Theories of Probability

CSCI/DSCI 575 Advanced Machine Learning

Department of Computer Science
Colorado School of Mines

The definition of probability

Probability

- Probability (or likelihood) is a measure of how likely it is that something will happen or that a statement is true.
- Probabilities are given a value between 0 (will not happen) and 1 (will happen).
- The higher the probability of an event, the more certain we are that the event will happen.

— from Wikipedia

A more formal definition

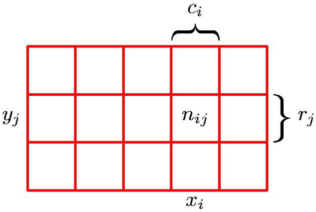
Probability P is a positive real number describing the likelihood of an event that satisfies:

- For any event, $P(X = e) \in [0, 1]$.
- For a mutually **exclusive** and **exhaustive** set of events E, $\sum_{X \in E} P(X) = 1$.

Starting from a general probability problem

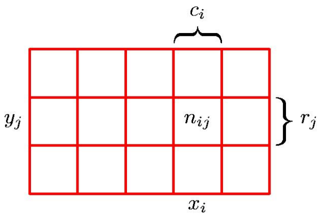
Considering a general setting for two random variables,

- X, which takes the values $\{x_i\}$ where $i = 1, \ldots, M$,
- and Y , which takes the values $\{y_j\}$ where $j = 1, \ldots, L$.



- In the illustration we have $M = 5$ and $L = 3$.
- If we consider a total number N of instances of these variables, then we denote the number of instances where $X = x_i$ and $Y = y_j$ by n_{ij}, which is the number of points in the corresponding cell of the array.
- The number of points in column i, corresponding to $X = x_i$, is denoted by c_i,
- and the number of points in row j, corresponding to $Y = y_j$, is denoted by r_j.

Joint probability



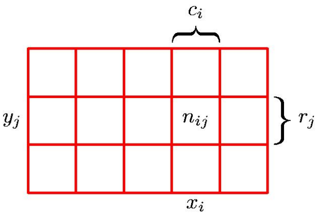
The probability that X will take the value x_i and Y will take the value y_j is written $P(X = x_i, Y = y_j)$ and is called the **joint probability** of $X = x_i$ and $Y = y_j$.

It is given by the number of points falling in the cell i, j as a fraction of the total number of points, and hence

$$P\left(X = x_i, Y = y_j\right) = P\left(X = x_i \cap Y = Y_j\right) = \frac{n_{ij}}{N} \quad ,$$

when $N \to \infty$.

Sum rule and marginal probability



The probability that X takes the value x_i irrespective of the value of Y is written as $P(X = x_i)$.

$P(X = x_i)$ is given by the fraction of the total number of points that fall in column i:

$$P\left(X = x_i\right) = \frac{c_i}{N} \ ,$$

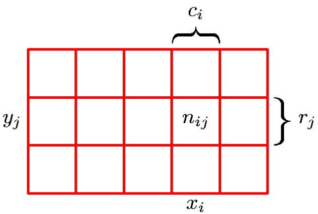
when $N \to \infty$.

Obviously, we have $c_i = \sum_j n_{ij}$, thus

$$P\left(X = x_i\right) = \sum_{j=1}^{L} P\left(X = x_i, Y = y_j\right) \ ,$$

which is called as **sum rule**. $P\left(X = x_i\right)$ is called as **marginal probability**.

Sum rule and marginal probability



The probability that X takes the value x_i irrespective of the value of Y is written as $P(X = x_i)$.

$P(X = x_i)$ is given by the fraction of the total number of points that fall in column i:

$$P\left(X = x_i\right) = \frac{c_i}{N} \ ,$$

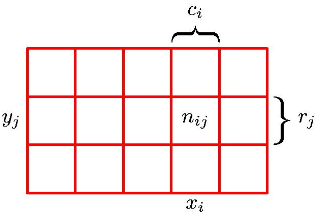
when $N \to \infty$.

Obviously, we have $c_i = \sum_j n_{ij}$, thus

$$P\left(X = x_i\right) = \sum_{j=1}^{L} P\left(X = x_i, Y = y_j\right) \ ,$$

which is called as **sum rule**. $P\left(X = x_i\right)$ is called as **marginal probability**.

Product rule and conditional probability



If we consider only those instances for which $X = x_i$, then the fraction of such instances for which $Y = y_j$ is written $P(Y = y_j|X = x_i)$ and is called the **conditional probability** of $Y = y_j$ given $X = x_i$.

$P(Y = y_j|X = x_i)$ is obtained by finding the fraction of those points in column i that fall in cell i, j:

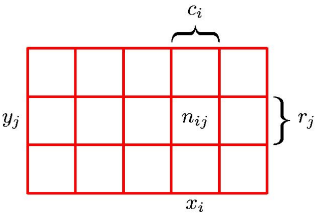
$$P(Y = y_j|X = x_i) = \frac{n_{ij}}{c_i}.$$

Then we can easily derive

$$P(Y = y_j, X = x_i) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N} = P(Y = y_j|X = x_i)P(X = x_i),$$

which is called as **product rule**.

Product rule and conditional probability



If we consider only those instances for which $X = x_i$, then the fraction of such instances for which $Y = y_j$ is written $P(Y = y_j | X = x_i)$ and is called the **conditional probability** of $Y = y_j$ given $X = x_i$.

$P(Y = y_j | X = x_i)$ is obtained by finding the fraction of those points in column i that fall in cell i, j:

$$P(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}.$$

Then we can easily derive

$$P(Y = y_j, X = x_i) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N} = P(Y = y_j | X = x_i) P(X = x_i),$$

which is called as **product rule**.

The rules of probability

The sum rule

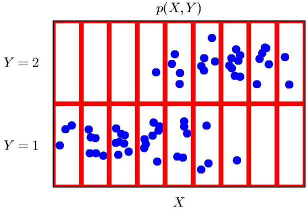
$$p(X) = \sum_Y p(X, Y) \quad .$$

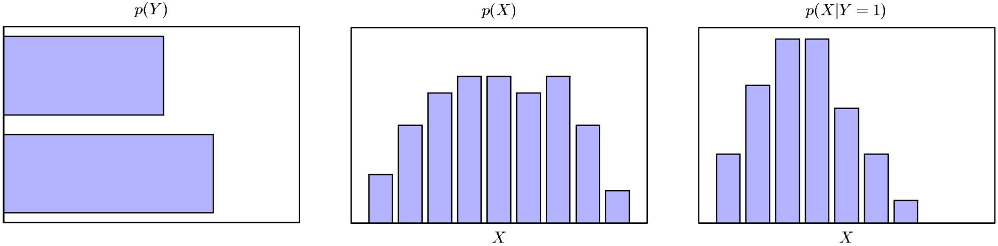
The product rule

$$p(X, Y) = p(X|Y)\, p(Y) = p(Y|X)\, p(X) \quad .$$

An illustrative example

An illustration of a distribution over two variables, X, which takes 9 possible values, and Y, which takes two possible values. The right figure shows a sample of 60 points drawn from a joint probability distribution over these variables. The remaining figures below show histogram estimates of the marginal distributions $P(X)$ and $P(Y)$, as well as the conditional distribution $P(X|Y=1)$ corresponding to the bottom row in the right figure.





Bayes' theorem

From the product rule, together with the symmetry property $P(X, Y) = P(Y, X)$, we immediately obtain the following relationship between conditional probabilities:

$$P(Y|X) = \frac{P(X|Y)\,P(Y)}{P(X)} \; ,$$

which is called as **Bayes' theorem** and plays a central role in pattern recognition and machine learning.

Using the sum rule, the denominator in Bayes' theorem can be expressed in terms of the quantities appearing in the numerator

$$P(X) = \sum_{Y} P(X|Y)\,P(Y) \; .$$

Bayes' theorem

From the product rule, together with the symmetry property $P(X, Y) = P(Y, X)$, we immediately obtain the following relationship between conditional probabilities:

$$P(Y|X) = \frac{P(X|Y)\,P(Y)}{P(X)} \ ,$$

which is called as **Bayes' theorem** and plays a central role in pattern recognition and machine learning.

Using the sum rule, the denominator in Bayes' theorem can be expressed in terms of the quantities appearing in the numerator

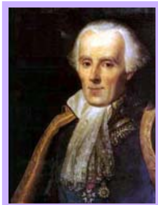
$$P(X) = \sum_{Y} P(X|Y)\,P(Y) \ .$$

A piece of history on probability

Thomas Bayes

1701–1761

Thomas Bayes was born in Tunbridge Wells and was a clergyman as well as an amateur scientist and a mathematician. He studied logic and theology at Edinburgh University and was elected Fellow of the Royal Society in 1742. During the 18$^{\text{th}}$ century, issues regarding probability arose in connection with gambling and with the new concept of insurance. One particularly important problem concerned so-called inverse probability. A solution was proposed by Thomas Bayes in his paper ‘Essay towards solving a problem in the doctrine of chances’, which was published in 1764, some three years after his death, in the *Philosophical Transactions of the Royal Society*. In fact, Bayes only formulated his theory for the case of a uniform prior, and it was Pierre-Simon Laplace who independently rediscovered the theory in general form and who demonstrated its broad applicability.



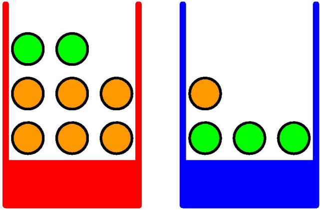
Pierre-Simon Laplace

1749–1827

It is said that Laplace was seriously lacking in modesty and at one point declared himself to be the best mathematician in France at the time, a claim that was arguably true. As well as being prolific in mathematics, he also made numerous contributions to astronomy, including the nebular hypothesis by which the earth is thought to have formed from the condensation and cooling of a large rotating disk of gas and dust. In 1812 he published the first edition of *Théorie Analytique des Probabilités*, in which Laplace states that “probability theory is nothing but common sense reduced to calculation”. This work included a discussion of the inverse probability calculation (later termed Bayes’ theorem by Poincaré), which he used to solve problems in life expectancy, jurisprudence, planetary masses, triangulation, and error estimation.

Probability of drawing a fruit from a jar

We have two colored boxes, and each contains two types of fruits (apples shown in green and oranges shown in orange).



We know that

$$P(B = r) = 4/10, \qquad \text{and} \qquad P(B = b) = 6/10.$$

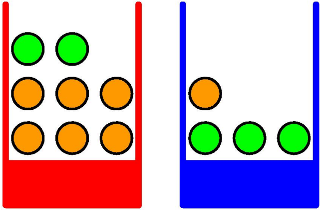
We also know that

$$P(F = a|B = r) = 1/4, \qquad P(F = o|B = r) = 3/4,$$
$$P(F = a|B = b) = 3/4, \qquad P(F = o|B = b) = 1/4,$$

What have you observed?

Probability of drawing a fruit from a jar

We have two colored boxes, and each contains two types of fruits (apples shown in green and oranges shown in orange).



We know that

$$P(B=r) = 4/10, \qquad \text{and} \qquad P(B=b) = 6/10.$$

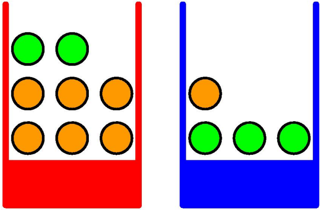
We also know that

$$P(F=a|B=r) = 1/4, \qquad P(F=o|B=r) = 3/4,$$
$$P(F=a|B=b) = 3/4, \qquad P(F=o|B=b) = 1/4,$$

What have you observed?

Probability of drawing a fruit from a jar

We have two colored boxes, and each contains two types of fruits (apples shown in green and oranges shown in orange).



What is the overall probability that the selection procedure will pick an apple?

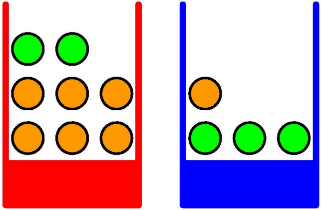
$$p(F = a) = p(F = a|B = r)p(B = r) + p(F = a|B = b)p(B = b) = \frac{1}{4} \times \frac{4}{10} + \frac{3}{4} \times \frac{6}{10} = \frac{11}{20} \ .$$

Given that we have chosen an orange, what is the probability that the box we chose was the red one?

$$p(B = r|F = o) = \frac{p(F = o|B = r)p(B = r)}{p(F = o)} = \frac{3}{4} \times \frac{4}{10} \times \frac{20}{9} = \frac{2}{3} \ .$$

Probability of drawing a fruit from a jar

We have two colored boxes, and each contains two types of fruits (apples shown in green and oranges shown in orange).



What is the overall probability that the selection procedure will pick an apple?

$$p(F=a) = p(F=a|B=r)p(B=r) + p(F=a|B=b)p(B=b) = \frac{1}{4} \times \frac{4}{10} + \frac{3}{4} \times \frac{6}{10} = \frac{11}{20} \ .$$

Given that we have chosen an orange, what is the probability that the box we chose was the red one?

$$p(B=r|F=o) = \frac{p(F=o|B=r)p(B=r)}{p(F=o)} = \frac{3}{4} \times \frac{4}{10} \times \frac{20}{9} = \frac{2}{3} \ .$$

Probability of drawing a fruit from a jar

The interpretation of Bayes' theorem.

- If we had been asked which box had been chosen before being told the identity of the selected item of fruit, then the most complete information we have available is provided by the probability $P(B)$.
  - We call this the **prior probability** because it is the probability available *before* we observe the identity of the fruit.
- Once we are told that the fruit is an orange, we can then use Bayes' theorem to compute the probability $P(B|F)$.
  - We call this the **posterior probability** because it is the probability obtained *after* we have observed F.

Probability of drawing a fruit from a jar

The interpretation of Bayes' theorem.

- If we had been asked which box had been chosen before being told the identity of the selected item of fruit, then the most complete information we have available is provided by the probability $P(B)$.
  - We call this the **prior probability** because it is the probability available *before* we observe the identity of the fruit.
- Once we are told that the fruit is an orange, we can then use Bayes' theorem to compute the probability $P(B|F)$.
  - We call this the **posterior probability** because it is the probability obtained *after* we have observed F.

Evidence is useful!

- The prior probability $P(B = r) = 4/10$ indicates we are likely to select the blue box,
- The posterior probability $P(B = r|F = o) = 2/3$ indicates we are likely to select the red box.

Copyright statement