Maximum Likelihood Estimation

The Logistic Regression is built on **Maximum Likelihood Estimation**

Probability vs Likelihood

- In probability we know the *parameters* which exactly describe the situation and how often things occur (Something must happen), they add up to **one** everything happen on the same universe
  For a single value of parameter something must happen
- Likelihoods are Probability of the **Observed data** under a hypothetical scenario.
- There are many likelihoods that do not add up to one and thus cannot be interpreted as probabilities
- **Likelihoods** depends on the parameter
  Likelihood of the parameter θ Fitting the given **Data** in other words...
  Given this **Observed data** what parameter θ make it probable, explains the data **Observed**
  - The **Maximum Likelihood** chooses the universe where are data would be most likely

Example :

- Flipping a coin one time and Observing **heads**

Consider these **universes**

1. **Probability** of heads is very small
2. **Probability** of heads or tails is fair
3. **Probability** of heads is 100%
   The **Maximum Likelihood** is the Probability of a coin that always lands on **heads** we maximize the chances of landing hands

- The probability of the data we observed is maximized
  But most of the time we add restrictions and study the probability of that happening with unknown parameters
- Probability of heads is p
- Probability of tails is $1-p$
- The goal is to estimate what p that maximizes the likelihood
- This solved using **Derivations** and studying where the graph maximize
- in a **Normal distribution** Setting the likelihood is maximize when we choose the **mean** parameter that best fit the data. The **MLE** of the mean of a normal distribution is the **Sample mean**. Searching for the value of the mean that makes the observed data most probable under the normal distribution
- If we have two observed probabilities the **MLE** try to maximize both of the probabilities

Machine Learning Example :

$$L(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p) = Value$$

- Where the $Value$ tell us how well the assumed *Estimated* **Coefficients** fits the data
- The *Best Fit* is the **Maximized Likelihood Estimates**

$$\hat{\beta}^{\text{MLE}} = \arg\max_{\beta} = L(\hat{\beta}_0, \hat{\beta}_1, \ldots \hat{\beta}_p)$$

$$L(\beta) \propto P(X_1, X_2, .., X_n|\beta)$$

- The **MLE** is the **Joint Probability** of X_i Given the estimated β

$$L(\beta) \propto \prod_{i=1}^{n} P(X_i|\beta)$$

- Calculating The product of all the Joint Probability will result in a very small value that machines will miss **Arithmetic Underflow**

$$\log(L(\beta)) \propto \log\left(\prod_{i=1}^{n} P(X_i|\beta)\right)$$

$$\log(L(\beta)) \propto \sum_{i=1}^{n} \log(P(X_i|\beta))$$

$$\hat{\beta}^{\text{MLE}} = \arg\max_{\beta} \sum_{i=1}^{n} \log(P(X_i|\beta)$$

- And This expression is used to derive the **Least Squares** methodResidual Sum of Squares, Sigmoid Function and other **Parametric approach** algorithms.