Maximum Likelihood Parameter Estimation

Ghazal Farhani

June 27, 2020

1 Max likelihood in the case of Bernoulli

The Likelihood of a sequence $D = x_1, x_2, ..., x_n$ (of for example coin tosses) can be written as:

$$\begin{aligned} P(D|\theta) &= \Pi_{n=1}^{N} \theta_n^x (1-\theta)^{1-x_n} = \theta_1^N (1-\theta)^{N_0} \\ N_1 &= \Sigma_{n=1}^{N} x_n \\ N_0 &= \Sigma_{n=1}^{N} (1 - x_n) \end{aligned} \tag{1}$$

In stead of likelihood we can find log of likelihood. The maximum log likelihood $\hat{\theta} = argmaxP(D|\theta)$ for the Bernoulli's distribution can be calculated as follows:

$$\begin{aligned} l(\theta) = logP(D|\theta) &= log\theta_1^N (1-\theta)^{N_0} \\ &= N_1 log\theta + N_0 log(1-\theta) \end{aligned} \tag{2}$$

Knowing $l(\theta)$, to calculate $\hat{\theta}$ the gradient of $l(\theta)$ with respect to θ should be found:

$$\frac{\partial l(\theta)}{\partial \theta} = \frac{N_1}{\theta} - \frac{N_0}{1-\theta} \tag{3}$$

And $\frac{\partial l(\theta)}{\partial \theta} = 0$ will result in:

$\hat{\theta} = \frac{N_1}{N}$

2 Bayesian Parameter Estimation

Instead of having a deterministic approach to estimate the maximum value for θ a probabilistic option can be considered based on a conditional probability approach:

$$\begin{aligned} P(D|\theta) &= \frac{P(\theta|D)P(\theta)}{P(D)} \\ P(D) &= \int P(D|\theta) d\theta \end{aligned} \tag{4}$$

$P(D)$ is the normalization factor which is independent of θ and is called marginal likelihood evidence, and $P(\theta)$ is the *a priori*.

2.1 Conjugate Prior

A *prior* is called conjugate if the result of multiplication of *prior* and $P(D|\theta)$ be a posterior with the same parametric family of *prior*. If the observations Bernoulli distribution:

$$P(D|\theta) \sim \theta^{N_1}(1-\theta)^{N_0} \tag{5}$$

Then an *a prior* of the form of Beta distribution will be desired:

$$\begin{aligned} Beta(\theta|\alpha_1, \alpha_0) &= \frac{\Gamma(\alpha_0+\alpha_1)}{\Gamma(\alpha_0)\Gamma(\alpha_1)}\theta^{\alpha_1-1}(1-\theta)^{\alpha_0-1} \\ \Gamma(x) &= \int u^{x-1}e^{-u}dx \end{aligned} \tag{6}$$

where $\frac{\Gamma(\alpha_0+\alpha_1)}{\Gamma(\alpha_0)\Gamma(\alpha_1)}$ is the normalized constant of Beta distribution:

$$\frac{1}{Z(\alpha_0, \alpha_1)} = \frac{\Gamma(\alpha_0+\alpha_1)}{\Gamma(\alpha_0)\Gamma(\alpha_1)} \tag{7}$$

Some properties of $Beta(\theta|\alpha_1, \alpha_0)$ is listed below:

$$\begin{aligned} mean &= \frac{\alpha_1}{\alpha_0+\alpha_1} \\ Var &= \frac{\alpha_1\alpha_0}{\alpha_1{\alpha_0}^2(\alpha_1+\alpha_0-1)} \end{aligned} \tag{8}$$

Using Beta *prior* the posterior can be written as follows:

$$P(\theta|D) \sim \theta^{N_1}(1-\theta)^{N_0}\theta^{\alpha_1-1}(1-\theta)^{\alpha_0-1} = \theta^{N_1+\alpha_1-1}(1-\theta)^{N_0+\alpha_0-1} \tag{9}$$

It is clear that the posterior is a Beta distribution as well, thus updating the posterior becomes easy as there is no need to calculate the normalization constant of posterior (save us from calculating the $P(D) = \int P(D|\theta)d\theta$).

Updating the posterior will be as follows:

$$\begin{aligned} P(\theta|D) &= \frac{P(D|\theta)P(\theta|\alpha_0, \alpha_1)}{P(D)} \\ &= \frac{1}{P(D)}\theta^{N_1}(1-\theta)^{N_0}\frac{1}{Z(\alpha_0, \alpha_1)}\theta^{\alpha_1-1}(1-\theta)^{\alpha_0-1} \\ &= Beta(\theta|N_1+\alpha_1, N_0+\alpha_0) \end{aligned} \tag{10}$$

where $\frac{1}{P(D)}\frac{1}{Z(\alpha_0,\alpha_1)}$ is the normalization constant for the posterior:

$$\frac{1}{Z(\alpha_0+N_0, \alpha_1+N_1)} = \frac{1}{P(D)}\frac{1}{Z(\alpha_0, \alpha_1)} \tag{11}$$

We can thus, calculate the marginal likelihood $P(D)$:

$$\begin{aligned} P(D) &= \frac{Z(\alpha_0+N_0, \alpha_1+N_1)}{Z(\alpha_0, \alpha_1)} \\ &= \frac{\Gamma(\alpha_0+N_0)\Gamma(\alpha_1+N_1)}{\Gamma(\alpha_0+\alpha_1+1)}\frac{\Gamma(\alpha_0+\alpha_1)}{\Gamma(\alpha_0)\Gamma(\alpha_1)} \end{aligned} \tag{12}$$

The updates on the posterior can be sequential; the first *prior* can be a Beta distribution where $_0 = \alpha_1 = 1$ (which is a uniform distribution). And α_0 and α_1 will be updated after some observations. For example, if a coin is tossed N times, and N_1 heads and N_0 tails were observed the posterior becomes:

$$\begin{aligned} P(\theta|\alpha_1, \alpha_0, N_1, N_0) &= Beta(\theta; \alpha_1 + N_1, \alpha_0 + N_0) = Beta(\theta; \alpha_1', \alpha_0') \\ \alpha_1' &= \alpha_1 + N_1 \\ \alpha_0' &= \alpha_0 + N_0 \end{aligned} \tag{13}$$

α_1' and α_0' are new *priors* for the next sequence (set of observations).

2.2 Predictive Distribution of Binomials