A.2 Probability

Probability is central to many of the algorithms used for Intelligent Systems. Probability can be thought of as the mathematics of random experiments. A random experiment is a an action that can result in many possible outcomes. For example, a system to recognize a certain person's face in color images can be designed using many images of that person. In this case, the action is taking the picture and the outcome is the corresponding digital image of the person. Every image is different, even if several images are taken at intervals of a few seconds.

Discrete probability seems straightforward conceptually. Assigning probabilities to outcomes of experiments with a finite number of outcomes, such as flipping coins or rolling dice, is very intuitive. By contrast, continuous probabilities, such as the probability that the temperature is exactly 30° or that the power associated with light radiated by the sun at a specific square meter on the earth's surface is exactly 1369.401957 watts is much more difficult to define. In fact, there are even discrete events for which the proper assignment of probability is non-intuitive. These concepts are all discussed in the following two sections.

A.2.1 Discrete, or Finite, Probability

In this section, Coin Flips and Rolling Dice are experiments used to illustrate and motivate the basic tenets of probabilities. A formal definition of discrete probability measures is given provided. Some definitions and computational examples of complete the section.

Every time one flips a coin, the outcome can either be heads or tails. We assume that the coin cannot land on it's edge. There are two probabilities, probability of heads, $P(H)$, and probability of tails, $P(T)$. Since this is all the possible outcomes, it is assumed that $P(H) + P(T) = 1$. Every time one rolls a die, there are 6 possible outcomes, $\mathcal{D} = \{1, 2, 3, 4, 5, 6\}$. It is assumed that, $\sum_{k=1}^{6} P(k) = 1$.

The values of the probabilities are only constrained by requiring that they are between 0 and 1. For example, the two sets of assignments $P(H) = 0.5$ and $P(T) = 0.5$ or $P(H) = 0.75$ and $P(T) = 0.25$ are both valid assignments of probabilities to the outcomes. $P(H) = 0.75$ and $P(T) = 0.75$ is not allowed because the probabilities should some to 1.

Once one specifies values of the individual outcomes, then one can assign probabilities to sets of outcomes. For example, the probability that the outcome is an even number can be thought of as the probability of the set $\mathcal{E} = \{2, 4, 6\}$ and that the outcome is an odd number as the probability of the set $\mathcal{O} = \{1, 3, 5\}$. Since these sets do not overlap, that is, since these sets are *disjoint* and since together they contain all the outcomes, it must be true that $P(\{2, 4, 6\}) + P(\{1, 3, 5\}) = 1$. For the same reason, the probabilities that the outcome is either 1 or 2, $P(\{1, 2\})$ or that the outcome is either 3 or 4, $P(\{1, 2\})$ should satisfy $P(\{1, 2\}) + P(\{1, 2\}) = P(\{1, 2, 3, 4\})$. Note that, in the language of sets, $\mathcal{E} =$ called the *complement* of $\mathcal{O}$ since $\mathcal{E} \cap \mathcal{O} = \emptyset$ and $\mathcal{E} \cup \mathcal{O} = \mathcal{D}$.

Outcomes that are defined by combining one or more individual outcomes, are called *events*. It is always assumed that there is an outcome, e.g. the coin does not disappear when it is flipped. This latter case is expressed mathematically by stating that the probability of the empty set is 0, $P(\{\}) = P(\phi) = 0$. These observations lead to two formal definitions. The first is the definition of a sample space. The sample space and associated notion of a σ-algebra means of identifying all the events of interest in a fashion that is compatible with how probabilities should behave. The sample space is the set of all outcomes of an experiment. One can think of it as the events that can be assigned probabilities. The second is the definition of a discrete probability measure.

Definition of **Sample Space**. A *Sample Space* is a set of all possible outcomes of an experiment.

Example. The sets $\{H, T\}$ and $\{1, 2, 3, 4, 5, 6\}$ are sample spaces.

Definition of **(Discrete) σ-algebra**. Let $\mathcal{X}$ represent a finite set and $2^{\mathcal{X}}$ denote the power set of $\mathcal{X}$, that is, $2^{\mathcal{X}} = \{A | A \subset \mathcal{X}\}$. A *Sample Space* $\mathcal{S}$ on $\mathcal{X}$ is a subset of $2^{\mathcal{X}}$, $\mathcal{S} \subseteq 2^{\mathcal{X}}$ with certain properties. The properties are given below; the format is for each property is the motivation for the property followed by the mathematical formulation of the property.

1. The probability that no event occurs can be assigned. $\emptyset \in \mathcal{S}$
2. If an event can be assigned probability, then the opposite of the event can be assigned probability. If $\mathcal{A} \in \mathcal{S}$, then $A^c \in \mathcal{S}$
3. If two events, $\mathcal{A}, \mathcal{B}$, can be assigned probability, then $\mathcal{A}$ **or** $\mathcal{B}$ can be assigned probability. If $\mathcal{A}, \mathcal{B} \in \mathcal{S}$, then $\mathcal{A} \cup \mathcal{B} \in \mathcal{S}$
4. If two events, $\mathcal{A}, \mathcal{B}$, can be assigned probability, then $\mathcal{A}$ **and** $\mathcal{B}$ can be assigned probability. If $\mathcal{A}, \mathcal{B} \in \mathcal{S}$, then $\mathcal{A} \cap \mathcal{B} \in \mathcal{S}$

Example. Note that $2^{\mathcal{X}}$ is a σ-algebra.

Example. The collection of sets $\{\emptyset, \{1, 2\}, \{3, 4\}, \{1, 2, 3, 4\}\}$ is a σ-algebra.

Definition of **Discrete Probability Measure**. Suppose $\mathcal{E}$ is a random experiment with potential outcomes represented by a finite set $\mathcal{X}$. Let $\mathcal{S}$ denote a sample space on $\mathcal{X}$. A *Probability Measure*, or *Probability Distribution*, on $\mathcal{X}$ is a function $P : \mathcal{S} \to [0, 1]$ with the properties that:

$$P(\emptyset) = 0,$$
$$P(\mathcal{S}) = 1,$$
$$\text{If } A, B \subset \mathcal{S} \text{ and } A \cap B = \phi, \text{ then }, P(A \cup B) = P(A) + P(B)$$

Example. Rolling a Die and a Coin. Suppose one person rolls a 4-sided die and another person flips a coin and the outcomes are observed. A sample space for this experiment is $\mathcal{S} = \{\{1, H\}, \{2, H\}, \{3, H\}, \{4, H\}, \{1, T\}, \{2, T\}, \{3, T\}, \{4, T\}\}$. A σ-algebra, Σ, can be constructed from this sample space by taking Σ to be the set of all unions and intersections of elements of $\mathcal{S}$. Let $N = |\mathcal{S}|$. There are $C_{N,1} = N$ ways to create unions of size 1, $C_{N,2}$ ways to create unions of size 2, etc., Therefore, there are $C_{N,1} + C_{N,2} + \ldots + C_{N,N-1} + 2 = 2^N$ elements in Σ.

Definition of **Frequency Histogram**. Let $\mathcal{E}$ be an experiment with a finite number of outcomes, call them $\mathcal{O}_1, \mathcal{O}_2, \ldots, \mathcal{O}_T$. Assume the experiment is performed N times and denote the number of times $\mathcal{O}_t$ occurs by N_t. Note that $\sum_{t=1}^{T} N_t = N$. The *Frequency Histogram*, or just *Histogram*, associated with these N experiments is the set of pairs $\mathcal{H} = \{(\mathcal{O}_1, N_1), (\mathcal{O}_2, N_2), \ldots, (\mathcal{O}_T, N_T)\}$. The frequency histogram is usually plotted as a bar chart. The numbers, N_t are sometimes called the *frequency of occurrence* of outcome $\mathcal{O}_t$ [1].

[1] The frequency of occurrence should not be confused with the frequency associated with wave phenomena.

Definition of **Normalized Histogram**. Let $\mathcal{H}$ be a frequency histogram. A *Normalized Histogram* $\mathcal{N}$ associated with $\mathcal{H}$ is calculated from $\mathcal{H}$ by dividing every N_t by N, so $\mathcal{N}$ = $\left\{\left(\mathcal{O}_1, \frac{N_1}{N}\right), \left(\mathcal{O}_2, \frac{N_2}{N}\right), \ldots, \left(\mathcal{O}_T, \frac{N_T}{N}\right)\right\}$. A normalized histogram is a discrete probability distribution.

Example. If one rolls a six-sided die N = 12 times and even numbers are the outcome 3 times each and odd numbers are the outcome 1 time each, then $N_1 = N_3 = N_6 = 1$, $N_2 = N_4 = N_6 = 3$. The normalized frequencies of occurrence are obtained by dividing by 12. The frequency and normalized histograms are shown in Figure A.5.

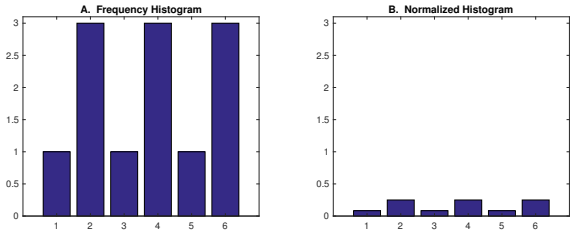

Figure A.5 A Frequency Histogram and the Normalized Histogram associated with it.

Example. Data Mining Example Suppose one has a computer program that searches the abstracts of reports by the National Academy of Science with the word Climate in the title and generates a frequency histogram of every "significant" word in the abstract.

Definition of **Moments of a Discrete Probability Distribution,** $P(x)$.. The moments, denoted $\mu_1, \mu_2, \ldots$ and centralized moments, denoted $\gamma_1, \gamma_2, \ldots$ are defined by

$$\mu_n = \sum_x \left(x^n P(x)\right) \qquad \text{for } n = 1, 2, \ldots$$

$$\gamma_n = \sum_x \left((x - \mu_1)^n P(x)\right) \qquad \text{for } n = 2, 3, \ldots$$

Definition of **Multinomial and Binomial Distributions**. Let $\mathcal{X}$ be a finite set with $|\mathcal{X}| = K$. A *Multinomial Distribution* on $\mathcal{X}$ is a distribution $P : \mathcal{X} \rightarrow [0, 1]$ where $\sum_{x \in \mathcal{X}} P(x) = 1$. A *Binomial Distribution* is a multinomial with $K = 2$. A multinomial can be represented as a set of ordered pairs $\mathcal{P}$ = $\{(x_1, P(x_1)), (x_2, P(x_2)), \ldots, (x_K, P(x_K))\}$.

There are many other discrete probability distributions but we the multinomial is most often used in intelligent systems.

A.2.2 Continuous, or Infinite, Probability

As mentioned previously, defining probabilities to experiments with infinitely many possible outcomes is not as straightforward as with finitely many possible outcomes. Fortunately, mathematicians worked out a method for a theory of probability that resolves these issues. A rigorous development cannot be followed here and would not be that useful for practitioners. However, there are a few notions that cause confusion and are therefore addressed here.

A.2.2.1 Overview of Theory of Probability Let $\mathcal{X}$ be a sample space. The definition of a σ-algebra is the same except that the properties of being closed under unions is extended to a *countably infinite* number of sets. If $\mathcal{S}$ is a σ-algebra and if $A_n \in \mathcal{S}$ for $n = 1, 2, \ldots,$ then $\cup_{n=1}^{\infty} A_n \in \mathcal{S}$. A non-rigorous definition of a probability measure is:

Definition of **Probability Measure or Distribution**. A *Probability Measure* is a function $P : \mathcal{S} \to [0, 1]$ with the properties that:

1. If $A \in \mathcal{S}$ then $P(A) \geq 0$.
2. $P(\mathcal{X}) = 1$ and $P(\emptyset) = 0$.
3. If $A_n \in \mathcal{S}$ for $n = 1, 2, \ldots$ and $n \neq m$ implies that $A_n \cap A_m = \emptyset$, then

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n)$$

Probability measures are also called *Probability Distributions*. It is usually the case that a Probability Measure is defined using integration.

Definition of **Probability Space**. A *Probability Space* is a triple $(\mathcal{X}, \mathcal{S}, P)$ where $\mathcal{X}$ is a set, $\mathcal{S}$ is a σ-algebra on $\mathcal{X}$, and P is a probability measure.

Example. Borel σ-algebra A rough, non-rigorous definition of the *Borel* σ-algebra, denoted by $\mathcal{B}_\sigma$ is that it is the set of all countable unions, countable intersections, and complements of open intervals on $\mathbb{R}$. Here, an element of the Borel σ-algebra will be called *measurable* which intuitively means that a probability can be assigned to the set. The probability can be assigned by integration, as described in what follows.

Definition of **Random Variable**. Suppose $\mathcal{P}$=$(\mathcal{X}, \mathcal{S}, P)$ is a probability space. A real-valued *Random Variable* on $\mathcal{P}$ is a function $R : \mathcal{X} \to \mathbb{R}$ with the property that if $B \subset \mathbb{R}$, then the set $R^{-1}(B)$ = $\{x \in \mathcal{X} | R(x) \in B\} \in \mathcal{S}$. This property enables one to assign probability to sets of outcomes of random variables by taking $P_R(B)$ = $P\left(R^{-1}(B)\right)$.

Definition of **Cumulative Distribution Function**. Assume R is a random variable with respect to some probability measure P and for any $x \in \mathbb{R}$, define $C_x = \{a \in \mathbb{R} | a \leq x\}$. The *Cumulative Distribution Function* (cdf) of R is the function $F(x) = P(C_x)$.

Definition of **Probability Density Function**.. Assume P is a probability distribution and $f : \mathbb{R} \to \mathbb{R}^+$. If f satisfies the property that, $\forall A \subset \mathcal{B}_\sigma$, $P(A)$ = $\int_A f(x)\, dx$, then f is called the *Probability Density Function*(pdf) of P.

Notice that, if F is a cumulative distribution function, then $F(x)$ = $\int_{-\infty}^{x} f(x)\, dz$.