Probability for Deep Learning
Part 1

Johar M. Ashfaque

A. Why Probability?

Probability is the mathematical language for quantifying uncertainty. We can apply probability theory to a diverse set of problems from flipping a coin to the analysis of computer algorithms. The starting point is to specify the sample space, the set of all the possible outcomes.

B. Sample Space and Events

The sample space Ω is the set of all the possible outcomes of an experiment. Events are subsets of Ω.

Example 1. If we flip a coin twice then $\Omega = \{HH, HT, TH, TT\}$. The event that the first flip gives a head is $A = \{HH, HT\}$.
Example 2. Let ω be the outcome of a measurement of some physical quantity, say temperature. Then $\Omega = \mathbb{R}$. The event that the measurement is larger than 10 but less than or equal to 23 is $A = (10, 23]$.

Given an event A, let A^c denote the complement of A. Informally, A^c can be read as "not A". The complement of Ω is the empty set $\emptyset$.

Ω	Sample Space
ω	outcome
A	event
A^c	complement of A
$A \cup B$	union (A or B)
$A \cap B$	intersection (A and B)
$A - B$	set difference (points in A that are not in B)
$A \subset B$	set inclusion (A is a subset of or equal to B)
$\emptyset$	null event (always false)

TABLE I. Sample Space and Events

We say that $A_1, A_2, ...$ are disjoint or mutually exclusive if $A_i \cap A_j = \emptyset$ whenever $i \neq j$. A partition of Ω is a sequence of disjoint sets $A_1, A_2, ...$ such that $\cup_{i=1}^{\infty} A_i = \Omega$.

C. Probability Measure

We want to assign a real number $\mathbb{P}(A)$ to every event A called the probability of A. We also call $\mathbb{P}$ a probability distribution or a probability measure. To qualify as a probability, $\mathbb{P}$ has to satisfy three axioms.

Definition .1 *A function $\mathbb{P}$ that assigns a real number $\mathbb{P}(A)$ to each event A is a probability distribution or a probability measure if it satisfies the following three axioms:*
- $\mathbb{P}(A) \geq 0$ *for every* A
- $\mathbb{P}(\Omega) = 1$

- *If $A_1, A_2, ...$ are disjoint then*

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

One can derive many properties of the function $\mathbb{P}$ from these axioms. Here are a few:
- $\mathbb{P}(\emptyset) = 0$
- $A \subset B \Longrightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$
- $0 \leq \mathbb{P}(A) \leq 1$
- $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$
- $A \cap B = \emptyset \Longrightarrow \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$

A less obvious property is given by the following Lemma.

Lemma .2 *For any events A and B*

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

Example 3. Flip two coins. Let H_1 be the event that heads occurs on flip 1 and let H_2 be the event that heads occurs on flip 2. If all outcomes are equally likely, that is

$$\mathbb{P}(\{H_1, H_2\}) = \mathbb{P}(\{H_1, T_2\}) = \mathbb{P}(\{T_1, H_2\}) = \mathbb{P}(\{T_1, T_2\}) = \frac{1}{4}$$

then

$$\mathbb{P}(H_1 \cup H_2) = \mathbb{P}(H_1) + \mathbb{P}(H_2) - \mathbb{P}(H_1 \cap H_2) = \frac{1}{2} + \frac{1}{2} - \frac{1}{4} = \frac{3}{4}.$$

D. Probability on Finite Sample Spaces

Suppose that the sample space $\Omega = \{\omega_1, ..., \omega_n\}$ is finite. For example, if a dice is thrown twice then Ω has 36 elements in total. If each outcome is equally likely then $\mathbb{P}(A) = \frac{A}{36}$ where A denotes the number of elements in A. The probability that the sum of the dice is 11 is $\frac{2}{36}$ since there are two outcomes that correspond to this event. In general, if Ω is finite and if each outcome is equally likely then

$$\mathbb{P}(A) = \frac{A}{\Omega}$$

which is called the uniform probability distribution. To compute probabilities, we need to count the number of points in an event A using combinatorial techniques. Given n objects, the number of ways of ordering these objects is $n!$. For convenience, we define $0! = 1$. We also define

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

which is the number of distinct ways of choosing k objects from n. For example, if we have a class of 20 people and we want to choose a committee of 3 students then there are

$$\binom{20}{3} = \frac{20!}{3!17!} = 1140$$

possible committees. Note the following property:

$$\binom{n}{0} = \binom{n}{n} = 1.$$

E. Independent Events

If we flip a fair coin twice, then the probability of two heads is $\frac{1}{2} \times \frac{1}{2}$. We multiply the probabilities because we regard the two tosses as independent. The formal definition of independence is as follows.

Definition .3 *Two events A and B are independent if*

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

Independence can arise in two distinct ways. Sometimes, we explicitly assume that the two event are independent. In other instances, we derive independence by verifying that $\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B)$ holds. Suppose that A and B are disjoint events, each with positive probability. Can they be independent? The answer is no. This follows since $\mathbb{P}(A)\mathbb{P}(B) > 0$ yet $\mathbb{P}(AB)\mathbb{P}(\emptyset) = 0$.

1) A and B are independent if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.
2) Independence is sometimes assumed and sometimes derived.
3) Disjoint events with positive probability are not independent.

TABLE II. Summary of Independence

F. Conditional Probability

Assuming that $\mathbb{P}(B) > 0$, we define the conditional probability of A given that B has occurred as follows.

Definition .4 *If $\mathbb{P}(B) > 0$ then the conditional probability of A given B is*

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

If A and B are independent events then

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

From the definition of conditional probability, we can write $\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B)$ and also

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

Often these formulae give us a convenient way to compute $\mathbb{P}(A \cap B)$ when A and B are not independent.

1) If $\mathbb{P}(B) > 0$ then the conditional probability of A given B is $\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$.
2) In general, $\mathbb{P}(A|B) \neq \mathbb{P}(B|A)$.
3) A and B are independent if and only if $\mathbb{P}(A|B) = \mathbb{P}(A)$.

TABLE III. Summary of Conditional Probability