1 Introduction

Today we continue our intro to the basic probability theory. Last time we defined conditional probability $\mathbf{P}(A|B)$ as probability of event A to happen if B happened:

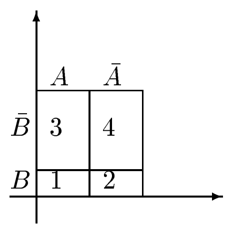
$$\mathbf{P}(A|B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)}$$

or we can rewrite it as

$$\mathbf{P}(B)\mathbf{P}(A|B) = \mathbf{P}(A \cap B),$$

sometimes it is useful to denote $\mathbf{P}(X|A)$ as $\mathbf{P}'(X)$. Let A_i is a partition of the Ω (i.e. $\cap A_i = \Omega$, $i \neq j \Rightarrow A_i \cap A_j = 0$) then $\mathbf{P}(B) = \sum \mathbf{P}(B|A_i)\mathbf{P}(A_i)$. If A and B are independent then $\mathbf{P}(A|B) = \mathbf{P}(A)$. The following example illustrates how can we use independence.

Assume we picked x and y independently from interval $[0, 1]$, let A denotes evenet that $x \leq 0.5$ and B denotes event that $y \leq 0.1$. This corresponds to the following picture:



Then we have

$$\mathbf{P}(A|B) = \frac{\mathbf{P}(A \cup B)}{\mathbf{P}(B)} = \frac{|1|}{|1| + |2|} = \frac{|1| + |3|}{|1| + |2| + |3| + |4|} = \mathbf{P}(A).$$

Another useful notion is notion of conditional independence:

$$\mathbf{P}(A \cup B|C) = \mathbf{P}(A|C)\mathbf{P}(B|C)$$

to shorten formula it is convinient to use P' as defined above, so we can write conditional independence in the form

$$\mathbf{P}'(A \cup B) = \mathbf{P}'(A)\mathbf{P}'(B).$$

The following theorem is also sometimes very useful:

Theorem 1 (Bayes) *If A_i is a partition and E is an event. Then we have*

$$\mathbf{P}(A_k|E) = \frac{\mathbf{P}(A_k \cup E)}{P(E)} = \frac{\mathbf{P}(E|A_k)P(A_k)}{\sum \mathbf{P}(E|A_i)\mathbf{P}(A_i)}$$

The following example is somehow a canonical example of usage of Bayes theorem.

Suppose there is a sequence of jobs, and two machine, each job is assigned randomly with probability $\mathbf{P}(M_1) = 0.001$ to the first machine, and $\mathbf{P}(M_2) = 0.999$, probability of waste for the first machine is $\mathbf{P}(R|M_1) = 0.01$ for the second is $\mathbf{P}(R|M_2) = 0.005$, what is probability that if given job is wasted it was performed by the second machine?

By the Bayes theorem we have

$$\mathbf{P}(M_2|R) = \frac{\mathbf{P}(R|M_2)\mathbf{P}(M_2)}{\sum_i \mathbf{P}(R|M_i)\mathbf{P}(M_i)}$$

.

2 Random variables

Formally random variable x can be defined as a function $x : \Omega \to \mathbf{Z}$ for the integer case or $x : \Omega \to R$ for continuius such that every set of form $\{w|x(w) \leq \alpha\}$ is an event (i.e. measurable). Then we can measure the probability that x is less then α

2.1 Discrete random variables

For discrete variables set $\{w|x(w) = \alpha\}$ is always an event, since we have

$$\{w|\ x(w) = \alpha\} = \{w|\ x(w) \leq \alpha\} - \{w|\ x(w) \leq \alpha - 1\}).$$

Function $F_x(\alpha) = \mathbf{P}(\{w|\ x(w) = \alpha\})$ is called probability mass function (PMF) of random variable x, basically $F_x(\alpha)$ is just probability that observed value of x. Also sometimes we will denote $\mathbf{P}(\{w|\ x(w) = \alpha\})$ as $\mathbf{P}(x = \alpha)$ which is also quite intuitive. Clearly (by definition of $\mathbf{P}$ function we have):

$$F_x(\alpha) \geq 0$$
$$\sum_\alpha F_x(\alpha) = 1$$

2.2 Continuous Variables

By analogy with PMF for continuous random variables we have CDF (cumulative distribution function):

$$F_x(\alpha) = \mathbf{P}(x \leq \alpha);$$

$F_x)$ is monotically nondecreasing function (just because, $\mathbf{P}(x \leq \alpha) \leq \mathbf{P}(x \leq \beta)$ if $\alpha < \beta$, $F_x(-\infty) = \lim_{\alpha \to -\infty} = 0$ and $F_x(\infty) = \lim_{\alpha \to -\infty} = 1$. Also sometimes it is useful to define probability density function (PDF) which is just derivative of $F_x(\alpha)$.

$$f_x(\alpha) = \frac{\mathbf{d}}{\mathbf{d}\alpha} F_x(\alpha)$$

This function has properties similar to PMF function, namely $f_x(\alpha) \geq 0$ and $\int_{-\infty}^{\infty} f_x(\alpha) d\alpha = 1$. However $f_x(\alpha)$ is *not* a probabilty of observing value α of random variable x. For example, if x is choosen uniformly at random from the interval $[0.5, 1.0]$ then $f_x(\alpha) = 2$ for $x \in [0.5, 1.0]$ and 0 elsewhere, but probability could not be more than one. In fact, probability of choosing any particular value in this interval is zero - this is another counter intuitive difference of continuousn random variables, and we can only count the probability that x falls withing any specified interval.

2.3 Expectation, moments and Central moments

Let us return to the discrete variables. Average value of x is called expectation of random variable x.

$$E[x] = \sum_{\alpha \in Z} \alpha F_x(\alpha)$$

We can also define expectation of function of random variable:

$$E[g(x)] = \sum_{\alpha \in Z} g(\alpha) F_x(\alpha)$$

If $g(x) = x^n$ then $E[g(x)]$ is called n-th moment of random variable. If $g(x) = (x - E[x])$ then $E[g(x)]$ is n-th central moment. First few moments have some distinguished meaning. Clearly first moment is just an expectation. First central moment is zero. Second central moment $E\big[(x - E(x))^2\big]$ is a variance σ_x^2, which is used to measure how much the random variable is spreaded near its expectation.

2.4 Multiple Random variables

Assume two discrete randomvariables x and y, then we can think of them as a random vector (x, y), and we can define distribution function as $F_{x,y}(\alpha, \beta) = P(x = \alpha, y = \beta$, which is called joint distribution. Clearly we have

$$\sum_{\alpha} \sum_{\beta} f_{x,y}(\alpha, \beta) = 1$$

.

Also we can do reverse operation, i.e. extract components from the random vectors:

$$f_y(\beta) = \textstyle\sum_{\alpha} f_{xy}(\alpha, \beta) f_x(\alpha) = \sum_{\beta} f_{xy}(\alpha, \beta)$$

These $f_x(\alpha)$ and $f_y(\beta)$ are called marginal distributions. And this operation is called marginalization of vector valued random variable. This operation is a particular case of more general operation. Assume we have random variable $f : \Omega \to Z^2$, and a function $g : Z^2 \to D$. So $g \circ f$ is D-valued random variable. If g is a projection function (i.e. $g(x, y) = x$), then $g \circ f$ is marginal.