Quantifying Uncertainty

CSCI/DSCI 575 Advanced Machine Learning

Department of Computer Science
Colorado School of Mines

Uncertainty is everywhere

- Will it be raining in tomorrow?
- What is the outcome of rolling dice?
- What is the speed of my car (in MPH)?
- ...

Uncertainty is everywhere

Uncertainty is a native property of the real-world data.

- Uncertainty arises through noise on measurements.
- Uncertainty also arises through the finite size of data sets (incomplete observation).

Probability theory is a foundation of machine learning.

- **Probability theory** provides a consistent framework for the **quantification and manipulation** of uncertainty and forms one of the central foundations for pattern recognition and machine learning.
- When combined with **decision theory**, it allows us to **make optimal predictions** given all the information available to us, even though that information may be incomplete or ambiguous.

Uncertainty is everywhere

Uncertainty is a native property of the real-world data.

- Uncertainty arises through noise on measurements.
- Uncertainty also arises through the finite size of data sets (incomplete observation).

Probability theory is a foundation of machine learning.

- **Probability theory** provides a consistent framework for the **quantification and manipulation** of uncertainty and forms one of the central foundations for pattern recognition and machine learning.
- When combined with **decision theory**, it allows us to **make optimal predictions** given all the information available to us, even though that information may be incomplete or ambiguous.

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, $\sum_{X \in E} P(X) = 1$.

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

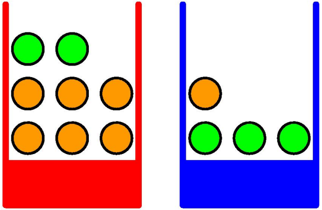
A more formal definition

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, $\sum_{X \in E} P(X) = 1$.

Probability of drawing a fruit from a jar

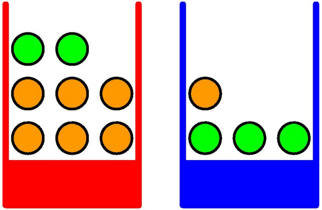
We have two colored boxes, and each contains two types of fruits (apples shown in green and oranges shown in orange).



- Suppose we randomly pick one of the boxes and from that box we randomly select an item of fruit.
- We have observed which sort of fruit it is.
- We could imagine repeating this process many times. Let us suppose
  - that in so doing we pick the red box 40% of the time and we pick the blue box 60% of the time,
  - and that when we remove an item of fruit from a box we are equally likely to select any of the pieces of fruit in the box.

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).

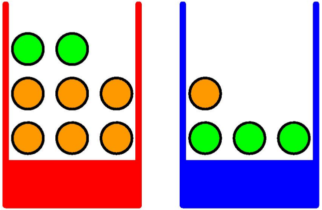


- Suppose we randomly pick one of the boxes and from that box we randomly select an item of fruit.
- We have observed which sort of fruit it is.
- We could imagine repeating this process many times. Let us suppose
  - that in so doing we pick the red box 40% of the time and we pick the blue box 60% of the time,
  - and that when we remove an item of fruit from a box we are equally likely to select any of the pieces of fruit in the box.

Probability of drawing a fruit from a jar

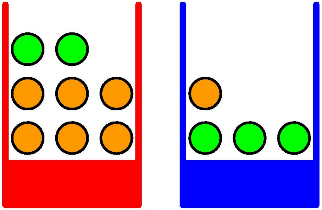
Formally, we define two random variables to denote the selected box by B and the selected fruit by F.

- $B \in \{b, r\}$ where b indicates the **blue** color and r indicates the **red** color;
- $F \in \{a, o\}$ where a indicates the **apple** and o indicates the **orange**;
- $P(B = r) = \frac{4}{10}$ and $P(B = b) = \frac{6}{10}$.



Probability of drawing a fruit from a jar

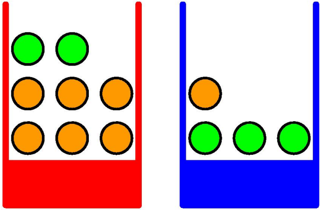
Given the available information that $P(B = r) = \frac{4}{10}$ and $P(B = b) = \frac{6}{10}$, can you answer the following questions?



- What is the overall probability that the selection procedure will pick an apple?
- Given that we have chosen an orange, what is the probability that the box we chose was the red one?

Probability of drawing a fruit from a jar

Given the available information that $P(B = r) = \frac{4}{10}$ and $P(B = b) = \frac{6}{10}$, can you answer the following questions?



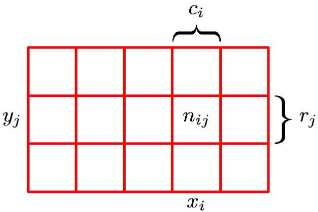
- What is the overall probability that the selection procedure will pick an apple?
- Given that we have chosen an orange, what is the probability that the box we chose was the red one?

Yes, but we need some statistical tools . . .

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 $j = 1, \ldots, L$.

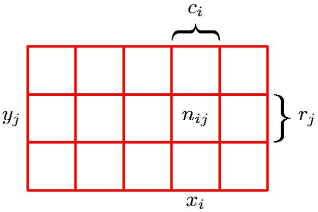


- In the illustration we have $M = 5$ and $L = 3$.
- If we consider a total number N of instances of these variables, then we denote the number of instances where $X = x_i$ and $Y = y_j$ by n_{ij}, which is the number of points in the corresponding cell of the array.
- The number of points in column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

Starting from a general probability problem

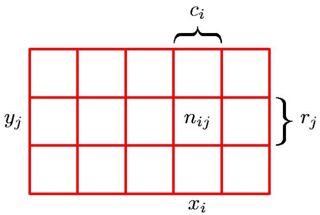
Considering a general setting for two random variables,

- 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 $N \to \infty$.

Copyright statement