Probability Theory

A key concept in the field of pattern recognition is that of uncertainity.

It arises both through noisy measurements, as well as through the finite datasets.

Probability Theory provides a consistent framework for quantification and manipulation of uncertainity and forms one central foundation to the world of pattern recognition.

When we combine decision theory, to be discussed later, with probability theory, it allows us to make much more optimal ~~discussion~~ decisions/ predictions given all the information available to us, even though the info might be incomplete or a bit ambiguous.

let us say that we are "equally likely" to pick up orange or apple from ~~a~~ two box.

Box 1 has 2 apple and 6 oranges

Box 2 has 3 apples and 1 orange.

Now suppose if we randomly select one of the two boxes and from that box we randomly select an item of fruit. After observing the fruit, we~~o~~ keep it back.

Let us suppose that we repeated this process multiple times. We picked ~~[illegible]~~ box 1 40% of time and box 2 60% of time.

In this example, identity of the box be an r.v, denoted by B.

B can either be r (box 1) or b (box 2)

Similarity, identity of fruit is another r.v and will be denoted by F.

F can be a (apple) or o (orange).

Probability is defined as

$$\frac{\text{event occured}}{\text{total number of trials}}$$

Then $p(r) = \frac{4}{10}$, $p(b) = \frac{6}{10}$.

or $p(B = r) = \frac{4}{10}$, $p(B = b) = \frac{6}{10}$

Note, probability $\in [0, 1]$

Also, if the events are mutually exclusive and if they include all the possible outcomes then all those probabilities sum up to one.

—— x ——

Derivation of sum
and product rule.

Let us consider two random variables X and Y.

Suppose X can take any values from 1 to M. ∴ X such that x_i, where $i = 1, \ldots, M$.

Y can take values such as

Y_j, $j = 1 \ldots L$.

Consider total 'N' number of trials in which we sample both X and Y.

Let number of trials in which $X = x_i$ and $Y = y_j$ be n_{ij}.

Also, let the number of trials in which X takes the value x_i (irrespective of what value Y takes) be c_i.

Similarily let number of trials for y_j be r_j.

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 as joint probability of $X = x_i$ and $Y = y_j$.

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N} \quad - (1)$$

Similarily the probability $p(X = x_i)$

$$p(X = x_i) = \frac{c_i}{N} \quad - (2)$$

$$\boxed{c_i = \Sigma_j n_{ij}}$$

from (1) and (2)

$$\boxed{p(X = x_i) = \sum_{j=1}^{L} p(X = x_i, Y = y_j)}$$

Sum rule of probability.

$P(X = x_i)$ is some time called as marginal probability, because it is obtained by marginalizing or summing out the other variables

If we consider only those instances for which $X = x_i$, then the instances where $Y = y_j$ is $P(Y = y_j | X = x_i)$ also known as conditional probability.

$$P(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i} \quad -(3)$$

From (1), (2) & (3)

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

which is the product rule of probability.

Precisely

Sum rule

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

Product rule

$$P(X, Y) = P(Y | X) \cdot P(X)$$

From product rule, and symmetry property

$$P(X,Y) = P(Y,X)$$

we obtain,

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

which is called Bayes' Theorem.

—— x ——.

Expectations and covariances

One of the most important operations involving probabilities is that of finding weighted averages of functions.

The average value of some function $f(x)$ under a p.d $p(x)$ is called the expectation of $f(x)$. is denoted by $\mathbb{E}(x)$

$$\mathbb{E}[f] = \sum_x p(x) f(x)$$

$$\mathbb{E}[f] = \int p(x) f(x) \, dx.$$

conditional expectation

$$\mathbb{E}_x[f|y] = \sum p(x|y) f(x)$$

Variance of $f(x)$

$$\text{Var}[f] = \mathbb{E}\left[\left(f(x) - \mathbb{E}[f(x)]\right)^2\right]$$

it provides a measure of how much variability there is in $f(x)$ and around its mean value $\mathbb{E}[f(x)]$.

$$\text{Var}[f] = \mathbb{E}[f(x)^2] - E[f(x)]^2$$

interms of x,

$$\text{var}[x] = \mathbb{E}[x^2] - (\mathbb{E}[x])^2$$

For two random variables x and y, the covariance is defined by

$$\text{cov}[x, y] = \mathbb{E}_{x,y}[\{x - E[x]\}\{y - \mathbb{E}[y]\}]$$

$$= E_{x,y}[xy] - E[x]E[y]$$

it expresses the extent to which x and y vary together. If independent $\text{cov} = 0$.