Binary Variables

Definition. If x is a random variable and $x \in \{0, 1\}$ then x is a binary variable.

Definition. Let $P(x = 1) = \mu$. Then the *Bernoulli* distribution is given by

$$\text{Bern}(x|\mu) = \mu^x (1 - \mu)^{1-x}.$$

Let $\mathcal{D} = x_1, ..., x_N$, then

$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} p(x_n|\mu)$$

which forms the likelihood function for $p(\mu|\mathcal{D})$. Then,

$$\ln p(\mathcal{D}|\mu) = (\ln \mu - (\ln(1 - \mu))) \left(\sum_{n=1}^{N} x_n \right) + N \ln(1 - \mu).$$

Since the above function depends on $\mathcal{D}$ only through the sum, $\sum_{n=1}^{N} x_n$ is called a *sufficient statistic*. Maximising the log will result in the sample mean as follows

$$\mu_{\text{ML}} = \frac{1}{N} \sum_{n=1}^{N} x_n.$$

If there is a μ chance of a binary variable being 1, then we can use a binomial distribution to compute the probability of drawing m instances of 1 out of a data set of N elements.

Definition. Given a size N and m, and a mean μ, the *binomial distribution* is defined as follows

$$\text{Bin}(m|N, \mu) = \binom{N}{m} \mu^m (1 - \mu)^{N-m}$$

where

$$\binom{N}{m} = \frac{N!}{m!(N - m)!}.$$

For the binomial distribution, we have the following properties

$$\begin{aligned} \mathbb{E}[m] &= N\mu \\ \text{var}[m] &= N\mu(1 - \mu). \end{aligned}$$

We now write a useful way of generalising the factorial to non-negative reals.

Definition. The *Gamma Distribution* is given by the following

$$\Gamma(x) = \int y^{x-1} e^{-y} \, dy$$

It is easy to show that the Gamma distribution possesses the property $\Gamma(x) = x!$ for any $x \in \mathbb{N}$.

For a Bayesian approach, we need to define a prior for μ. Due to various useful properties, we use the following.

Definition. Given hyper-parameters a and b, the *Beta Distribution* is given by the following.

$$\text{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1}(1-\mu)^{b-1}$$

We define the prior $p(\mu) = \text{Beta}(\mu|a,b)$.

The Beta distribution has the following properties.

$$\int_0^1 \text{Beta}(\mu|a,b) \, d\mu = 1.$$

$$\mathbb{E}[\mu] = \frac{a}{a+b}$$

$$\text{var}[\mu] = \frac{ab}{(a+b)^2(a+b+1)}$$

The posterior can then be deduced by multiplying the beta function with the binomial likelihood, then normalising.

Proposition. Given the parameters m, l, a, b, where $l = N - m$, the binomial likelihood and beta prior, we have

$$p(\mu|m,l,a,b) \propto \mu^{m+a-1}(1-\mu)^{l+b-1}.$$

Proof. Recall Baye's theorem.

$$p(\mu|\mathcal{D}) = \frac{p(\mathcal{D}|\mu)p(\mu)}{p(\mathcal{D})}$$

Hence we have

$$\begin{aligned} p(\mu|m,l,a,b) \propto p(\mathcal{D}|\mu)p(\mu) &= \text{Bin}(m|N,\mu) \, \text{Beta}(\mu|a,b). \\ &= \frac{N!}{(N-m)!m!} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1}(1-\mu)^{b-1} \mu^m (1-\mu)^{N-m} \end{aligned}$$

Since $\frac{N!}{(N-m)!m!} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$ is just a constant, we can delete it. Hence,

$$\begin{aligned} p(\mu|m,l,a,b) &\propto \mu^{a-1}(1-\mu)^{b-1}\mu^m(1-\mu)^l \\ &\propto \mu^{m+a-1}(1-\mu)^{l+b-1}. \end{aligned}$$

■

We can also observe that posterior $p(\mu|m,l,a,b)$, when normalised, is just a Beta distribution, since

$$p(\mu|m,l,a,b) = \frac{\Gamma(m+a+l+b)}{\Gamma(m+a)\Gamma(l+b)}\mu^{m+a-1}(1-\mu)^{l+b+1} = \text{Beta}(\mu|m+a,l+b).$$

Hence the posterior can act like the prior upon observations of more data.

Hence we can sequentially train our model by giving it more and more data, and constantly updating the Beta distribution. We now predict as best as we can, the outcome of the next trial, then we use the rules of probability.

$$p(x=1|\mathcal{D}) = \int_0^1 p(x=1|\mu)p(\mu|\mathcal{D})d\mu = \int_0^1 \mu p(\mu|\mathcal{D})d\mu = \mathbb{E}[\mu|\mathcal{D}].$$

It follows that

$$p(x=1|\mathcal{D}) = \frac{m+a}{m+a+l+b}.$$