pr[O, A_1]/\Pr[O] = \Pr[O, A_1]/(\Pr[O, A_1] + \Pr[O, A_2] + \Pr[O, A_3]) = \Pr[O, A_1]/(\Pr[O|A_1] \times \Pr[A_1] + \Pr[O|A_2] \times \Pr[A_2] + \Pr[O|A_3] \times \Pr[A_3])$.

$\Pr[O, A_1] = S_{O_1 B_2 B_5 O_4}$, $\Pr[A_1] = S_{O_1 B_2 B_3 B_4}$, $\Pr[O|A_1] = S_{O_1 B_2 B_5 O_4}/S_{O_1 B_2 B_3 B_4}$.

It can also present quantitative probabilities by areas, e.g., $\Pr[O, A_2] > \Pr[O] \times \Pr[A_2]$, $\Pr[O, A_3] < \Pr[O] \times \Pr[A_3]$.

[Figure 16: see original paper]

Remark 3.34.
(1) When $\Pr[A|O]$ is not easy to compute, we can compute it from $\Pr[O|A]$ (together with $\Pr[A]$ and $\Pr[O]$). Besides, when $\Pr[O]$ is not easy to compute, we can compute it from $\Pr[O|A]$ (together with $\Pr[A]$) and $\Pr[O|B]$ (together with $\Pr[B]$).
(2) Recall that $\Pr[A|O]$ is the so-called posterior probability, which can be computed from the so-called prior probabilities such as $\Pr[A]$.

The next proposition states a general case.

Proposition 3.35. If $\Pr[\bigcup_{i=1}^{n} A_i] = 1$ and $\forall i, j \in [1, n]$, $\Pr[A_i \cap A_j] = 0$ (i.e., $A_1, ..., A_n$ are mutually exclusive), then $\Pr[A_i|O] = \frac{\Pr[O|A_i] \times \Pr[A_i]}{\sum_{j=1}^{n} \Pr[O|A_j] \times \Pr[A_j]}$.

Proof. Straightforward. $\Pr[\bigcup_{i=1}^{n} A_i] = 1$, thus $\Pr[(\bigcup_{i=1}^{n} A_i) \cap O] = \Pr[\bigcup_{i=1}^{n} (A_i \cap O)] = \Pr[O]$. $\forall i, j \in [1, n]$, $\Pr[A_i \cap A_j] = 0$, thus $\Pr[\bigcup_{i=1}^{n} (A_i \cap O)] = \sum_{i=1}^{n} \Pr[A_i, O]$. Therefore, $\Pr[A_i|O] = \Pr[A_i, O] / \Pr[O] = \Pr[O|A_i] \times \Pr[A_i] / \sum_{j=1}^{n} \Pr[O|A_j] \times \Pr[A_j]$. □

It is worth noting that many textbooks on Bayes theorem do not explicitly emphasize the condition $\Pr[\bigcup_{i=1}^{n} A_i] = 1$ and $\forall i, j \in [1, n]$, $\Pr[A_i \cap A_j] = 0$. They always assume $\Pr[O] = \Pr[\bigcup_{i=1}^{n} (A_i \cap O)]$ by default, which is NOT proper in some cases.

Next, we emphasize other cases for computing $\Pr[O]$ from the conditional probabilities.

Theorem 3.36. If $\Pr[A \cup B] = 1$ and $\Pr[A \cap B] \neq 0$, then $\Pr[A|O] = \frac{\Pr[O|A] \times \Pr[A]}{\Pr[O|A] \times \Pr[A] + \Pr[O|B] \times \Pr[B] - \Pr[O|A \cap B] \times \Pr[A \cap B]}$.

Proof. If $\Pr[A \cup B] = 1$ and $\Pr[A \cap B] \neq 0$, then $\Pr[O] = \Pr[(A \cup B) \cap O] = \Pr[A, O] + \Pr[B, O] - \Pr[A, B, O]$, which is due to Lemma 3.13. Thus, $\Pr[O] = \Pr[O|A] \times \Pr[A] + \Pr[O|B] \times \Pr[B] - \Pr[O|A \cap B] \times \Pr[A \cap B]$.

Therefore, $\Pr[A|O] = \Pr[A, O] / \Pr[O] = \frac{\Pr[O|A] \times \Pr[A]}{\Pr[O|A] \times \Pr[A] + \Pr[O|B] \times \Pr[B] - \Pr[O|A \cap B] \times \Pr[A \cap B]}$.
□

The conditions in Theorem 3.36 can be extended, e.g., extending two to more events, changing $\Pr[A \cap B] \neq 0$ to A_i $(i = 1, 2, ..., n)$ being mutually exclusive, and loosening the condition $O \subseteq \bigcup_{i=1}^{n} A_i$ to $\bigcup_{i=1}^{n} A_i \subseteq O$. Thus, the theorem for more general conditions is obtained as follows.

Theorem 3.37. If $\forall i, j \in [1, n]$, $\Pr[A_i \cap A_j] = 0$, then $\Pr[A_i|O] = \frac{\Pr[O|A_i] \times \Pr[A_i]}{\sum_{j=1}^{n} \Pr[O|A_j] \times \Pr[A_j] + \Pr[O|\bigcup_{i=1}^{n} A_i] \times \Pr[\bigcup_{i=1}^{n} A_i]}$.

Proof. The key point is to compute $\Pr[O]$. $\Pr[O] = \Pr[(\bigcup_{i=1}^{n} A_i) \cap O] = \Pr[\bigcup_{i=1}^{n}(A_i \cap O)] = \sum_{i=1}^{n} \Pr[A_i, O]$, which is due to Lemma 3.13. Together with $\Pr[A_i \cap A_j] = 0$ for $i, j \in [1, n]$, we have $\Pr[\bigcup_{i=1}^{n}(A_i \cap O)] = \sum_{i=1}^{n} \Pr[A_i, O] = \sum_{i=1}^{n} \Pr[O|A_i] \times \Pr[A_i]$. □

Geometric Representation. Fig. 17 [Figure 17: see original paper] depicts how to compute O by $A \cap O$ and $B \cap O$ (and others, if applicable), which is critical in Bayes theorem in a geometric way. The relations between A, B, and O.

[Figure 17: see original paper]

Theorem 3.37 is a general version of Bayes theorem and is visualized in Fig. 17 (the middle and rightmost panels).

Indeed, the last item in the proof of Theorem 3.37 can be written as follows: $\Pr[O|\bigcup_{i=1}^{n} A_i] \times \Pr[\bigcup_{i=1}^{n} A_i] = \Pr[\bigcap_{i=1}^{n} A_i, O] = \Pr[A_1, A_2, ..., A_n, O]$.

On the basis of Theorem 3.37, we explore the most general case—$\bigcup_{i=1}^{n} A_i \subseteq O$, and A_i $(i = 1, 2, ..., n)$ are NOT mutually exclusive. The key point hereby is how to compute $\Pr[O]$ from $\Pr[O, A_i]$ that have overlaps.

Theorem 3.38. $\forall i \in [1, n]$, $\Pr[A_i|O] = \frac{\Pr[O|A_i] \times \Pr[A_i]}{\lambda}$, where

$$\lambda = \sum_{i=1}^{n} \Pr[O|A_i] \times \Pr[A_i] - \sum_{1 \le i < j \le n} \Pr[O|A_i, A_j] \times \Pr[A_i, A_j] + \cdots + (-1)^{n+1} \Pr[O|A_1, A_2, ..., A_n] \times \Pr[A_1, A_2, ..., A_n]$$

Proof. We only need to prove the computation of $\Pr[\bigcup_{i=1}^{n}(A_i \cap O)]$ is the same as Theorem 3.37. Recall Eq. 1, we have

$$\Pr[\bigcup_{i=1}^{n}(A_i \cap O)] = \sum_{i=1}^{n} \Pr[A_i \cap O] - \sum_{1 \le i < j \le n} \Pr[A_i \cap O, A_j \cap O] + \sum_{1 \le i < j < k \le n} \Pr[A_i \cap O, A_j \cap O, A_k \cap O] + \cdots + (-1)^{n+1} \Pr[A$$

□

Remark 3.39.
(1) If A_i $(i = 1, ..., n)$ are not independent, joint probabilities, e.g., $\Pr[A_i, A_j]$, cannot be computed from $\Pr[A_i]$ and $\Pr[A_j]$ directly as $\Pr[A_i, A_j] \neq \Pr[A_i] \times \Pr[A_j]$.
(2) We can also write λ as follows: $\lambda = \sum_{i=1}^{n} \Pr[O|A_i] \times \Pr[A_i] - \Pr[A_1, A_2, O] - \Pr[A_1, A_3, O] - \cdots - \Pr[A_{n-1}, A_n, O] + \cdots + (-1)^{n+1} \Pr[A_1, A_2, ..., A_n, O] + \Pr[A_1, ..., A_n, O]$. It is easy to understand that λ includes entire yet non-duplicated contributions for $\Pr[O]$.
(3) Let $\Delta_1 = -\Pr[A_1, A_2, O] - \cdots - \Pr[A_{n-1}, A_n, O] + \cdots + (-1)^{n+1} \Pr[A_1, A_2, ..., A_n, O] +$

$\Pr[A_1, A_2, ..., A_n, O]$. If $\Delta_1 \ll \sum_{i=1}^{n} \Pr[O|A_i] \times \Pr[A_i]$, then $\lambda \approx \sum_{i=1}^{n} \Pr[O|A_i] \times \Pr[A_i]$. The equation degenerates to the normal form (see Proposition 3.35).
(4) If A_i are mutually exclusive, then $\Delta_1 = \Pr[A_1, A_2, ..., A_n, O]$. $\lambda = \sum_{i=1}^{n} \Pr[O|A_i] \times \Pr[A_i] + \Pr[A_1, A_2, ..., A_n, O]$. The equation degenerates to the form in Theorem 3.37.
(5) Let $\Delta_2 = \sum_{1 \le i < j \le n} \Pr[A_i, A_j, O] + \sum_{1 \le i < j < k \le n} \Pr[A_i, A_j, A_k, O] + \cdots + \Pr[A_1, A_2, ..., A_n, O] + \Pr[A_1, ..., A_n, O]$. If $\Delta_2 \ll \sum_{i=1}^{n} \Pr[O|A_i] \times \Pr[A_i]$, then $\lambda \approx \sum_{i=1}^{n} \Pr[O|A_i] \times \Pr[A_i]$ because $\Delta_1 < \Delta_2$. The equation degenerates to the normal form (see Proposition 3.35).

3.6. Posterior Probability or the Impact of Observation to Hypothesis

Suppose there exist only two events in the observation domain as a basic version. That is, $\Pr[A] + \Pr[\overline{A}] = 1$. Hereby, we want to explore when $\Pr[A] < \Pr[A|O]$ or $\Pr[A] > \Pr[A|O]$.

Remark 3.40.
(1) After observing event O, the estimated probability of A has changed from x to a new value $\Pr[A|O]$ (the so-called posterior probability).
(2) If A and O are independent, then $\Pr[A|O] = \Pr[A, O]/\Pr[O] = \Pr[A] \times \Pr[O]/\Pr[O] = \Pr[A]$. Thus, the observation of O does not change the probability estimation of event A.
(3) If $\Pr[A|O] = 0$, then $\Pr[A, O] = 0$, which means A and O are mutually exclusive. The observation of O implies the impossibility of event A.

As $\Pr[A|O] = \frac{\Pr[O|A] \times \Pr[A]}{\Pr[O]} = \frac{\Pr[O|A]}{\Pr[O]} \times \Pr[A]$, whether $\frac{\Pr[O|A]}{\Pr[O]} > 1$ or not is critical. $\Pr[A|O] > \Pr[A] \Leftrightarrow \frac{\Pr[O|A]}{\Pr[O]} > 1 \Leftrightarrow \Pr[O|A] > \Pr[O]$, as discussed in Fig. 12.

Next, we will explore when a new observation may increase (or decrease) the probability of an old hypothesis (e.g., $\Pr[A|O] > \Pr[A]$).

Recall Theorem 3.7 and Theorem 3.10 (and Fig. 15). Let $\Pr[A] = a \in (0, 1]$, $\Pr[O|A] = b \in (0, 1]$, and $\Pr[O|\overline{A}] = c \in (0, 1]$ (or $\Pr[\overline{O}|\overline{A}] = c' \in (0, 1]$). Suppose a, b, c are given, then $\Pr[A|O] = a \times b/(a \times b + (1 - a) \times c)$ and $\Pr[A|\overline{O}] = a \times (1 - b)/(a \times (1 - b) + (1 - a) \times (1 - c))$.

Suppose a, b, c' are given, then $\Pr[A|O] = a \times b/(a \times b + (1 - a) \times (1 - c'))$ and $\Pr[A|\overline{O}] = a \times (1 - b)/(a \times (1 - b) + (1 - a) \times c')$.

Theorem 3.41. Suppose $\Pr[A] = a \in (0, 1]$, $\Pr[O|A] = b \in (0, 1]$, and $\Pr[O|\overline{A}] = c \in (0, 1]$. Then $\Pr[A|O] > \Pr[A]$ if and only if $b > c$.

Proof. $\Pr[A|O] > \Pr[A] \Leftrightarrow a \times b/(a \times b + (1 - a) \times c) > a \Leftrightarrow b/(b + (1/a - 1) \times c) > a \Leftrightarrow b > a \times b + (1 - a) \times c \Leftrightarrow b \times (1 - a) > (1 - a) \times c \Leftrightarrow b > c$. □

Geometric Representation. The geometric representation for Theorem 3.41 is shown in Fig. 18 [Figure 18: see original paper].

Theorem 3.42. Suppose $\Pr[A] = a \in (0,1]$, $\Pr[O|A] = b \in (0,1]$, and $\Pr[O|\overline{A}] = c \in (0,1]$. Then $\Pr[A|\overline{O}] > \Pr[A]$ if and only if $b < c$.

Proof. $\Pr[A|\overline{O}] > \Pr[A] \Leftrightarrow a \times (1-b)/(a \times (1-b) + (1-a) \times (1-c)) > a \Leftrightarrow (1-b)/((1-b)+(1/a-1) \times (1-c)) > a \Leftrightarrow 1-b > (1-b) \times a + (1-a) \times (1-c) \Leftrightarrow (1-b) \times (1-a) > (1-a) \times (1-c) \Leftrightarrow 1-b > 1-c \Leftrightarrow b < c$. □

Geometric Representation. The geometric representation for Theorem 3.42 is shown in Fig. 19 [Figure 19: see original paper].

Theorem 3.43. Suppose $\Pr[A] = a \in (0,1]$, $\Pr[O|A] = b \in (0,1]$, and $\Pr[\overline{O}|\overline{A}] = c' \in (0,1]$. Then $\Pr[A|O] > \Pr[A]$ if and only if $b + c' > 1$.

Proof. $\Pr[A|O] > \Pr[A] \Leftrightarrow a \times b/(a \times b + (1-a) \times (1-c')) > a \Leftrightarrow b > a \times b + (1-a) \times (1-c') \Leftrightarrow b \times (1-a) > (1-a) \times (1-c') \Leftrightarrow b > 1-c' \Leftrightarrow b+c' > 1$. □

Geometric Representation. The geometric representation for Theorem 3.43 is shown in Fig. 20 [Figure 20: see original paper].

Theorem 3.44. Suppose $\Pr[A] = a \in (0,1]$, $\Pr[O|A] = b \in (0,1]$, and $\Pr[\overline{O}|\overline{A}] = c' \in (0,1]$. Then $\Pr[A|\overline{O}] > \Pr[A]$ if and only if $b + c' < 1$.

Proof. $\Pr[A|\overline{O}] > \Pr[A] \Leftrightarrow a \times (1-b)/(a \times (1-b) + (1-a) \times c') > a \Leftrightarrow 1-b > (1-b) \times a + (1-a) \times c' \Leftrightarrow (1-b)(1-a) > (1-a) \times c' \Leftrightarrow 1-b > c' \Leftrightarrow b+c' < 1$. □

Geometric Representation. The geometric representation for Theorem 3.44 is shown in Fig. 21 [Figure 21: see original paper].

4.1. Multiple Observations

Multiple observations may have multiple conditional probabilities. Suppose there exist t observations, namely, $O_1, O_2, ..., O_t$.

Generally, suppose there exist observations $O_1, O_2, ..., O_t$. For $i = 1, 2, ..., t$, we have $\Pr[A|O_1, O_2, ..., O_t] = \Pr[A, O_1, ..., O_t] / \Pr[O_1, ..., O_t]$.

Theorem 4.1. $\Pr[A|O_1, O_2, ..., O_t] = 1 - \frac{\Pr[A \cup (O_1, ..., O_t)] - \Pr[A]}{\Pr[O_1, ..., O_t]}$.

Proof. $\Pr[A, O_1, ..., O_t] = \Pr[A] + \Pr[O_1, ..., O_t] - \Pr[A \cup (O_1, ..., O_t)]$.

$\Pr[A|O_1, O_2, ..., O_t] = \Pr[A, O_1, ..., O_t] / \Pr[O_1, ..., O_t] = \frac{\Pr[A] + \Pr[O_1, ..., O_t] - \Pr[A \cup (O_1, ..., O_t)]}{\Pr[O_1, ..., O_t]} = 1 - \frac{\Pr[A \cup (O_1, ..., O_t)] - \Pr[A]}{\Pr[O_1, ..., O_t]}$. □

Remark 4.2.
(1) If observations happen sequentially, then each observation will result in $\Pr[A|O_i]$, $i = 1, 2, ..., t$, which could be larger or smaller as the observation may "support" the hypothesis or not. In this case, multiple observations degenerate to the accumulation of one observation. (The geometric graph for multiple observations could be a three-dimensional overlay of the two-dimensional graph

for one observation.)
(2) If observations happen simultaneously, we can view t observations as a single combined observation. In this case, it again degenerates to one observation. Simply speaking, let $O' = O_1 \cap O_2 \cap ... \cap O_t$, then $\Pr[A|O_1, O_2, ..., O_t] = \Pr[A, O']/\Pr[O']$.
(3) To compute $\Pr[A|O_1, O_2, ..., O_t]$ and differentiate the contributions of observations, $\Pr[A, O_1, ..., O_t]$ and $\Pr[O_1, ..., O_t]$ need to be computed. The simplest case is that all t observations are mutually independent and A is also mutually independent with the t observations. We have $\Pr[A|O_1, O_2, ..., O_t] = \Pr[A, O_1, ..., O_t]/\Pr[O_1, ..., O_t] = \Pr[A] \times \Pr[O_1] \times ... \times \Pr[O_t]/(\Pr[O_1] \times ... \times \Pr[O_t]) = \Pr[A]$.
(4) If all t observations are not mutually independent, then $\Pr[O_1, ..., O_t]$ must be computed by Eq. 1. Similarly, if A is also not mutually independent with the t observations, then $\Pr[A, O_1, ..., O_t]$ must be computed by Eq. 1.

Geometric Representation. If observations are mutually exclusive, then similar to Fig. 16, Fig. 22 [Figure 22: see original paper] shows joint probabilities after three observation events O_1, O_2, O_3 with original hypothesis (i.e., $\Pr[O_i, A]$, $\Pr[O_i, \overline{A}]$, $i = 1, 2, 3$) in a geometric way. We can see that $\Pr[A|O_1] = \Pr[A]$, $\Pr[A|O_2] > \Pr[A]$, and $\Pr[A|O_3] < \Pr[A]$.

[Figure 22: see original paper]

4.2. General Case with Multiple Hypothesis and Multiple Observations

The general case is that there exist multiple hypotheses and multiple observations, and the question is how to visualize them.

Geometric Representation. Fig. 23 [Figure 23: see original paper] shows joint probabilities after three observation events O_1, O_2, O_3 with original three hypotheses A_1, A_2, A_3 (i.e., $\Pr[O_i, A_i]$, $i = 1, 2, 3$) in a geometric way.

[Figure 23: see original paper]

Remark 4.3.
(1) If and only if $\Pr[A_i \cap (\bigcup_{j=1}^{3} O_j)] = \emptyset$, A_i $(i = 1, 2, 3)$ are mutually exclusive, and O_j $(j = 1, 2, 3)$ are mutually exclusive, then $\forall i = 1, 2, 3$, $\Pr[A_i] = \sum_{j=1}^{3} \Pr[A_i, O_j]$.
(2) If and only if $\Pr[O_j \cap (\bigcup_{i=1}^{3} A_i)] = \emptyset$, A_i $(i = 1, 2, 3)$ are mutually exclusive, and O_j $(j = 1, 2, 3)$ are mutually exclusive, then $\forall j = 1, 2, 3$, $\Pr[O_j] = \sum_{i=1}^{3} \Pr[A_i, O_j]$.
(3) $\forall i \in [1, 3]$, $j \in [1, 3]$, $\Pr[O_j|A_i] = \Pr[A_i, O_j]/\Pr[A_i]$, $\Pr[A_i|O_j] = \Pr[A_i, O_j]/\Pr[O_j]$.
(4) $\sum_{i=1}^{3} \Pr[A_i] \in [0, 1]$. $\sum_{j=1}^{3} \Pr[O_j] \in [0, 1]$.
(5) $\Pr[O_j|A_i] = \Pr[A_i, O_j]/\Pr[A_i] = \Pr[A_i, O_j]/\sum_{j=1}^{3} \Pr[A_i, O_j]$. $\Pr[A_i|O_j] =$

$\Pr[A_i, O_j]/\Pr[O_j] = \Pr[A_i, O_j]/\sum_{i=1}^{3}\Pr[A_i, O_j]$.
(6) More generally, we can extend case (1) to $\Pr[O_j] = \sum_{i=1}^{3}\Pr[A_i, O_j] + \Pr[(\bigcup_{i=1}^{3} A_i), O_j]$. Similarly, $\Pr[A_i] = \sum_{j=1}^{3}\Pr[A_i, O_j] + \Pr[A_i, (\bigcup_{j=1}^{3} O_j)]$.
(7) Although the above remarks discuss the case of A_i and O_j $(i, j = 1, 2, 3)$ in the graph, they can be easily extended to the general case where A_i and O_j $(i, j = 1, 2, ..., n)$.

Next, we discuss a general case for conditional probability, roughly speaking, like a "chain".

Proposition 4.4. $\Pr[A_1] \times \Pr[A_2|A_1] \times \Pr[A_3|A_1, A_2] \times \Pr[A_4|A_1, A_2, A_3] \times \cdots \times \Pr[A_n|A_1, A_2, ..., A_{n-1}] = \Pr[A_1, A_2, ..., A_n]$.

Proof. Straightforward. It is due to $\Pr[A_{i+1}|A_1, ..., A_i] \times \Pr[A_1, ..., A_i] = \Pr[A_1, ..., A_{i+1}]$. □

Theorem 4.5. Suppose $\Pr[A_1, ..., A_i] = a_i$ $(i = 1, 2, ..., n)$. Then $\Pr[A_i|A_1, A_2, ..., A_{i-1}] = a_i/a_{i-1}$ for $i = 2, 3, ..., n$.

Proof. Straightforward. For $i = 2$, $\Pr[A_2|A_1] = \Pr[A_1, A_2]/\Pr[A_1] = a_2/a_1$. Similarly, $\Pr[A_i|A_1, A_2, ..., A_{i-1}] = \Pr[A_1, A_2, ..., A_i]/\Pr[A_1, A_2, ..., A_{i-1}] = a_i/a_{i-1}$ for $i = 3, 4, ..., n$. □

Corollary 4.6. Suppose $\Pr[A_1, ..., A_i] = a_i$ $(i = 1, 2, ..., n)$. Then $a_1 \geq a_2 \geq a_3 \cdots \geq a_n$.

Proof. Straightforward. $\Pr[A_1] = a_1 \geq \Pr[A_1, A_2] = a_2$, as $A_1 \cap A_2 \subseteq A_1$. Similarly, $\Pr[A_1, A_2, ..., A_{i-1}] = a_{i-1} \geq \Pr[A_1, A_2, ..., A_i] = a_i$ for $i = 2, 3, ..., n$. □

Corollary 4.7. If $a_i/a_{i-1} = \alpha$, then $a_i = \alpha^{i-1} \times a_1$, where $\Pr[A_1, ..., A_i] = a_i$ $(i = 1, 2, ..., n)$.

Proof. $\Pr[A_i|A_1, A_2, ..., A_{i-1}] = a_i/a_{i-1} = \alpha$. $a_2 = \Pr[A_1, A_2] = \Pr[A_2|A_1] \times \Pr[A_1] = a_2/a_1 \times a_1 = \alpha \times a_1$. $a_3 = \Pr[A_1, A_2, A_3] = \Pr[A_3|A_1, A_2] \times \Pr[A_1, A_2] = a_3/a_2 \times a_2 = \alpha \times a_2 = \alpha^2 \times a_1$. Iteratively, $a_i = \Pr[A_1, A_2, ..., A_i] = \Pr[A_i|A_1, A_2, ..., A_{i-1}] \times \Pr[A_1, A_2, ..., A_{i-1}] = \alpha^{i-1} \times a_1$. □

Theorem 4.8. If $\Pr[A_i|A_{i-1}] = \alpha$ $(i = 1, 2, ..., n)$, and $\Pr[A_{i-1}, \overline{A_i}] = 0$, then $\Pr[A_i] = \alpha^{i-1} \times \Pr[A_1]$.

Proof. $\Pr[A_{i-1}, A_i] = \Pr[A_i|A_{i-1}] \times \Pr[A_{i-1}] = \alpha \times \Pr[A_{i-1}]$. $\Pr[A_{i-1}, \overline{A_i}] = 0$, thus $\Pr[A_i] = \Pr[A_{i-1}, A_i] + \Pr[\overline{A_{i-1}}, A_i] = \alpha \times \Pr[A_{i-1}] + 0 = \alpha \times \Pr[A_{i-1}]$. Iteratively, $\Pr[A_{i-1}] = \alpha \times \Pr[A_{i-2}]$. Thus, $\Pr[A_i] = \alpha^{i-1} \times \Pr[A_1]$. □

The next proposition discusses the correlation between one observation (among multiple observations) and the hypothesis.

Proposition 4.9. $\Pr[H|O_1, O_2] > \Pr[H|O_1] \Leftrightarrow \Pr[O_2|H, O_1] > \Pr[O_2|O_1]$.

Proof. $\Pr[H|O_1, O_2] > \Pr[H|O_1] \Leftrightarrow \Pr[H, O_1, O_2]/\Pr[O_1, O_2] > \Pr[H, O_1]/\Pr[O_1] \Leftrightarrow \Pr[H, O_1, O_2]/\Pr[H, O_1] > \Pr[O_1, O_2]/\Pr[O_1] \Leftrightarrow \Pr[O_2|H, O_1] > \Pr[O_2|O_1]$.

The above shows that H and O_2 have strong correlations. □

4.3. Inference

If A is looked at as a reason and O as a result, conditional probabilities may have more explanations. Suppose A_i, $i = 1, 2, ..., n$ and O_j, $j = 1, 2, ..., m$.

$\Pr[O_j|A_i]$ is a probability from reason A_i to result O_j. Suppose $\Pr[A_i] = a_i$, $\Pr[O_j|A_i] = p_{i,j}$, $\Pr[O_j] = o_j$.

Fig. 24 [Figure 24: see original paper] shows conditional probabilities after sequential events $\Pr[A_i]$ $(i = 1, 2, ..., n)$ in a mapping way.

[Figure 24: see original paper]

Remark 4.10.
(1) Usually, $\sum_{j=1}^{m} \Pr[O_j|A_i] = 1$. That is, $\sum_{j=1}^{m} p_{i,j} = 1$. The reason is from the definition of conditional probability.
(2) Usually, $\Pr[\bigcup_{i=1}^{n} A_i] = 1$. Otherwise, $\Pr[\bigcup_{i=1}^{n} A_i] \neq 0$. Hereby $\Pr[\bigcup_{i=1}^{n} A_i]$ can be added as the probability of an extra reason (i.e., $A_{n+1} = \bigcup_{i=1}^{n} A_i$).
(3) Usually, $\forall i, j \in [1, n]$, $\Pr[A_i \cap A_j] = 0$. That is, A_i $(i = 1, 2, ..., n)$ are mutually exclusive. Thus, $\Pr[O_j] = \sum_{i=1}^{n} \Pr[O_j|A_i] \times \Pr[A_i] = \sum_{i=1}^{n} p_{i,j} \times a_i$. This is the so-called total probability formula.
(4) Usually, $\sum_{j=1}^{m} O_j = 1$. Otherwise, $\Pr[\bigcup_{j=1}^{m} O_j] \neq 0$. Hereby $\Pr[\bigcup_{j=1}^{m} O_j|A_i]$ can be added as an extra result (i.e., $O_{m+1} = \bigcup_{j=1}^{m} O_j$) if the probability is available.
(5) $\Pr[A_i|O_j]$ is usually of interest if $\Pr[O_j|A_i]$ and $\Pr[A_i]$ are given. This is the main concern of Bayes probability. For example, in Fig. 24, $\Pr[A_i|O_j] = \Pr[A_i|O_j]/\Pr[O_j]$. Roughly speaking, if we view $\Pr[O_j|A_i] \times \Pr[A_i]$ as a "branch" from A_i to O_j in probability, then $\Pr[A_i|O_j]$ is the ratio of the "branch" (from A_i to O_j) over all branches (from A_i $(i = 1, 2, 3)$ to O_j).
(6) We can build a matrix $P_{n\times m}$, where the element at row i and column j is $P[i, j] = \Pr[O_j|A_i] = p_{i,j}$. Thus, $\forall i \in [1, n]$, $\sum_{j=1}^{m} p_{i,j} = \sum_{j=1}^{m} P[i, j] = \sum_{j=1}^{m} \Pr[O_j|A_i] = 1$.
(7) $A_{1\times n} \times P_{n\times m} = O_{1\times m}$, where $A_{1\times n} = [a_1, a_2, ..., a_n]$ and $O_{1\times m} = [o_1, o_2, ..., o_m]$.

The last point in Remark 4.10 is nontrivial; we thus prove it.

Theorem 4.11. Suppose $A_{1\times n} \times P_{n\times m} = O_{1\times m}$, where $A_{1\times n} = [a_1, a_2, ..., a_n]$ and $O_{1\times m} = [o_1, o_2, ..., o_m]$. If $\forall i \in [1, n]$, $\sum_{j=1}^{m} p_{i,j} = 1$, then $\sum_{i=1}^{n} a_i = 1$ and $\sum_{j=1}^{m} o_j = 1$.

Proof. $\forall j \in [1, m]$, $o_j = \sum_{i=1}^{n} a_i \times p_{i,j}$ due to $A_{1\times n} \times P_{n\times m} = O_{1\times m}$. $\sum_{j=1}^{m} o_j = \sum_{j=1}^{m} \sum_{i=1}^{n} a_i \times p_{i,j} = \sum_{i=1}^{n} (a_i \times \sum_{j=1}^{m} p_{i,j}) = \sum_{i=1}^{n} (a_i \times 1) = \sum_{i=1}^{n} a_i = 1$. □

Remark 4.12.
(1) In the Markov chain context, A and O are matrices with the same

dimension (i.e., $n = m$), and P is the so-called transitive probability matrix. If $\lim_{z\to\infty} A \times P^z = A'$, then A' will be the probabilities for the asymptotically stable status, which is the so-called stationary distribution. The computation of $\lim_{z\to\infty} P^z$ usually requires matrix eigenvalue decomposition of P. More specifically, $P = X^{-1}P'X$ where P' is a diagonal matrix with eigenvalues (e.g., $\lambda_1, \lambda_2, ..., \lambda_n$) if applicable. Thus, $P^z = X^{-1}P'^zX$ and P'^z is a diagonal matrix with eigenvalues to the power of z (e.g., $\lambda_1^z, \lambda_2^z, ..., \lambda_n^z$).
(2) Given $A_{1\times n}$ and $P_{n\times m}$, we can compute $O_{1\times m}$, and $\Pr[A_i|O_j] = \Pr[A_i, O_j]/\Pr[O_j] = \Pr[O_j|A_i] \times \Pr[A_i]/\sum_{i=1}^{n} \Pr[O_j|A_i] \times \Pr[A_i] = p_{i,j} \times a_i/o_j$. We thus give a graph to show this kind of probability change in Fig. 25 [Figure 25: see original paper].

[Figure 25: see original paper]

Finally, we discuss a typical application of Bayes theorem. n samples with t features are denoted as a vector $X_i = [x_{i,1}, x_{i,2}, ..., x_{i,t}]$ $(i = 1, 2, ..., n)$. Given X_i and its class $C_i \in \{c_1, c_2, ..., c_m\}$, a classifier can be "trained." Thus, given a new sample denoted as X', the classifier can return the most suitable $c_{j'}$, $j' \in [1, m]$, which is the so-called Naive Bayes Classification.

Remark 4.13.
(1) Naive Bayes classification is that, given $X' = [x'_1, x'_2, ..., x'_t]$, to find the most suitable class, e.g., $c_{j'}$ $(j' \in [1, m])$, such that $\Pr[c_{j'}|X']$ is maximal, which is the so-called Maximum Likelihood Estimation.
(2) Given $X_i = [x_{i,1}, x_{i,2}, ..., x_{i,t}]$ $(i = 1, 2, ..., n)$ and its class $C_i \in \{c_1, c_2, ..., c_m\}$, $\Pr[c_u]$ $(u = 1, 2, ..., m)$ and $\Pr[x'_v|c_u]$ $(v = 1, 2, ..., t)$ can be computed. More specifically, $\forall u = 1, 2, ..., m$, $\Pr[c_u] = \frac{|\{i|\exists i\in[1,n], C_i = c_u \in \{c_1, c_2, ..., c_m\}\}|}{|\{i|\exists i\in[1,n], C_i = c_u \in \{c_1, c_2, ..., c_m\}\}|}$. $\forall v = 1, 2, ..., t$, $\Pr[x'_v|c_u] = \frac{|\{i|\exists i\in[1,n], x_{i,v} = x'_v, C_i = c_u \in \{c_1, c_2, ..., c_m\}\}|}{|\{i|\exists i\in[1,n], C_i = c_u \in \{c_1, c_2, ..., c_m\}\}|}$, where $|S|$ denotes the number of elements in set S.
(3) Suppose $x'_1, x'_2, ..., x'_t$ are given and the class is c_u, $u = 1, ..., m$. Then $\Pr[x'_1, x'_2, ..., x'_t|c_u] = \Pr[x'_1|c_u] \times \Pr[x'_2|c_u] \times ... \times \Pr[x'_t|c_u]$, as x'_v, $v = 1, ..., t$ are mutually independent and independent of c_u. This is the so-called "naive" assumption, which means that all features in any vector are mutually independent. The non-naive computation should follow Theorem 3.26, Theorem 3.21, and its corollaries, which makes the computation much more complex. Recall that the naive version is proved in Theorem 3.18.
(4) In the naive Bayes classifier, find $j' \in [1, m]$ such that $f = \arg\max_{j'\in[1,m]} \Pr[c_{j'}|X'] = \Pr[c_{j'}] \times \Pr[X'|c_{j'}]/\Pr[X']$. As $\Pr[X']$ remains the same as a constant, it can be ignored. Due to the naive assumption in (3), $f = \arg\max_{j'\in[1,m]} \Pr[c_{j'}] \times \prod_{v=1}^{t} \Pr[x'_v|c_{j'}]$. (If $\Pr[x'_v|c_{j'}] = 0$, then it can be intentionally enlarged to a small value for the multiplication.)

Acknowledgement

The research was financially supported by the Key Laboratory of Computational Science and Application of Hainan Province (No. JSKX202302), the Open Topic

of Hunan Engineering Research Center of Geographic Information Security and Application (No. HNGISA2023001), the Key Laboratory of Data Protection and Intelligent Management, Ministry of Education, Sichuan University and also the Fundamental Research Funds for the Central Universities (No. SCU2023D008), and the National Natural Science Foundation of China (No. 61972366).

References

[1] Dimitri P. Bertsekas and John N. Tsitsiklis, *Introduction to Probability* (2nd Edition), Posts and Telecom Press, 2008.
[2] Morris H. DeGroot and Mark J. Schervish, *Probability and Statistics* (4th Edition), Pearson Press, 2011.
[3] Will Kurt, *Bayesian Statistics the Fun Way: Understanding Statistics and Probability with Star Wars, LEGO, and Rubber Ducks*, No Starch Press, 2019.
[4] Bayes' Theorem, https://www.3blue1brown.com/lessons/bayes-theorem#title, Published Dec. 22, 2019, Updated Aug. 28, 2025.

Note: Figure translations are in progress. See original paper for figures.