1. Maximum Likelihood Estimation is a relatively simple method of constructing an estimator for an unknown parameter θ. Maximum likelihood estimation (MLE) can be applied in most problems, it has a strong intuitive appeal, and often yields a reasonable estimator of θ. Furthermore, if the sample is large, the method will yield an excellent estimator of θ. For these reasons, the method of maximum likelihood is probably the most widely used method of estimation in statistics.

Maximum Likelihood Estimation (MLE) is a principle that estimates the parameters of a statistical model, which makes the observed data most probable. In other words, MLE maximizes the data likelihood.

MLE, is a traditional probabilistic approach that can be applied to data belonging to any distribution, i.e., Normal, Poisson, Bernoulli, etc. With prior assumption or knowledge about the data distribution, Maximum Likelihood Estimation helps find the most likely-to-occur distribution parameters. For instance, let us say we have data that is assumed to be normally distributed, but we do not know its mean and standard deviation parameters. Maximum Likelihood Estimation iteratively searches the most likely mean and standard deviation that could have generated the distribution. Moreover, Maximum Likelihood Estimation can be applied to both regression and classification problems.

Therefore, Maximum Likelihood Estimation is simply an optimization algorithm that searches for the most suitable parameters. Since we know the data distribution a priori, the algorithm attempts iteratively to find its pattern. The approach is much generalized, so that it is important to devise a user-defined Python function that solves the particular machine learning problem.

Parameter Estimation: Estimating the Probability of Heads

Let's assume we have a random variable X representing a coin. We can estimate the probability that it will turn up heads ($X = 1$) or tails ($X = 0$).

Task: Estimate the probability of heads $\theta = P(X = 1)$

Evidently, if $P(X = 1) = \theta$, then $P(X = 0) = 1 - \theta$. Since we do not know the "true" probability of heads, i.e. $P(X = 1) = \theta$, we will use $\hat{\theta}$ to refer to its estimate.

Question: What is the probability of $\theta = P(X = 1)$?

In general, Maximum Likelihood Estimation principle asks to choose parameter θ that maximizes $P(Data|\theta)$, or in other words maximizes the probability of the observed data. We assume that θ belongs to the set $\Theta \subset \mathbb{R}^n$.

Therefore,

$$\hat{\theta}_{MLE} = arg\ max P(Data|\theta) \tag{1}$$

In regards to our coin flip example, if we flip the coin repeatedly, we observe that:

1. It turns up heads α_1 times
2. It turns up tails α_0 times

Intuitively, we can estimate the $P(X = 1)$ from our training data (number of tosses) as the fraction of flips that ends up heads:

$$P(X = 1) = \frac{\alpha_1}{\alpha_1 + \alpha_0} \tag{2}$$

For instance, if we flip the coin 40 times, seeing 18 heads and 22 tails, then we can estimate that:

$$\hat{\theta} = P(X = 1) = \frac{18}{40} = 0.45 \tag{3}$$

And if we flip it 5 times, observing 3 heads and 2 tails, then we have:

$$\hat{\theta} = P(X = 1) = \frac{3}{5} = 0.6 \tag{4}$$

2. Since the sample is (3, 0, 2, 1, 3, 2, 1, 0, 2, 1), the likelihood is

$$L(\theta) = P(X = 3)P(X = 0)P(X = 2)P(X = 1)P(X = 3)P(X = 2)P(X = 1)P(X = 0)P(X = 2)P(X = 1) \tag{5}$$

Substituting from the probability distribution given above, we have

$$L(\theta) = \prod_{i=1}^{n} P(X_i|\theta) = \Big(\frac{2\theta}{3}\Big)^2 + \Big(\frac{\theta}{3}\Big)^3 + \Big(\frac{2(1-\theta)}{3}\Big)^3 + \Big(\frac{1-\theta}{3}\Big)^2$$

Clearly, the likelihood function L(θ) is not easy to maximize.

Let us look at the log likelihood function

$$l(\theta) = logL(\theta) = \sum_{i=1}^{n} logP(X_i|\theta) \tag{6}$$

$$= 2\Big(log\frac{2}{3} + log\theta\Big) + 3\Big(log\frac{1}{3} + log\theta\Big) + 3\Big(log\frac{2}{3} + log(1-\theta)\Big) + 2\Big(log\frac{1}{3} + log(1-\theta)\Big) \tag{7}$$

$$= C + 5\ log(\theta) + 5\ log(1-\theta) \tag{8}$$

where C is a constant which does not depend on θ. It can be seen that the log likelihood function is easier to maximize compared to the likelihood function.

Let the derivative of $l(\theta)$ with respect to θ be zero:

$$\frac{dl(\theta)}{d\theta} = \frac{5}{\theta} - \frac{5}{1-\theta} = 0 \tag{9}$$

and the solution gives us the MLE, which is $\hat{\theta} = 0.5$

3. Before actually solving the problem, let's establish some notation and terms.We can think of counting the number of heads in 100 tosses as an experiment. For a given value of p, the probability of getting 55 heads in this experiment is the binomial probability.

$$P(55\ heads) = \binom{100}{55} p^{55}(1-p)^{45} \tag{10}$$

The probability of getting 55 heads depends on the value of p, so let's include p in by using the notation of conditional probability:

$$P(55\ heads|p) = \binom{100}{55} p^{55}(1-p)^{45} \tag{11}$$

You should read $P(55\ heads|p)$ as: "the probability of 55 heads given p", or more precisely as "the probability of 55 heads given that the probability of heads on a single toss is p." Here are some standard terms we will use as we do statistics.

(a) Experiment: Flip the coin 100 times and count the number of heads.

(b) Data: The data is the result of the experiment. In this case it is '55 heads'.

(c) Parameter(s) of interest: We are interested in the value of the unknown parameter p.

(c) Likelihood, or likelihood function: this is $P(data|p)$. Note it is a function of both the data and the parameter p.

In this case the likelihood is

$$P(55\ heads|p) = \binom{100}{55} p^{55}(1-p)^{45} \tag{12}$$

Notes:

1. The likelihood $P(data|p)$ changes as the parameter of interest p changes.

2. Look carefully at the definition. One typical source of confusion is to mistake the likelihood $P(data|p)$ for $P(p|data)$. We know from our earlier work with Bayes' theorem that $P(data|p)$ and $P(p|data)$ are usually very different.

Definition: Given data the maximum likelihood estimate (MLE)for the parameter p is the value of p that maximizes the likelihood $P(data|p)$. That is, the MLE is the value of p for which the data is most likely.

answer: For the problem at hand, we saw above that the likelihood

$$P(55\ heads \mid p) = \binom{100}{55} p^{55}(1-p)^{45} \tag{13}$$

We'll use the notation $\hat{p}$ for the MLE. We use calculus to find it by taking the derivative of the likelihood function and setting it to 0.

$$\frac{d\ P(data \mid p)}{dp} = \binom{100}{55}(55p^{54}(1-p)^{45} - 45p^{55}(1-p)^{44}) = 0 \tag{14}$$

Solving this for p we get

$$55p^{54}(1-p)^{45} = 45p^{55}(1-p)^{44} \tag{15}$$

$$55(1-p) = 45p \tag{16}$$

$$55 = 100p \tag{17}$$

$$the\ MLE\ is\ \hat{p} = 0.55 \tag{18}$$

Note:

1. The MLE for p turned out to be exactly the fraction of heads we saw in our data.

2. The MLE is computed from the data. That is, it is a statistic.

3. Officially you should check that the critical point is indeed a maximum. You can do this with the second derivative test.

Log likelihood:

If is often easier to work with the natural log of the likelihood function. For short this is simply called the log likelihood. Since $ln(x)$ is an increasing function, the maxima of the likelihood and log likelihood coincide.

Example 2. Redo the previous example using log likelihood.

answer: We had the likelihood

$$P(55\ heads \mid p)) = \binom{100}{55} p^{55}(1-p)^{45} \tag{19}$$

Therefore the log likelihood is

$$ln(P(55heads \mid p)) = ln\binom{100}{55} + 55\ ln(p) + 45\ ln(1-p) \tag{20}$$

Maximizing likelihood is the same as maximizing log likelihood. We check that calculus gives us the same answer as before:

$$\frac{d(log\ likelihood)}{dp} = \frac{d}{dp}\Big[ln\binom{100}{55} + 55\ ln(p) + 45\ ln(1-p)\Big] = 0 \tag{21}$$

$$\frac{55}{p} - \frac{45}{1-p} = 0 \tag{22}$$

$$\Rightarrow \hat{p} = 0.55 \tag{23}$$

4. For continuous distributions, we use the probability density function to define the likelihood.We show this in a few examples. In the next section we explain how this is analogous to what we did in the discrete case.

We need to be careful with our notation. With five different values it is best to use subscripts. Let X_i be the lifetime of the i^{th} bulb and let x_i be the value X_i takes. Then each X_i has PDF $f_{X_i}(x_i) = \lambda e^{-\lambda x_i}$. We assume the lifetimes of the bulbs are independent,so the joint PDF is the product of the individual densities:

$$f(x_1, x_2, x_3, x_4, x_5)|\lambda) = (\lambda e^{-\lambda x_1})(\lambda e^{-\lambda x_2})(\lambda e^{-\lambda x_3})(\lambda e^{-\lambda x_4})(\lambda e^{-\lambda x_5}) = \lambda^5 e^{-\lambda(x_1+x_2+x_3+x_4+x_5)} \tag{24}$$

Note that we write this as a conditional density, since it depends on λ. Viewing the data as fixed and λ as variable, this density is the likelihood function. Our data had values

$$x_1 = 2,\ x_2 = 3,\ x_3 = 1,\ x_4 = 3,\ x_5 = 4 \tag{25}$$

So the likelihood and log likelihood functions with this data are:

$$f(2,3,1,3,4)|\lambda) = \lambda^5 e^{-13\lambda} = ln(f(2,3,1,3,4)|\lambda)) = 5\ ln(\lambda) - 13\lambda \tag{26}$$

Finally we use calculus to find the MLE:

$$\frac{d(log\ likelihood)}{d\lambda} = \frac{f'(2,3,1,3,4)|\lambda)}{f(2,3,1,3,4)|\lambda)} = \frac{5}{\lambda} - 13 = 0 \Rightarrow \hat{\lambda} = \frac{5}{13} \tag{27}$$

5. Solution: Marginal PMFs are calculated as follows:

$p_X(1) = \Sigma_y\ p_{XY}(1, y) = \frac{1}{2}$

Similarly, $p_X(3) = \frac{1}{2}$
$p_Y(2) = \Sigma_x\ p_{XY}(x, 2) = \frac{1}{2}$

$p_Y(4) = \frac{1}{2}$

$Expectation$:

$E[X] = \frac{1}{2} \times 1 + \frac{1}{2} \times 3 = 2$

$E[Y] = \frac{1}{2} \times 2 + \frac{1}{2} \times 4 = 3$

$E[X^2] = \frac{1}{2} \times 1^2 + \frac{1}{2} \times 3^2 = 5$

$E[Y^2] = \frac{1}{2} \times 2^2 + \frac{1}{2} \times 4^2 = 10$

$Variance$:

$Var[X] = E[X^2] - (E[X])^2 = 5 - 2^2 = 5 - 4 = 1$

$Var[Y] = E[Y^2] - (E[Y])^2 = 10 - 3^2 = 10 - 9 = 1$

$Co - variance$**:**

$$Cov[X, Y] = E[\{X - E(X)\}\{Y - E(Y)\}]$$

$$= E[X \cdot \{Y - E[Y]\} - E[X] \cdot \{Y - E[Y]\}]$$

$$= E[\{X \cdot Y - X \cdot E(X)\}\{Y - E(Y)\}]$$

$$= E[X \cdot Y - X \cdot E[Y] - E[X] \cdot Y + E[X] \cdot [Y]$$

$$= E[X \cdot Y - E[X] \cdot E(Y) - E[X] \cdot E[Y] + E[X] \cdot [Y]$$

$$= E[X \cdot Y] - E[X] \cdot E(Y) - E[X] \cdot E[Y] + E[X] \cdot [Y]$$

$$= E[X \cdot Y] - E[X] \cdot E[Y]$$

Now we calculate $E[X \cdot Y]$ as follows:

$$E[X \cdot Y] = \Sigma_x \Sigma_y xy f_{XY}(x, y) = (1 \times 2) \times \frac{1}{2} + (3 \times 4) \times \frac{1}{2} = 7$$

$$Cov[X, Y] = 7 - 2 \times 3 = 7 - 6 = 1$$

or rearranging, we get

$$Cov[X, Y] = \frac{1}{2}(1 \times 2 - 2 \times 3) + \frac{1}{2}(3 \times 4 - 2 \times 3) = 1$$

So: $\Sigma(X, Y) = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$