Lecture 4: Parameter Estimation (Part I)

Statistical Methods for Data Science

Yinan Yu

Department of Computer Science and Engineering

November 9, 2023

Today

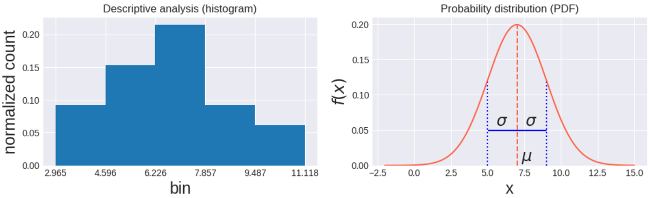
1. Mathematical modeling
2. Parameter estimation
   - Maximum likelihood estimation (MLE)
   - Likelihood and likelihood function
   - Joint probability distribution
   - Independence
3. Summary

Learning outcome

- Be able to explain different components in a mathematical model $y = g(x; \theta \mid h)$
- Understand the purpose and general steps of parameter estimation
- Be able to explain these concepts: joint probability distribution, independent and identically distributed (i.i.d.) random variables, likelihood function, maximum likelihood

Recap: three questions from lecture 2

Jack suggested to use a Gaussian distribution to model your data.



- ✅ Question 1: Why should I use probability distributions instead of histograms?
- ✅ Question 2: How do you know if my data follows a Gaussian distribution?
- ? Question 3: How do I find the unknown parameters?

In today's lecture, we are going to address question 3.

Today

What you will learn from this section

In the previous section, we have touched upon the topic of choosing a probabilistic model to describe a given data set.

Generally speaking, given a data set and a problem to be solved, you need to formulate the solution mathematically so that you can write a computer program to solve the problem. This is the main task for a data scientist.

This section aims to help you get started by providing explicit components and steps for formulating mathematical models.

Terminology

- What is mathematical modeling? - Mathematical modeling is to *describe* a system using the language of mathematics in order to solve **a range of problems**.
- What the *description* looks like in data science:

$$y = g(x; \theta \mid h)$$

- Left hand side:
  - y: target or label - what you want to predict; **a result** that answers the question at hand
- Right hand side:
  - x: variables or features - placeholder for data in order to solve *a range of* problems; **the input**
  - g: model - a mathematical function that can be used to solve a given range of problems - a model is given by domain experts or derived from your assumption; can be selected from established models; **known** except for some parameters
  - h: hyperparameters - part of the model g (given or derived from your assumption); **known** (but you might need to "guess" them first)
  - θ: parameters - part of the model g; in a data-driven paradigm θ is **unknown**; need to be estimated from data
- Symbols:
  - Semicolon (";") is used to emphasize that θ is not known for free - it needs to be estimated
  - Bar ("|" pronounced "*given*") is used to indicate that h is known to you
- Note: x, y, θ and h are not necessarily scalars; they can be multiple scalars, vectors or more complex data structures; g can be complex functions, for instance, a machine learning model or a deep neural network.

Five questions

Overwhelmed? Take it easy! Here is something that helps you get started!
Answer these five questions in the language of mathematics step by step:

- 1) What do we want to predict, i.e. what is the target y?
- 2) What are the variables x?
- 3) What is the mathematical function g that relates variables x to the target y?
- 4) Are there any hyperparameters h in the function g? How do we choose them?
- 5) What are the unknown parameters θ in g? **How do we estimate them from data?**

Probabilistic modeling

Model data using probability distributions
Example:

$$y = g(x_1, x_2; \mu, \sigma) = P(x_1 \leq weight \leq x_2) = \int_{x_1}^{x_2} f_X(t)dt = \int_{x_1}^{x_2} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt$$

- 1) What do we want to predict, i.e. what is the target y? - The probability of the event $x_1 \leq weight \leq x_2$
- 2) What are the variables x? - x_1 and x_2
- 3) What is the mathematical function g that relates variables x to the target y? - The integral of the Gaussian PDF
- 4) Are there any hyperparameters h in the function g? How do we choose them? - There is none in this case
- 5) What are the unknown parameters - μ and σ **How do we estimate them from data?**

General steps of parameter estimation for probabilistic models

- Note: the estimate of θ is denoted as $\hat{\theta}$.
- General steps for parameter estimation for a probabilistic model g
  - a) Describe the **experiments**
  - b) Describe the **data** generated from the experiments
  - c) Describe the **random variables**
  - d) Identify **parameters of interest** θ
  - e) Choose an **estimation method**, e.g. MLE/MAP
  - f) **Compute** $\hat{\theta}$ typically by solving an optimization problem
    - Closed-form solution for simple cases
    - Iterative methods for general cases
  - g) Evaluation: estimate and report the uncertainty of $\hat{\theta}$ (later)
- **Underlying assumption**: the data we use for parameter estimation is generated from the same distribution as the data we use for prediction.

Today

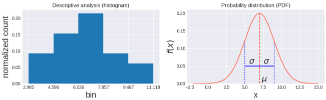
1. Mathematical modeling
2. Parameter estimation
    - Maximum likelihood estimation (MLE)
    - Likelihood and likelihood function
    - Joint probability distribution
    - Independence
3. Summary

Overview

Given a dataset and a problem to be solved, now you know how to choose a probability distribution. However, the model has unknown parameters. In this section, you will learn how to estimate these parameters from data.

- There are two important parameter estimation methods: 1) the **maximum likelihood estimation (MLE)** and 2) the **maximum a posteriori estimation**.
- Concepts such as likelihood function, independent and identically distributed random variables, prior, posterior, Bayes' rule, etc are important building blocks for future machine learning models.

Overview (cont.)



- In a Gaussian distribution, what are the parameters to be estimated? mean μ and standard deviation σ
- The **maximum likelihood estimates** are the sample mean $\bar{x}$ and the sample standard deviation s for parameters μ and σ, respectively.

$$\hat{\mu} = \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\hat{\sigma} = s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

- Straightforward for Gaussian distribution! Gaussian is great!
- However, it is not straightforward for all distributions - it is important to properly understand the MLE framework.

Maximum likelihood estimation (MLE)

Simplest case study: estimate one parameter given one observation

- **Model** g (cf. lecture 3):
  - **Assumption**: a duck's weight is drawn from a Gaussian distribution with standard deviation σ and mean μ

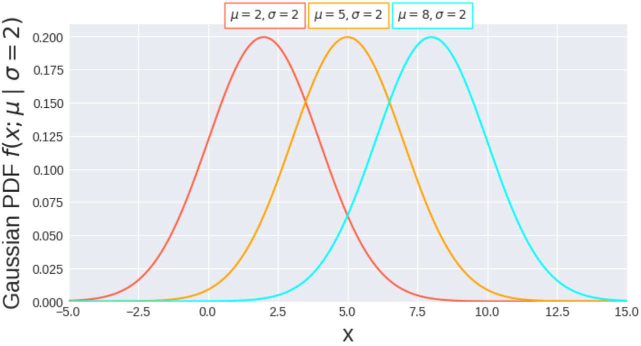
  To simplify the problem for illustration purposes, let's only look at one parameter for now:
  - We assume that σ is known to us: $\sigma = 2$
  - Unknown parameter: μ

  We want to estimate this unknown parameter by collecting some data from experiments.
- **Experiment**: we weigh a duck and observe its weight
- **Data**: the duck weighs 4 kg
- **Random variable**: $X = x$ if a duck weighs x kg
- **Parameter of interest**: μ
- **Estimation method**: the maximum likelihood estimation for μ
- **Compute** $\hat{\mu}_{MLE}$ by maximizing the **likelihood function**

Can you guess what result we are going to get? $\hat{\mu}_{MLE} = 4$

Intuition

Which Gaussian distribution is most "likely" to be the underlying model for the given data $x = 4$?

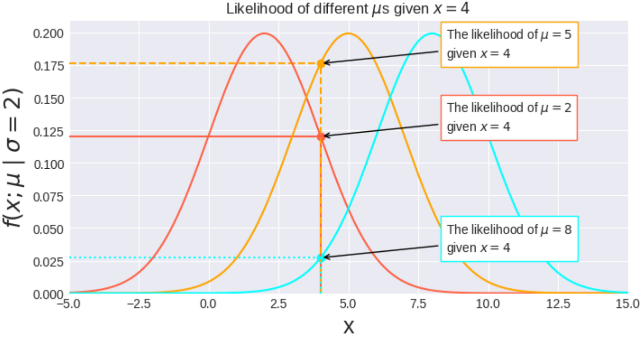


Likelihood and likelihood function

🚨 Terminology alert 🚨 - likelihood

Assumption (reminder): weights x follow a Gaussian distribution with unknown parameter μ and known $\sigma = 2$

- **Likelihood of μ** given data $x = 4$ is $f(x = 4; \mu \mid \sigma = 2)$



A nonrigorous note on functions and variables

- Let g be a function that relates input variables x to a target y:

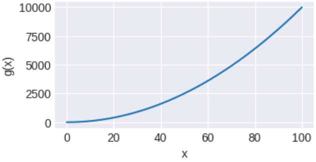
$$y = g(x)$$

- Typically, we care about the behavior of y for **all possible values for x**. This is called **generalization** in machine learning.
- Even if we add parameters θ and hyperparameters h to g, $g(x; \theta \mid h)$ is still a function of x.
- In a plot, we typically place the variable on the x-axis!
- If we are interested in the behavior of y in terms of θ, we can construct a different function L that takes θ as the variables $y = L(\theta)$ to relate θ to y.

A nonrigorous note on functions and variables (cont.)

- Example: $y = g(x) = x^2$
- In Python, **all possible values for x** means something like this:

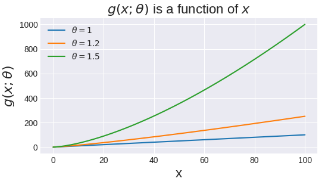
```
# Assume x can take any value between 0 and 100
xmin, xmax = 0, 100
N = 10000 # ideally, N should be infinity. But sadly, computers are discrete
          # so N has to be finite.
x = np.linspace(xmin, xmax, num=N) # all possible values for x
# Plot a function
def g(t):
    return np.power(t, 2)
y = g(x)
plt.plot(x, y)
```



A nonrigorous note on functions and variables (cont.)

- Now we add a parameter θ to g: $y = g(x; \theta) = x^{\theta}$

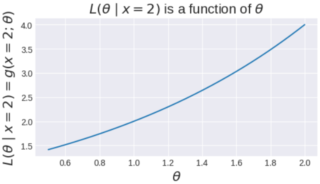
```
def g_theta(t, theta):
    return np.power(t, theta)
xmin, xmax = 0, 100 # assume x can take any value between 0 and 100
N = 10000
x = np.linspace(xmin, xmax, num=N) # all possible values for x
y = g_theta(x, 1)
plt.plot(x, y)
y = g_theta(x, 1.2)
plt.plot(x, y)
y = g_theta(x, 1.5)
plt.plot(x, y) # x is still on the x-axis
```



A nonrigorous note on functions and variables (cont.)

- Now we define a new function: $y = L(\theta \mid x = 2) = g(x = 2; \theta) = 2^{\theta}$

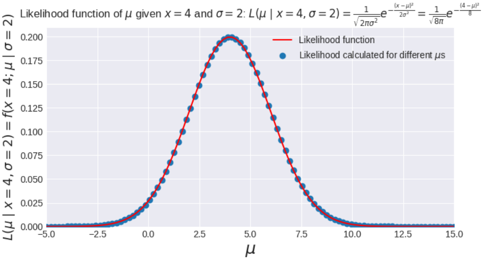
```
def L(t):
    return g_theta(2, t)
# Now theta is the variable! So we need to get all possible values for theta
# Assume theta can take any value between 0.5 and 2
theta_min, theta_max = 0.5, 2
N = 10000
thetas = np.linspace(theta_min, theta_max, num=N) # all possible values for theta
y = L(thetas)
plt.plot(thetas, y) # theta is on the x-axis now
```



Terminology alert - likelihood function

- **Likelihood function of μ** given data $x = 4$ for $-\infty \leq \mu \leq \infty$:

$$
\begin{aligned}
L(\mu \mid x=4, \sigma=2) &= f(x=4; \mu \mid \sigma=2) \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \\
&= \frac{1}{\sqrt{8\pi}} e^{-\frac{(4-\mu)^2}{8}}
\end{aligned}
$$

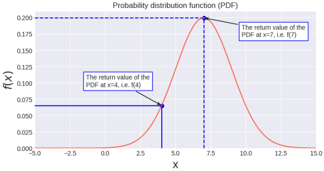


- A tiny note about symbol (most abstract), definition (less abstract) and computation (concrete - something you can implement it in Python)

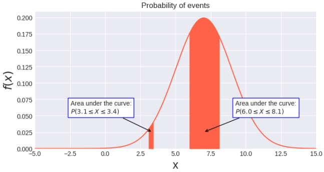
Recall: probability density function and probability of events

Gaussian distribution with $\mu = 7$, $\sigma = 2$

- Probability density function $f(x \mid \mu = 7, \sigma = 2)$:



- Probability of events $P(x_1 \leq X \leq x_2)$:



Probability of events vs likelihood function

- Probability of events given $\mu = 7$ and $\sigma = 2$:

$$\begin{aligned} g(x_1, x_2 \mid \mu = 7, \sigma = 2) &= P(x_1 \leq X \leq x_2) \\ &= \int_{x_1}^{x_2} f_X(t)dt = \int_{x_1}^{x_2} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt \\ &= \int_{x_1}^{x_2} \frac{1}{\sqrt{8\pi}} e^{-\frac{(t-7)^2}{8}} dt \end{aligned}$$

  Here x_1 and x_2 are the **variables** - when we change x_1 and x_2, we get a different probability $g(x_1, x_2 \mid \mu = 7, \sigma = 2)$.
- Likelihood function for a given observation $x = 4$ (with known $\sigma = 2$):

$$L(\mu \mid x = 4, \sigma = 2) = \frac{1}{\sqrt{8\pi}} e^{-\frac{(4-\mu)^2}{8}}$$

  Here μ is the **variable** - when we change μ, we get a different likelihood $L(\mu \mid x = 4, \sigma = 2)$.

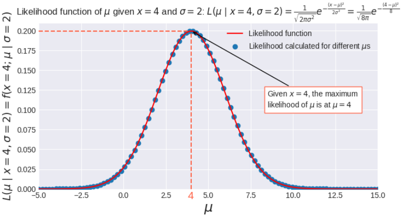
Maximum likelihood

From the likelihood function

$$L(\mu \mid x = 4, \sigma = 2) = \frac{1}{\sqrt{8\pi}} e^{-\frac{(4-\mu)^2}{8}}$$

We can now define the **maximum likelihood** of μ given $x = 4$:

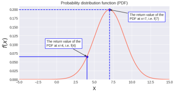
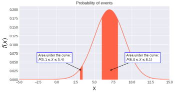
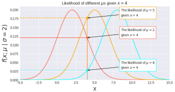
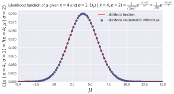
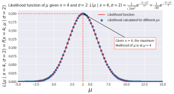
$$\text{the maximum likelihood of } \mu = \max(L(\mu \mid x = 4, \sigma = 2))$$



The value of μ that maximizes the likelihood function is called the **maximum likelihood estimation** (MLE) of μ. In this case, $\hat{\mu}_{\text{MLE}} = 4$.

Note: $\hat{}$ here means that $\hat{\mu}$ is an estimate instead of the true value μ.

Comparison

Probability density function	
Probability of events	
Likelihood of a parameter given data	
Likelihood function of a parameter given data	
Maximum likelihood estimation	

Summary: what have we done so far?

- We observe one data point $x = 4$.
- We assume that duck weights are drawn from a *Gaussian distribution* with known $\sigma = 2$ and unknown μ. We need to estimate μ.
- We write down the likelihood function:
  $L(\mu \mid x = 4, \sigma = 2) = \frac{1}{\sqrt{8\pi}} e^{-\frac{(4-\mu)^2}{8}}$.
- The maximum likelihood estimation of μ is defined as:

$$\hat{\mu}_{\text{MLE}} = \arg\max_{\mu} L(\mu \mid x = 4, \sigma = 2) = \arg\max_{\mu} \frac{1}{\sqrt{8\pi}} e^{-\frac{(4-\mu)^2}{8}} \quad (1)$$

Remaining questions

- We can't estimate the whole distribution from only one data point $x = 4$! What if we have more than one observation?
- How can we maximize the likelihood function and find the value of $\hat{\mu}_{MLE}$ analytically?
- What if σ is also unknown?
- What about discrete distributions?

Case study: parameter estimation given more observations

- **Model**:
  - Assumption: a duck's weight is drawn from a Gaussian distribution with known standard deviation $\sigma = 2$ and unknown mean μ
- **Experiment**: we observe 20 ducks
- **Data**:

duck id	1	2	3	4	$\cdots$	19	20
weight	6.98	5.43	2.97	7.07	$\cdots$	4.63	7.27

- **Parameter of interest**: μ
- **Estimation method**: maximum likelihood estimation
- **Compute** $\hat{\mu}_{MLE}$ by maximizing the likelihood function
- Recall: when we only have one observation $x = 4$, the likelihood function looks like this

$$L(\mu \mid x = 4, \sigma = 2) = f(x = 4; \mu \mid \sigma = 2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{8\pi}} e^{-\frac{(4-\mu)^2}{8}}$$

- Educated guess 🤔 - now we have more observations, the likelihood function probably should look like this:

$$L(\mu \mid x_1 = 6.98, \cdots, x_{20} = 7.27, \sigma = 2) = \boxed{f(x_1 = 6.98, \cdots, x_{20} = 7.27; \mu \mid \sigma = 2)}$$

Joint probability distribution

🚨 Terminology alert 🚨 - joint probability distribution 👯

Given two random variables X and Y, we use their **joint probability distribution** to characterize their behaviors:

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y) \quad \text{joint CDF}$$

- X, Y discrete: joint PMF $f_{X,Y}(x, y) = P(X = x, Y = y)$
- X, Y continuous: joint PDF $f_{X,Y}(x, y)$

- Bummer: these expressions are usually quite hard to obtain...
- Solution: we impose some assumptions to make the calculation easier.

Independence

Independence 😼

- Recall independent events: two events A and B are independent if and only if

$$P(A \text{ and } B) = P(A \cap B) = P(A)P(B)$$

New! **Independent random variables**: random variables X, Y are independent if and only if

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y) = F_X(x)F_Y(y)$$

  - X, Y discrete:

$$f_{X,Y}(x, y) = P(X = x, Y = y) = P(X = x)P(Y = y) = f_X(x)f_Y(y)$$

    where $f_{X,Y}(x, y)$ is the joint PMF
  - X, Y continuous:

$$f_{X,Y}(x, y) = f_X(x)f_Y(y)$$

    where $f_{X,Y}(x, y)$ is the joint PDF
- This idea generalizes to more than two random variables

Independence 😼

Any number of random variables:

- Given n random variables $X_1, X_2, \cdots, X_n$ with CDF $F_{X_i}(x_i)$,

$$F_{X_1,\cdots,X_n}(x_1, \cdots, x_n) = \prod_{i=1}^{n} F_{X_i}(x_i)$$

  where $F_{X_1,\cdots,X_n}(x_1, \cdots, x_n)$ is the joint CDF

  - X_i discrete with PMF $f_{X_i}(x)$:

$$f_{X_1,\cdots,X_n}(x_1, \cdots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)$$

    where $f_{X_1,\cdots,X_n}(x_1, \cdots, x_n)$ is the joint PMF

  - X_i continuous with PDF $f_{X_i}(x)$:

$$f_{X_1,\cdots,X_n}(x_1, \cdots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)$$

    where $f_{X_1,\cdots,X_n}(x_1, \cdots, x_n)$ is the joint PDF

- Now we have turned the joint probability distribution into multiplications of things we know how to compute. Yay!

Back to the case study

The likelihood function

$$L(\mu \mid x_1 = 6.98, \cdots, x_{20} = 7.27, \sigma = 2) = \boxed{f(x_1 = 6.98, \cdots, x_{20} = 7.27; \mu \mid \sigma = 2)}$$

- **Model**:
  - Assumption: the weight is drawn from a Gaussian distribution with known standard deviation $\sigma = 2$ and unknown mean μ
- **Experiment**: we weigh 20 ducks
- **Data**:

duck id	1	2	3	4	$\cdots$	19	20
weight	6.98	5.43	2.97	7.07	$\cdots$	4.63	7.27

- **Random variable**: we define 20 random variables X_i : duck weight $\to \mathbb{R}$, where X_i are **independent and identically distributed (i.i.d)** Gaussian random variables
  - $X_1, \cdots, X_{20}$ are independent - [new!] assumption:

$$f_{X_1,\cdots,X_{20}}(x_1 = 6.98, \cdots, x_{20} = 7.27) = f_{X_1}(x = 6.98) \cdots f_{X_{20}}(x = 7.27)$$

  - $X_1, \cdots, X_{20}$ are identically distributed - they have the same PDF:

$$f_{X_1}(x; \mu \mid \sigma) = \cdots = f_{X_{20}}(x; \mu \mid \sigma) = f(x; \mu \mid \sigma)$$

  where $\sigma = \sigma_1 = \sigma_2 = \cdots = \sigma_{20} = 2$ and $\mu = \mu_1 = \mu_2 = \cdots = \mu_{20}$.
- **Parameter of interest**: μ
- **Estimation method**: maximum likelihood estimation
- **Compute** $\hat{\mu}_{MLE}$ by maximizing the likelihood function

Today

Summary

So far:

- Data types and data containers
- Descriptive data analysis: descriptive statistics, visualization
- Probability distributions, events, random variables, PMF, PDF, parameters
- CDF, Q-Q plot, how to compare two distributions (data vs theoretical, data vs data)
- Modeling
- Parameter estimation: maximum likelihood estimation (to be continued...)

Next (part II):

- Maximum a posteriori estimation

Before next lecture:

- Conditional probability, i.i.d. random variables

Until next time!