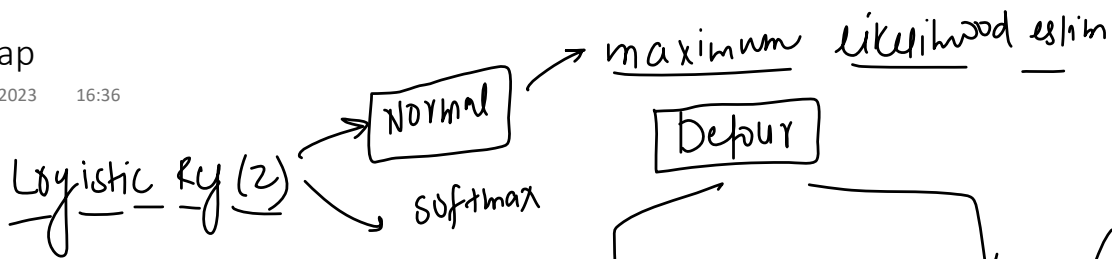


Recap

03 July 2023 16:36



- 1) Prob vs like
- 2) maximum like

General topic

Important

Interview

Log

linear

Neural networks

→ confusing

→ mathematical

hard topic

intuition

→ overview

Plan of attack

MLE → Likelihood

→ MLE

MLE + ML

log reg

Why

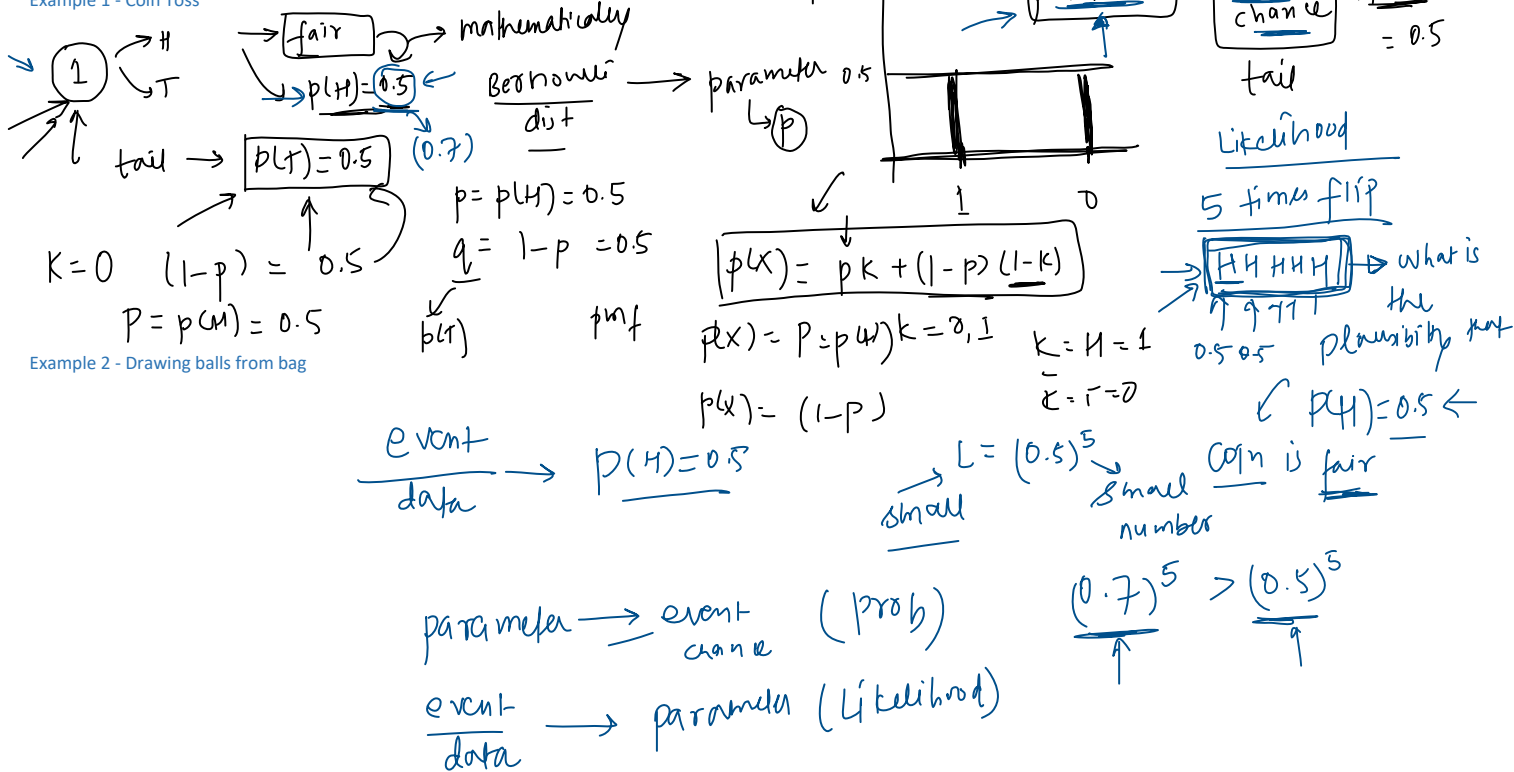
faq

Some Examples

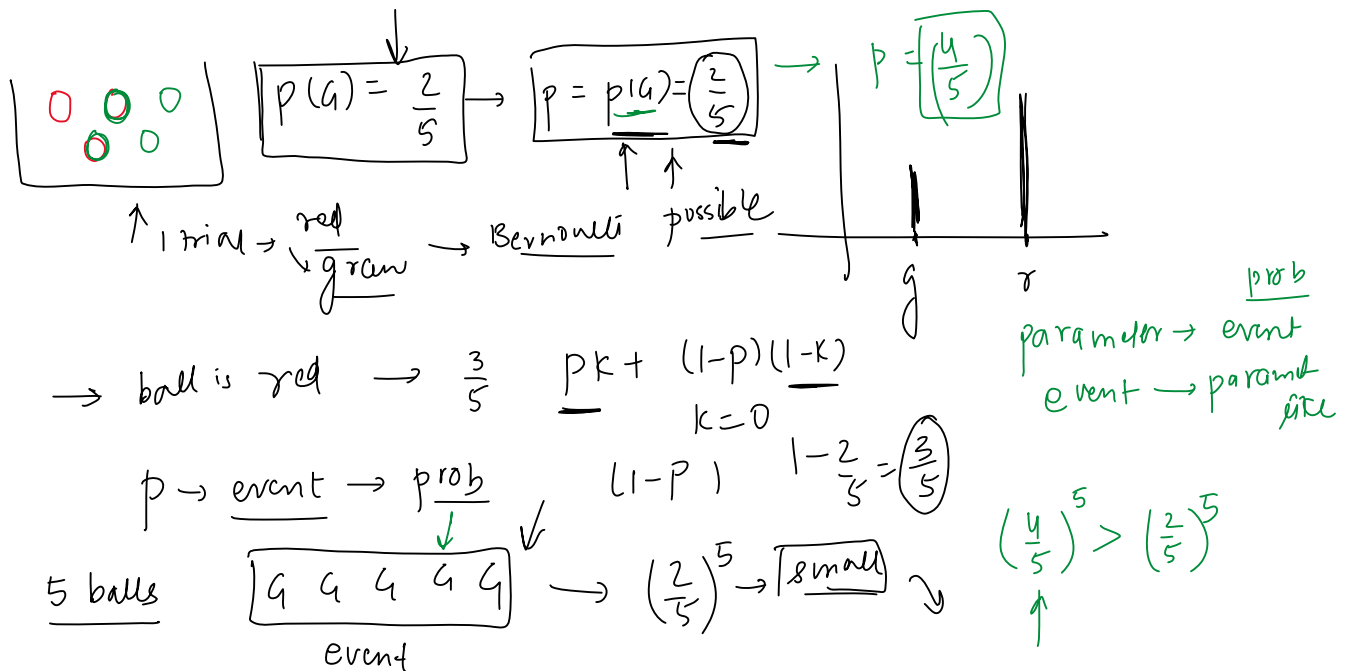
01 July 2023 07:47

random \rightarrow bernoulli \rightarrow (P) \rightarrow Prob. distribution

Example 1 - Coin Toss



Example 2 - Drawing balls from bag



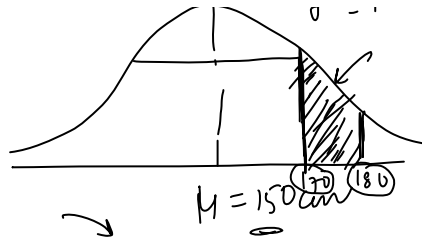
3) Discrete \rightarrow bernoulli
 \rightarrow continuous



↳ continuous

Normal distribution

heights



dist $\rightarrow (\mu, \sigma) \rightarrow$

$$p(130 < X < 180) = 0.02$$



130

very small

$$L(\mu, \sigma | X=100) =$$

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

= like

$$1.47 \times 10^{-7}$$

parameter \rightarrow event (prob)

event \rightarrow parameter (likelihood)

data _____

Probability Vs Likelihood

01 July 2023 09:37

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

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

More Definitions

A **probability** quantifies how often you observe a certain outcome of a test, given a certain understanding of the underlying data.

A **likelihood** quantifies how good one's model is, given a set of data that's been observed.

Probabilities describe test outcomes, while likelihoods describe models.

01 July 2023 10:08

$p = 0.5$
 $p = 0.6$

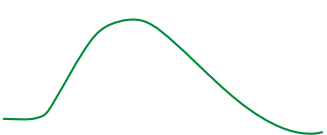
Age

| |
|----|
| 35 |
| 71 |
| 82 |

→ like max/min normal

→

↳ @ Age 71 → Sharp



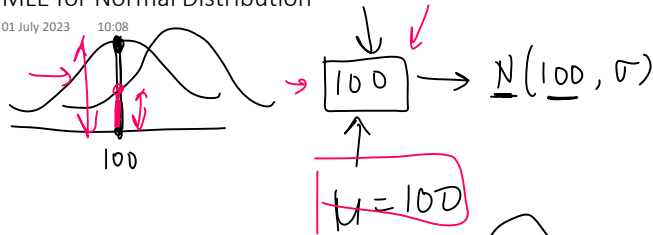
2nd

maximum like

$\mu = 100$

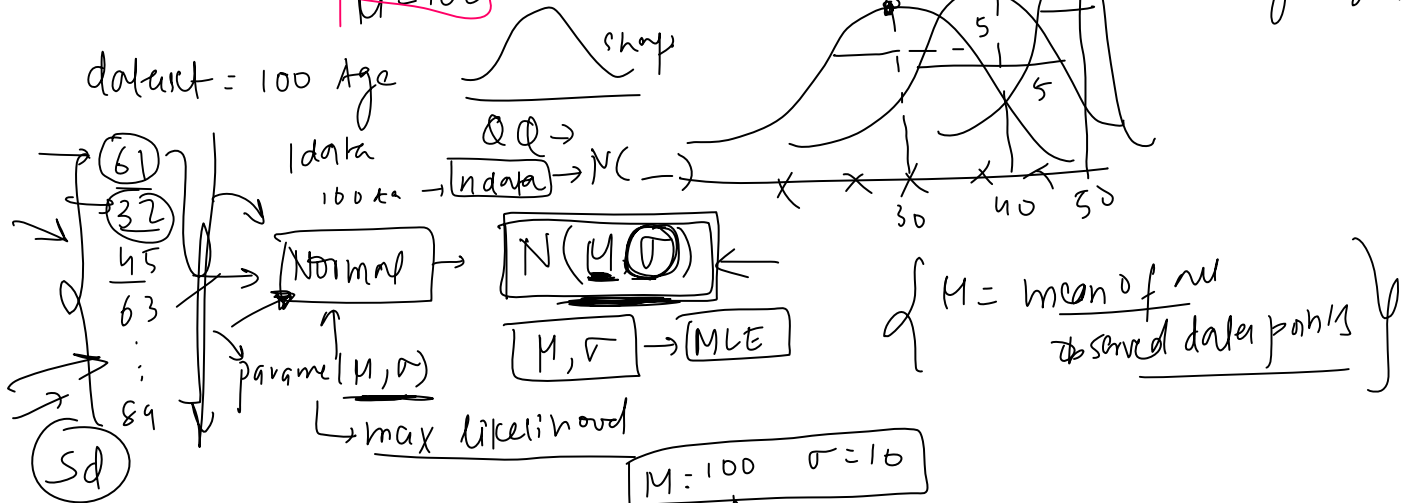
MLE for Normal Distribution

01 July 2023 10:08



max
a, b, c

log a, log b, log c



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

$$\mu = 100, \sigma = 20$$

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

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

$$\log(a \cdot b) = \log a + \log b$$

log

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

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

$$= 0.89$$

log-likelihood

$$\log(L(\mu, \sigma | x_1, x_2, \dots, x_n)) =$$

$$\log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_1-\mu)^2}{2\sigma^2}}\right) + \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_2-\mu)^2}{2\sigma^2}}\right) + \dots + \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_n-\mu)^2}{2\sigma^2}}\right)$$

$$\log(ab) = \log a + \log b$$

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

$$\log a^b = b \log a$$

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

$$\log a^b = b \log a$$

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

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

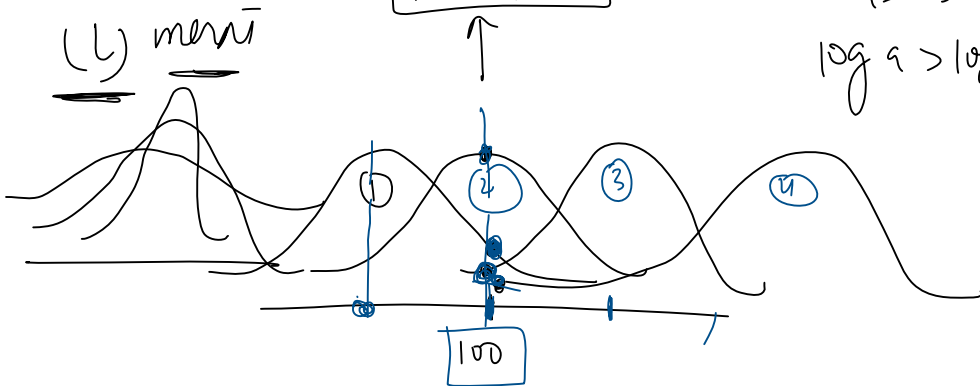
$$= -\frac{1}{2} \log 2\pi - \frac{1}{2} \log \sigma^2 - \frac{(x_1 - \mu)^2}{2\sigma^2}$$

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

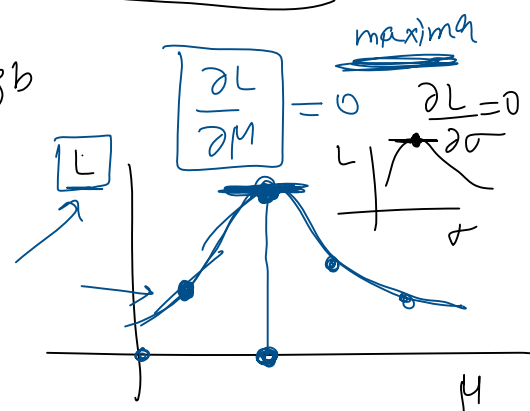
$$-\frac{1}{2} \log 2\pi - \log \sigma - \frac{(x_1 - \mu)^2}{2\sigma^2} - \frac{1}{2} \log 2\pi - \log \sigma - \frac{(x_2 - \mu)^2}{2\sigma^2}$$

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

maximize $\log(L)$ w.r.t μ and σ



$$\log a > \log b$$



$$\frac{\partial \log(L)}{\partial \mu} = \frac{2(x_1 - \mu)}{2\sigma^2} + \frac{(x_2 - \mu)}{\sigma^2} + \frac{(x_3 - \mu)}{\sigma^2} + \dots + \frac{(x_n - \mu)}{\sigma^2} = 0$$

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

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

$$\mu = x_1 + x_2 + \dots + x_n$$

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

$$\log \sigma = \frac{1}{\sigma}$$

$$[\sigma^{-2}] = -2\sigma^{-3}$$

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

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

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

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

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

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

$$\{50, 63, 65, \dots, 100\} \rightarrow \text{Normal}$$

$$\rightarrow N(\text{avg}, \text{std})$$

→ high

→ 27, 28, 31, 35 - ~ 100

distribution \rightarrow MLE

MLE

 μ, σ

- 1) assume some dist
- 2) use a likelihood
- 3) find the value for parameters that maximizes that likelihood

$$= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

 χ, N, opt

Output

 ml

mu → parameter
distribution

| | | | | | |
|-------|-------|-------|---------|-------|-------|
| x_1 | x_2 | x_3 | \dots | x_n | y |
| | | | | | y_n |

Bernoulli

→ 1) find out the dist of $y | X$

→ 2) decide to apply a ml model
parametric in nature → logit

logistic
regression

```

graph TD
    A[Parameter] --> B["β₀ β₁"]
  
```

non-parametric \rightarrow decision tree

3) You randomly decide some values $\beta_0, \beta_1, \beta_2, \dots, \beta_n$

4) select a likelihood function

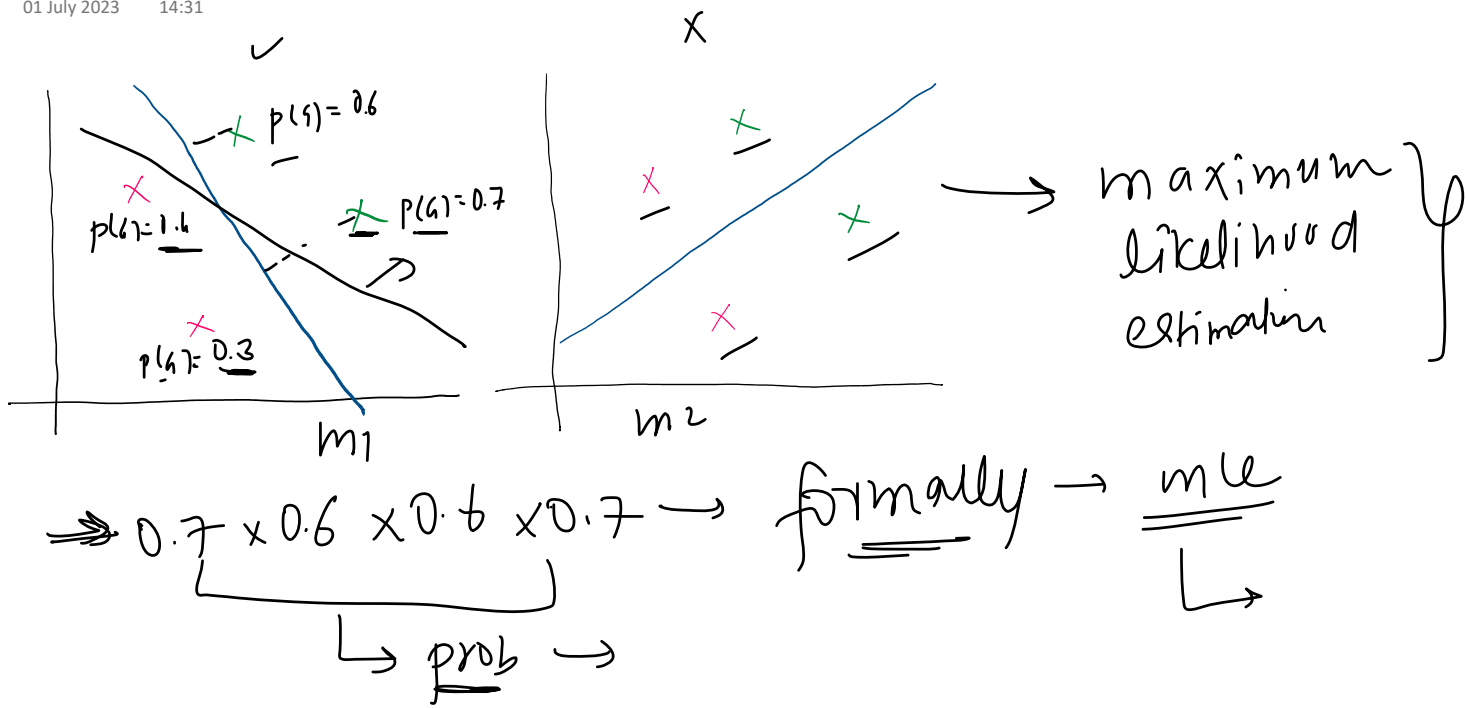
$$p^nf = \underline{pk + (1-p)(1-k)}$$

5) values of $\beta_0 \beta_1 \beta_2 \dots \beta_k$

\rightarrow maximum value

MLE in Logistic Regression

01 July 2023 14:31



Tasks

03 July 2023 15:50

Some Important Questions

01 July 2023 17:15

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, particularly those that are parametric (i.e., defined by a set of parameters), but it's not applicable to all machine learning algorithms.

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

However, there are some machine learning algorithms that don't rely on MLE. For example:

1. **Non-parametric methods:** Some machine learning methods, such as k-Nearest Neighbors (k-NN) and Decision Trees, are non-parametric and do not make strong assumptions about the underlying data distribution. These methods don't have a fixed set of parameters that can be optimized using MLE.
2. **Unsupervised learning algorithms:** Some unsupervised learning algorithms, like K-means clustering, use different objective functions, not necessarily tied to a probability distribution.
3. **Reinforcement Learning:** Reinforcement Learning methods generally don't use MLE, as they are more focused on learning from rewards and punishments over a sequence of actions rather than fitting to a specific data distribution.

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

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

Maximum Likelihood Estimation (MLE) is a method of estimating the parameters of a statistical model to maximize the likelihood function, which is conceptually similar to minimizing 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 functions in machine learning are closely related. Many common loss functions can be derived from the principle of maximum likelihood estimation under certain assumptions 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 perspectives to look at the same problem.

In many machine learning algorithms, the aim is to minimize the difference 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 parameters that give us the closest predictions to the actual values.

On the other hand, when we look at the problem from a statistical perspective, we talk in terms of maximizing the likelihood of seeing the observed data given the model parameters. 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 functions can be derived from the principle of MLE under certain assumptions 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 methods and unsupervised learning algorithms 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.

In summary, while the concepts of loss function minimization and maximum likelihood estimation are closely related and often equivalent, the concept of a loss function is more flexible and computationally convenient, which is why it's more commonly used in the machine learning community.

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 functions used in many common machine learning algorithms, and it helps us understand why these loss functions work and under what assumptions.
2. **Interpretability:** The MLE framework gives us a way to interpret our model parameters. The MLEs are the parameters 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 tools 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 on our machine learning tasks.

In short, while you can often get by with a practical understanding of loss functions and optimization algorithms in applied machine learning, understanding MLE can be extremely valuable for gaining a deeper understanding of how and why these models work.