Probability Vs Likelihood

Some Example:

Example 1: Coin Toss

(1) → H
↳ T

→ fair
↳ $P(H) = 0.5$

→ mathematically

Bernauli → 1 parameter
Distribution ↳ (p)

Tail = $P(t) = 0.5$

$p = P(H) = 0.5$

$q = 1 - p = 0.5$ ← Tail

$$P(x) = p^k + (1-p)(1-k)$$

$k = 0, 1$

* for Head $k = 1$

$P(x) = p$ ↳ Probability of Head

$k=0$ $P(x) = (1-p)$

* If we know distribution then we also know parameter and also know event and chance then probability

Likelihood

5 times coin flip

HHHHH → what is plasibility that $P(H) = 0.5$ and coin is fair.

↑ we get 5 Head

So, what is the chances that this is fair coin

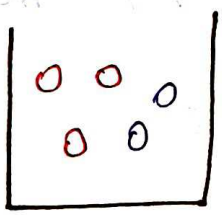
H H H H H ⇒ $L = (0.5)^5$

↑ ↑ ↑ ↑ ↑ ↳ very small

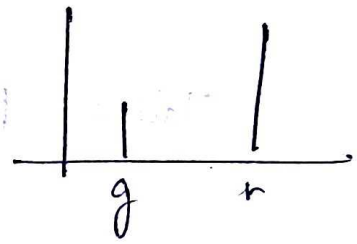
or or or or or

* The likelihood that the coin is fair given 5 tims Head is ~~coup~~ coin is fair.

Example → 2



$P(G) = \frac{2}{5}$



1 trial → red → Berauli

↘ green

ball is red → $\frac{3}{5}$ $pk + (1-p)(1-k)$

$k=0$

$(1-p) = 1 - \frac{2}{5} = \frac{3}{5}$

Likelihood

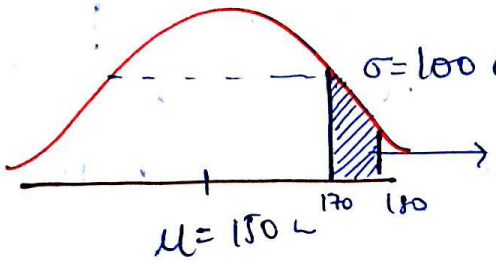
5 balls ↔ [G G G G G] → $\left(\frac{2}{5}\right)^5$ → small

Event

Example 3 continous

↳ Normal Distribution

dist → (μ, σ)



Find area to get probability of getting weight between 170 - 180 using μ, σ.

Likelihood

100 cm → N(150, 10) (μ, σ)

↑ event

$$L(\mu,\sigma \mid X=100) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{\frac{-(X-\mu)^2}{2\sigma^2}}$$ (X = 100) $\mu=150$, $\sigma=10$

$$= 1.4\times10^{-7}$$

↳ very small

parameter → event (prob)

event → parameter (likelihood)

Defination:

Probability: This is a measure of the chance that certain event will occur out of the all possible events. It's usually presented as a ratio or fraction, and it range from 0 (means the event will not happen) to 1 (means the event is certain to happen).

Likelihood: In statistical context, likelihood a function that measures the plausibility of a particular parameter value given some observed data. It quantifies how well a specific outcome support specific parameter values.

Maximum Likelihood Estimation (MLE)

coin toss

$p = 0.5$

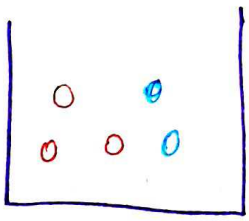
↓

HHHHH → $(0.5)^5 = L(p = 0.5 \mid HHHHH)$

if $p = 0.6 \rightarrow (0.6)^5 = L(p = 0.6 \mid HHHHH)$

which one better $(0.5)^5$ or $(0.6)^5$

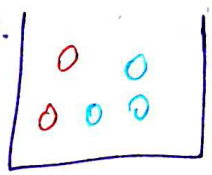
$(0.5)^5 < (0.6)^5$



$P(B) = \frac{2}{5}$

↳ BBBBB

$L = (p = 2/5 \mid BBBBB) = \left(\frac{2}{5}\right)^5$



$P(B) = \frac{3}{5}$

$L = (p = 2/5 \mid BBBBB) = \left(\frac{3}{5}\right)^5$

which one better

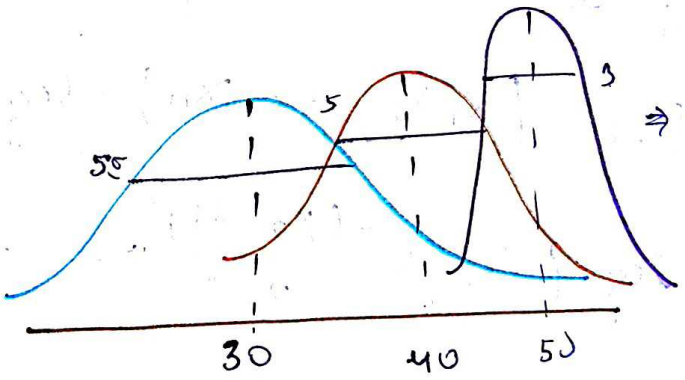
MLE for Normal Distribution

dataset = 100 or n

61
32
45
63
.
.
.
89

→ observation

* We know n data points follow normal Distribution but which μ, σ value follow normal distribution



⇒ which one is highest peak value.

$L(\mu, \sigma | x_1, x_2, x_3 \dots x_n)$ $\mu = 110$, $\sigma = 20$

$L(\mu = 100, \sigma = 10 | x_1 = 61, x_2 = 32, \dots x_n = 89) = 0.89$

$L(\mu = 110, \sigma 20 | x_1 = 61, x_2 = 32, \dots x_n = 89) = 0.79$

* Maximum likelihood is good 0.89

Let assume

we have only x_1 data

$$L(\mu, \sigma | x_1) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_1 - \mu)^2}{2\sigma^2}}$$

If we have x_n independent (dataset and x_1 not depend on x_2)

$$L(\mu, \sigma \mid x_1, x_2, x_3 \cdots x_n) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_1-\mu)^2}{2\sigma^2}} \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_2-\mu)^2}{2\sigma^2}}$$

$$\times \cdots\cdots \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_n-\mu)^2}{2\sigma^2}}$$

* log both side (log likelihood)

$$\log(L(\mu, \sigma \mid x_1, x_2 \cdots x_n) = \boxed{\log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_1-\mu)}{2\sigma^2}}\right)} + \text{①}$$

$$\log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_2-\mu)^2}{2\sigma^2}}\right) + \cdots + \log\left(\frac{1}{\sqrt{2\pi\sigma}} e^{\frac{(x_n-\mu)^2}{2\sigma^2}}\right)$$

* focus on ① then apply all of them

$$\log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_1-\mu)^2}{2\sigma^2}}\right) = \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \log e^{-\frac{(x_1-\mu)^2}{2\sigma^2}}$$

$$= \log(2\pi\sigma^2)^{-1/2} - \frac{(x_1-\mu)^2}{2\sigma^2}$$

$$= -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(x_1-\mu)^2}{2\sigma^2}$$

$$= -\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2 - \frac{(x_1-\mu)^2}{2\sigma^2}$$

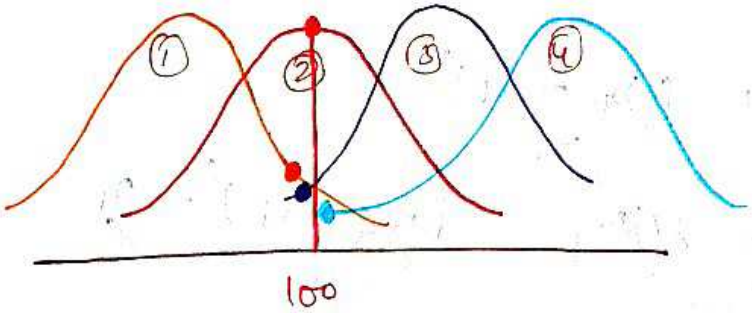
$$= -\frac{1}{2}\log 2\pi - \frac{2}{2}\log\sigma - \frac{(x_1-\mu)^2}{2\sigma}$$

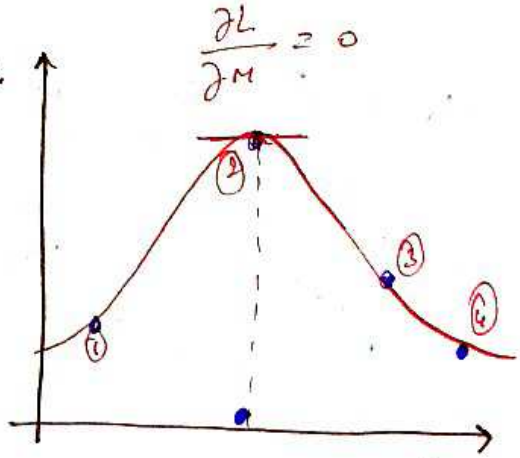
$$\log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_1-\mu)^2}{2\sigma^2}}\right) = -\frac{1}{2}\log 2\pi \log\frac{(x_1-\mu)^2}{2\sigma^2}$$

* Now apply for all eqn

$$= -\frac{1}{2}\log 2\pi - \log\sigma - \frac{(x_1-\mu)^2}{2\sigma^2} - \frac{1}{2}\log 2\pi - \log\sigma - \frac{(x_2-\mu)^2}{2\sigma^2} - \cdots - \frac{1}{2}\log 2\pi - \log\sigma - \frac{(x_n-\mu)^2}{2\sigma^2}$$

$$\log(L) = \frac{n}{2}\log 2\pi - n\log\sigma - \frac{(x_1-\mu)^2}{2\sigma^2} - \frac{(x_2-\mu)^2}{2\sigma^2} \cdots \frac{(x_n-\mu)^2}{2\sigma} \quad - (1)$$





eqn (1) differentiate w.r.t μ

$$= 0 - 0 + \frac{2(x_1-\mu)}{2\sigma^2} + \frac{(x_2-\mu)}{\sigma^2} + \cdots + \frac{x_n-\mu}{\sigma^2} = 0$$

$$(x_1 - \mu) + (x_2 - \mu) + \cdots + (x_n - \mu) = 0$$

$$x_1 + x_2 + x_3 + \cdots + x_n - n\mu = 0$$

$$n\mu = x_1 + x_2 + \cdots + x_n$$

$$\boxed{\mu = \frac{x_1 + x_2 + \cdots + x_n}{n}}$$

eqn ① is differentiate by σ

$$\frac{\partial \log(L)}{\partial \sigma} = 0 - \frac{n}{\sigma} + \frac{(x_1 - \mu)^2}{\sigma^3} + \frac{(x_2 - \mu)^2}{\sigma^3} + \cdots + \frac{(x_n - \mu)^2}{\sigma^3}$$

$$\frac{(x_1 - \mu)^2}{\sigma^3} + \frac{(x_2 - \mu)^2}{\sigma^3} + \cdots + \frac{(x_n - \mu)^2}{\sigma^3} = \frac{n}{\sigma}$$

$$\frac{(x_1 - \mu)^2}{\sigma^2} + \frac{(x_2 - \mu)^2}{\sigma^2} + \cdots + \frac{(x_n - \mu)^2}{\sigma^2} = n$$

$$\boxed{\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}} \Rightarrow \text{Variance}$$

$$\boxed{\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}} = \text{Std}}$$

MLE in Machine Learning

X_1	X_2	X_3 ...	X_n	(Y) → output → distribution.
:	:	:	:	Y
				N

Y, N → Bernauli

1.) Find out the distribution of Y/X

2.) decide to apply a ML model parametric in nature
   ↳ Logistic Regression

3.) You randomly some value → $\beta_0\ \beta_1\ \beta_2 \dots \beta_n$

4.) Select a likelihood function.
   ex:- pmf = $p^k + (1-p)(1-k)$

(5) Value of $\beta_0\ \beta_1\ \beta_2\ \beta \dots \beta_n$
   ↳ maximum value.

MLE in Logistic Regression

X_1	X_2	X_3	----	X_n	Y
					1
					0
					1
					0

n input cols

Y → bernauli distribution

1 query → 1 or 0 → bernauli distribution

100 querys 1 or 0 → Binomial distribution

1.) Assume a distribution for y (Bernauli distribution)

2.) Logistic Regression → parameter → β_0 β_1 β_2 - - β_n

3.) Random Values → β_0, β_1 - - - - β_n

* assume all y's are independent

Y/x → parameter in β

$$L(Y/X;\beta) = Pk + (1-P)(1-k) \times P$$

$\therefore k = Y$

$P = P(1)$

$1-P = P(0)$

$$L[Y/X;\beta] = P_1 Y_1 + (1-P_1)(1-Y_1)] \times P$$

for all rows of Y

$$L(Y/X;\beta) = P_1 Y_1 + (1-P_1)(1-Y_1) \times P_2 Y_2 + (1-P_2 \times 1-Y) \times$$

$$\text{------} P_n (Y_n)' + (1-P_n)(1-Y_n)$$

for easy mathematics

$$pk + (1-p)(1-k) \xrightarrow{\text{Change}} p^y (1-p)^{(1-y)}$$

↳ output is same

$$L(Y/X;\beta) = p_1^{y_1} (1-p_1)^{(1-y_1)} \times p_2^{y_2} (1-p_2)^{(1-y_2)} \cdots \times p_n^{y_n} (1-p_n)^{(1-y_n)}$$

Compact formula

$$L(Y/X;\beta) = \prod_{i=1}^{n} p_i^{y_i} (1-p_i)^{(1-y_i)}$$

log both side

$$\log(L) = \sum_{i=1}^{n} \log \left(p_i^{y_i} (1-p_i)^{(1-y_i)} \right) = \sum_{i=1}^{n} \log p_i^{y_i} + \log (1-p_i)^{(1-y_i)}$$

(a) (b)

$$\boxed{\log(L) = \sum_{i=1}^{n} y_i \log p_i + (1-y_i) \log (1-p_i)}$$

↳ maximum → Find → β

↳ Gradient Descent

Some Important Question

1. is MLE a general concept applicable to all machine learning algorithms.

Maximum Likelihood Estimation (MLE) is a general statistical concept that can be applied to many machine learning algorithms, particular those that are parametric, but it's not applicable to all machine learning algorithms.

MLE is commonly used in algo such as linear regression, logistic regression and neural network among others. These algorithms use MLE to find the optimal value of parameter that best fit the training data.

However, there are some machine learning algo that don't rely on MLE, for example:

(1) Non-parametric methods
(2) Unsupervised learning algo
(3) Reinforcement learning.

2. How is MLE related to the concept of loss function?

In machine learning, a loss function measures how well a model's predictions allign with the actual values. The goal of training a machine learning model is often to find the model parameter that minimize the loss function.

Maximum likelihood Estimation (MLE) is a method of estimation the parameter of a statistical model to maximize the likelihood function, which is conceptually similar to minimize a loss function, which is conceptualy similar to minimize a loss function. In fact, for many common models, minimizing the loss function is equivalent to maximizing the likelihood function.

MLE and the concept of loss function in machine learning are closely related. Many common loss function can be derived from the principle of maximum likelihood estimation under certain assumption about the data or the model. By minimizing these loss functions, we're effectively performing maximum likelihood estimation.

3. Then why does loss function exist, why don't we maximize likelihood.

The confusion arises from the fact that we're using two different perspective to look at the same problem.

In many machine learning algorithms, the aim is to minimize the different between the predicted and actual values, and this is typically represented by a loss function. When we talk about minimizing the loss function, it's essentially the same as saying we're trying to find the best model parameter that gives us the closest prediction to the actual values.

On the other hand, when we looked at the problem from statistical prespective, we talk in terms ~~the same as maximizing the likelihood~~ function. ~~In fact, many common loss function~~ can be ~~derived from the principle of MLE~~ under ~~certain assumption about data~~. Of maximizing the likelihood of seeing the observed data given the model parameter. This is represented by a likelihood function.

For many models, these two perspectives are equivalent — minimizing the loss function is the same as maximizing the likelihood function. In fact, many common loss function can be derived from the principle of MLE under certain assumption about the data.

So why do we often talk about minimizing the loss function instead of maximizing the likelihood? There are a few reasons:

1. Computational Reasons: It's often easier and more computationally efficient to minimize a loss function than to maximize a likelihood function. This is particularly true when working with complex models like neural networks.

2. Generalization: The concept of a loss function is more general and can be applied to a wider range of problems. Not all machine learning problems can be framed in terms of maximizing a likelihood. For example, many non-parametric method and unsupervised learning algo don't involve likelihoods.

3. Flexibility: Loss functions can be easily customized to the specific needs of a problem. For instance, we might want to give more weight to certain types of errors, or we might want to use a loss function that is robust to outliers.

4. Then why study about maximum likelihood at all?

The study of Maximum Likelihood Estimation (MLE) is essential for several reasons, despite the prevalence of loss functions in machine learning.

1. Statistical foundation: MLE provides a strong statistical foundation for understanding machine learning models. It gives a principled way of deriving the loss function used in many common machine learning algorithms, and it helps us understand why these loss function work and under what assumption.

2. Interpretability: The MLE framework gives us a way to interpret our model parameter. The MLEs are the parameter that make the observed data most likely under our model, which can be a powerful way of understanding what our model has learned.

3. Model Comparison: MLE gives us a way to compare different models on the same dataset. This can be done using tool like the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC), which are based on the likelihood function and can help us choose the best model for our data.

4. Generalization to Other Methods: MLE is a specific case of more general methods, like Expectation Maximization and Bayesian Inference which are used in more complex statistical modelling. Understanding MLE can provide a stepping stone to these more advanced topics.

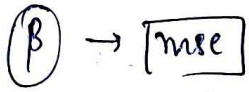
5. Deeper Understanding: Lastly, understanding MLE can give us a deeper understanding of our models, leading to better intuition, better model selection and ultimately, better performance our machine learning tasks

Task → ① Likelihood function of softmax reg

② MLE to linear Regression

normal distribution

by hint



Assumptions Logistic Regression

1. Binary Logistic Regression requires the dependent variable to be binary: That means the outcome variable must have two possible outcomes, such as "yes" vs "no", "success" vs "failure", "span" vs "not span", etc.

2. Independence of observations: The observations should be independent of each other. In other words, the outcome of one instances should not affect the outcome of another.

3. Linearity of Independent variable and log odds:

Although Logistic Regression does not require the dependent and independent variables to be related linearly, it requires that the independent variables are linearly ~~and~~ related to the log odds.

4. Absence of multicollinearity

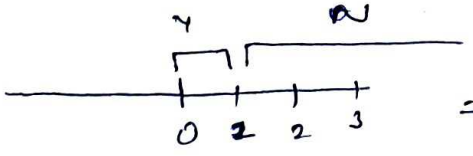
5. Large sample size.

Odds and log(odds)

Odds: The odds of an event is the ratio of the probability of the event happening (p) to the probability of the event not happening (1-p). It's a way of expressing the likelihood of an event - If the odds are greater that 1, the event is more likely to happen than not and vice-versa

→ 3 odds → $P(3) = \frac{1}{6}$ and $P(!3) = \left(\frac{5}{6}\right)$

$$\text{odds} = \frac{1/6}{5/6} = \frac{1}{5}$$



= 2 yes ∞ No

not symmetric

Hence log for symmetric

Another Interpretation of Logistic Regression

$\log(\text{odds}) = \log\left(\frac{p}{1-p}\right)$

$\beta_0 = 0 \quad \beta_1 = 1 \quad \beta_2 = 2$

cgpa	iq	placed	$\hat{y}$
8	80	1	0.93
-	-	0	0.37

$P(y=1) = p$

$\hat{y} = P(y=1) = p = \frac{1}{1+e^{-\beta x}}$

$\beta_0 + \beta_1 x_1 + \beta_2 x_2$

$\downarrow \quad \downarrow \quad \downarrow$

$0 \quad 1 \times 8 \quad + \; 2 \times 80$

$p = \frac{1}{1+e^{-\beta x}} \Rightarrow 1 + e^{-\beta x} = \frac{1}{p}$

$e^{-\beta x} = \frac{1}{p} - 1 \Rightarrow e^{-\beta x} = \frac{1-p}{p}$

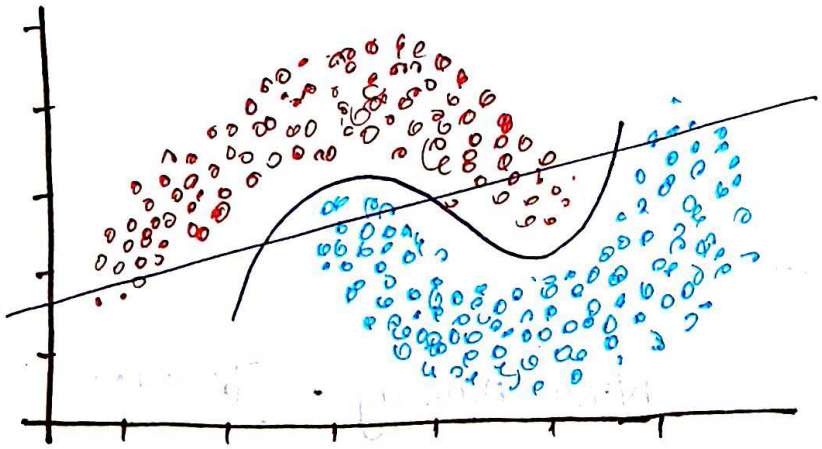
$\frac{1}{e^{\beta x}} = \frac{1-p}{p} \Rightarrow \frac{p}{1-p} = e^{\beta x}$

$\log\left(\frac{p}{1-p}\right) = \log\left(e^{\beta x}\right)$

$\log\left(\frac{p}{1-p}\right) = e^{\beta x}$

$\log\left(\frac{p}{1-p}\right) = \beta x$

Polynomial Features



Polynomial feature

$x \rightarrow$ degree $= 2$

x^0	x^1	x^2
↑	↑	↑
β_1	β_2	β_3

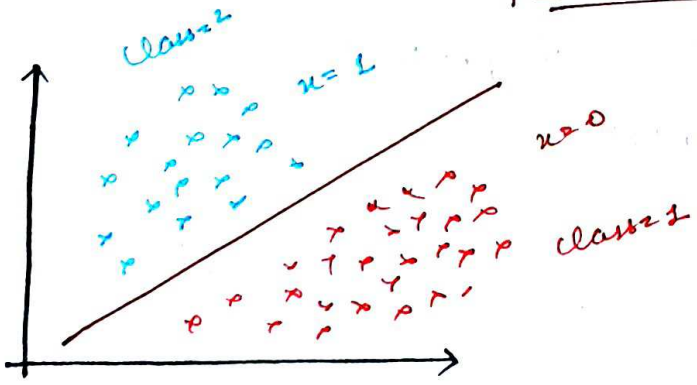
code

β_0

degree ↑↑ → Overfitting

degree ↓↓ → underfitting

Performance Metrics



Dataset

Actual	Predicted
x_1	x_2	y	$\hat{y}$
-	-	0	1
-	-	1	1
-	-	1	1
-	-	0	0
-	-	1	1
-	-	0	1
-	-	1	0