Probability Theory

Let X be a random variable with possible values $x_1, ..., x_M$ and Y with $y_1, ..., y_L$. Then we denote the probability of $X = x_i$ by $P(X = x_i)$ which is abbreviated to $P(x_i)$. Then, $P(x|y)$ denotes the probability of x given y is true.

Sum Rule.

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

Product Rule.

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

Corollary.

$$P(X) = \sum_Y P(Y|X)P(X)$$

Definition. X and Y are said to be *independent* if $P(X, Y) = P(X)P(Y)$.

Bayes' Theorem.

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

In the above equation, X is a hypothesis and Y is the evidence we know is true. We want to know what is the probability of the hypothesis given the evidence. Here, $P(X)$ is the *prior*, which tells us the probability of X happening before we observed the evidence, then $P(Y|X)$ is the *likelihood*, the chance of the evidence given the hypothesis. $1/P(Y)$ is the scaling factor to ensure the probability is normalised. Finally, $P(X|Y)$ is called the *posterior*.

Probability Densities

Now we move to continuous probabilities.

Definition. Suppose x is a real-valued random variable. If the probability of x falling in the interval $(x, x + \delta x)$ is $p(x)\delta x$ as δx approaches 0, then $p(x)$ is the *probability density* over x if the two of the following conditions hold:

1. $p(x) \geq 0$ for all $x \in \mathbb{R}$.
2. $\int_{-\infty}^{\infty} p(x) dx = 1$.

Trivially, we have

$$P(x \in (a, b)) = \int_a^b p(x) dx$$

and we define the cumulative distribution function given by

$$P(z) = \int_{-\infty}^z p(x) dx,$$

which satisfies $P'(x) = p(x)$. We can also have multivariable probability densities over $\mathbb{R}^D$, where we just integrate over $\mathbb{R}^D$ instead of $\mathbb{R}$.

Continuous Sum Rule.

$$p(x) = \int p(x, y) dy$$

Product rule is the exact same formula as the discrete version.

Expectation and Covariance

Definition. Let X be a random variable and $f : X \longrightarrow \mathbb{R}$, and let $p(x)$ be a probability distribution over X. Then, the *expectation* of f is given by

$$\mathbb{E}[f] = \sum_{x \in X} p(x) f(x)$$

if X is discrete, and

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

if continuous.

The intuition is to compute the average of f but weighing each value by its probability. Note that if we consider functions to be vectors, then $\mathbb{E}[\cdot]$ is a linear map.

We can also approximate $\mathbb{E}[f] \approx \frac{1}{N} \sum_{n=1}^N f(x_n)$ where $\{x_1, ..., x_N\}$ is N samples drawn from X. If we take the limit as N approaches infinity then the approximation becomes an equivalence.

Definition. The *variance* of f is defined as follows

$$\operatorname{var}[f] = \mathbb{E}\left[(f - \mathbb{E}[f])^2\right]$$

Using the linearity of $\mathbb{E}$ its trivial to show that

$$\mathrm{var}[f] = \mathbb{E}[f^2] - \mathbb{E}[f]^2.$$

Definition. For two random variables x and y, the *covariance* is defined as follows

$$\mathrm{cov}[x, y] = \mathbb{E}_{x,y}[(x - \mathbb{E}[x])(y - \mathbb{E}[y])] = \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y].$$

Notation. $\mathrm{cov}[x] := \mathrm{cov}[x, x]$.