Lecture 4: Parameter Estimation

Statistical Methods for Data Science

Yinan Yu

Department of Computer Science and Engineering

November 12, 2020

Today

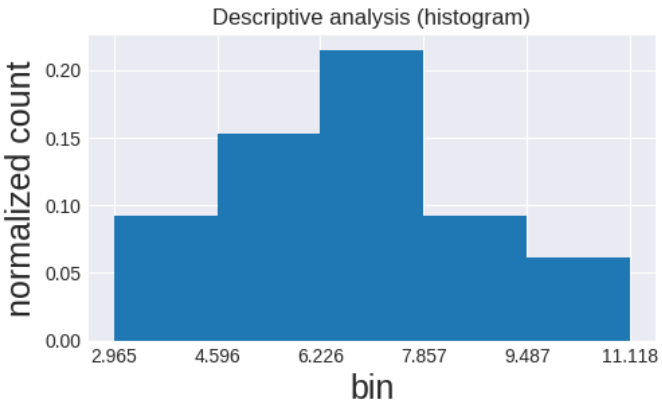
1. Parameter estimation
   - Maximum likelihood estimation (MLE)
   - Maximum a posteriori estimation (MAP)
   - MLE vs MAP
2. Summary

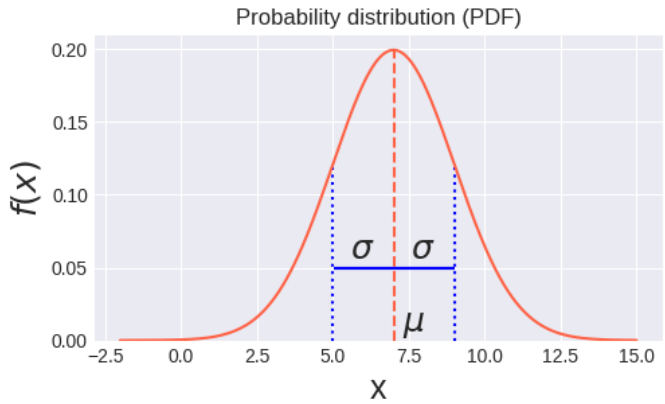
Learning outcome

- Understand the purpose and general steps of parameter estimation
- Be able to explain these terminologies: joint probability distribution, independent and identically distributed (i.i.d.) random variables, likelihood function, maximum likelihood, prior, posterior
- Be able to formulate MLE and MAP
- Be able to implement simple cases in Python

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

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

- There are two dominant parameter estimation methods for probabilistic models: 1) the **maximum likelihood estimation (MLE)** and 2) the **maximum a posteriori estimation**.
- After this section, you should be able to 1) formulate a parameter estimation problem as MLE and MAP; 2) implement solutions for simple cases in Python.
- 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, e.g. naive Bayes classifiers.

Recall: five steps for modeling from lecture 3

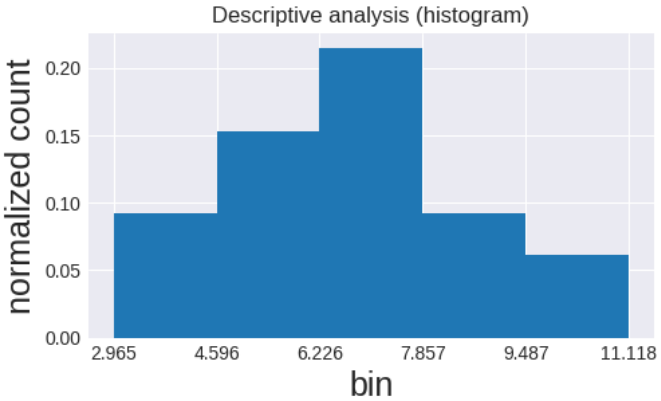
- 1) What do we want to predict, i.e. what is the target y?
- 2) What are the variables x?
- 3) What is the mathematical function g that relates variables x to the target y?
- 4) Are there any hyperparameters h in the function g? How do we choose them?
- 5) What are the unknown parameters θ in g? **How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

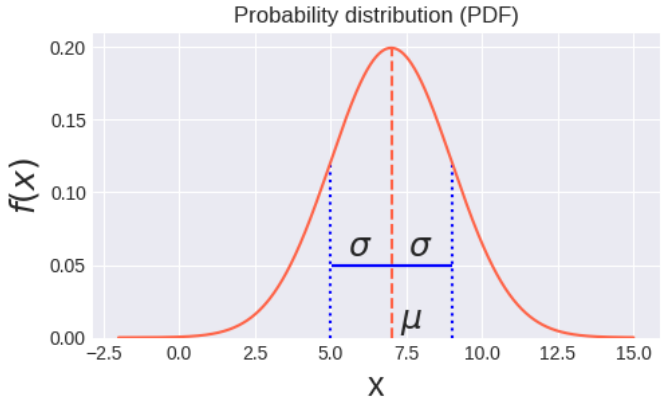
General steps of parameter estimation

- 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) Choose a **parameter 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.

Maximum likelihood estimation (MLE)

Why is it important?





- In the Gaussian case, the maximum likelihood estimates are the sample mean and the sample standard deviation for parameters μ and σ, respectively.
- These estimations match our intuition.
- Maximum likelihood estimation is to put this intuition into a theoretical framework in order to get a better understanding of their analytical properties.
- Although we are using simple examples to illustrate the concept, this framework can be applied to any distributions with more complex data.

What you will learn from this subsection

Terminology

- Joint probability distribution
- Independence
- Independent and identically distributed (i.i.d.) random variables
- Likelihood, likelihood function and maximum likelihood
- How to find the maximum likelihood estimation

Scenarios covered:

- Continuous distribution:
  1) Gaussian distribution with one parameter given one observation
  2) Gaussian distribution with one parameter given more than one observations
  3) Gaussian distribution with more than one parameters given more than one observations
- Discrete distribution:
  4) Bernoulli distribution

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

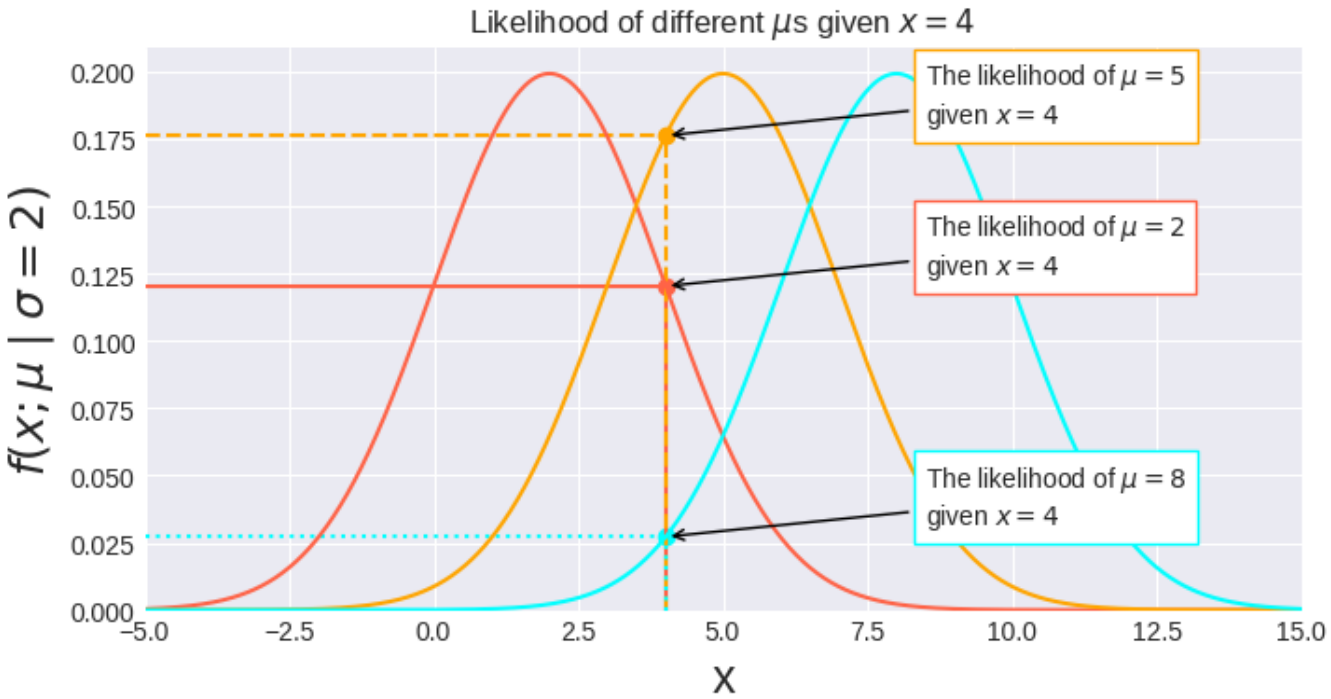
  We want to estimate this unknown parameter by running some experiments.
- **Experiment**: we weigh a duck and observe its weight
- **Data**: the duck weighs 4 kg
- **Random variable**: $X = x$ if a duck weighs x kg
- **Parameter of interest**: μ
- **Estimation method**: the maximum likelihood estimation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

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

Terminology alert - likelihood

Gaussian distribution with unknown parameter μ and known $\sigma = 2$

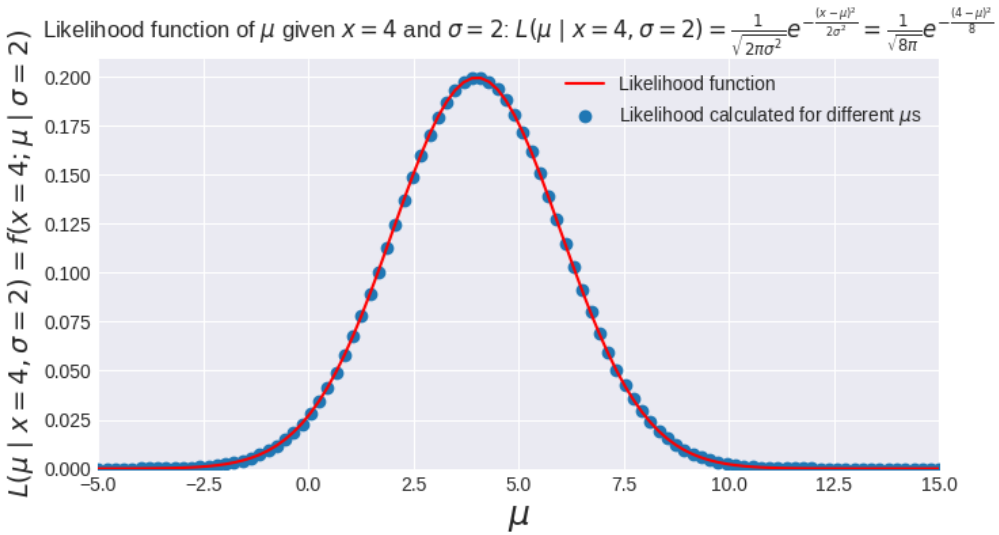
- Likelihood of μ given data $x = 4$: $f(x = 4; \mu \mid \sigma = 2)$:



Terminology alert - likelihood function

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

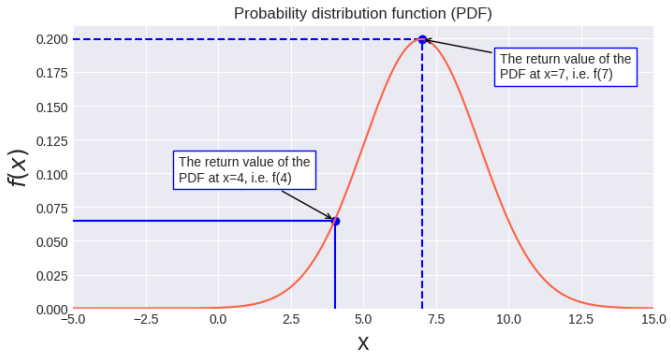
$$\begin{aligned} L(\mu \mid x = 4, \sigma = 2) &= f(x = 4; \mu \mid \sigma = 2) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \\ &= \frac{1}{\sqrt{8\pi}} e^{-\frac{(4-\mu)^2}{8}} \end{aligned}$$



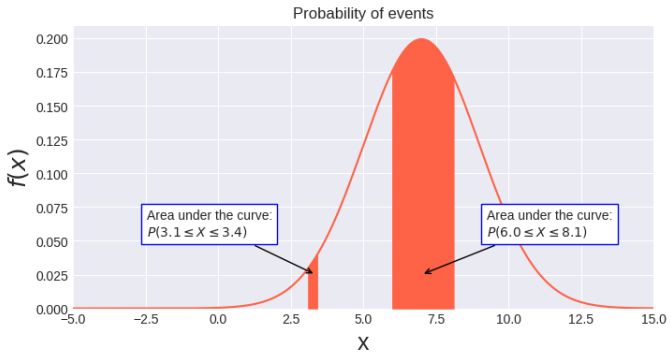
Recall: probability density function and probability of events

Gaussian distribution with known μ, σ

- 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 $L(\mu \mid x = 4, \sigma = 2)$.

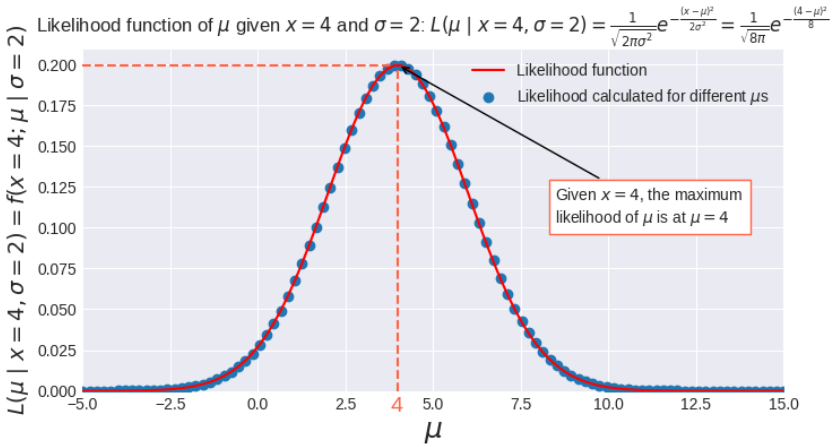
Maximum likelihood

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 $x = 4$:

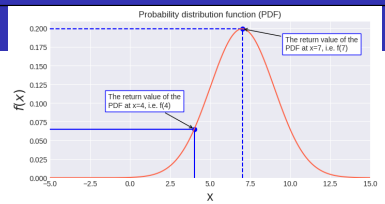
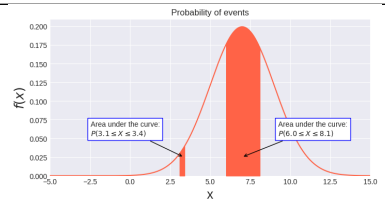
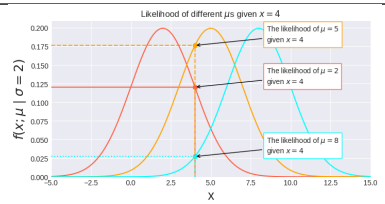
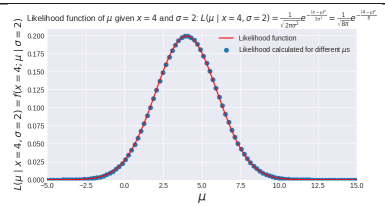
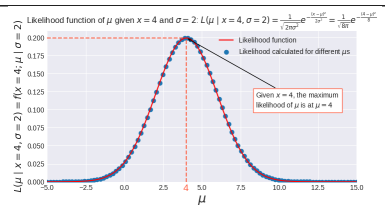
the maximum likelihood of $\mu = \max(L(\mu \mid x = 4, \sigma = 2))$



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 estimate 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:

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

🚨 Terminology alert 🚨 - joint probability distribution 👯

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

- X, Y discrete:

$$f_{X,Y}(x, y) = P(X = x, Y = y) \quad \text{joint PMF}$$

- X, Y continuous:

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

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

Independence

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

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

- 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

- 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 $\rightarrow \mathbb{R}$, where X_i are **independent and identically distributed (i.i.d)** Gaussian random variables
  - $X_1, \cdots, X_{20}$ are independent:

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

  - $X_1, \cdots, X_{20}$ 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

Case study: the likelihood function given i.i.d. observations

- Put everything together, the likelihood function is expressed as

$$\begin{aligned} & L(\mu \mid x_1 = 6.98, x_2 = 5.43, \cdots, x_{20} = 7.27, \sigma = 2) \\ = \; & f(x_1 = 6.98, x_2 = 5.43, \cdots, x_{20} = 7.27; \mu \mid \sigma = 2) \\ = \; & f(6.98; \mu \mid \sigma = 2) f(5.43; \mu \mid \sigma = 2) \cdots f(7.27; \mu \mid \sigma = 2) \\ = \; & \frac{1}{\sqrt{8\pi}} e^{-\frac{(6.98-\mu)^2}{8}} \frac{1}{\sqrt{8\pi}} e^{-\frac{(5.43-\mu)^2}{8}} \cdots \frac{1}{\sqrt{8\pi}} e^{-\frac{(7.27-\mu)^2}{8}} \\ = \; & (\frac{1}{\sqrt{8\pi}})^{20} e^{-\frac{(6.98-\mu)^2+\cdots+(7.27-\mu)^2}{8}} \end{aligned}$$

- The MLE can be computed as:

$$\begin{aligned} \hat{\mu}_{MLE} &= \arg\max_{\mu} L(\mu \mid x_1 = 6.98, x_2 = 5.43, \cdots, x_{20} = 7.27, \sigma = 2) \\ &= \arg\max_{\mu} (\frac{1}{\sqrt{8\pi}})^{20} e^{-\frac{(6.98-\mu)^2+\cdots+(7.27-\mu)^2}{8}} \end{aligned}$$

- From page 8, the only step left is f) compute $\hat{\mu}_{MLE}$.

Compute $\hat{\mu}_{MLE}$ from the likelihood function 

The maximum likelihood estimate of μ is the value that maximizes the likelihood function

$$
\begin{aligned}
\hat{\mu}_{MLE} &= \arg\max_{\mu} L(\mu \mid x_1 = 6.98, x_2 = 5.43, \cdots, x_{20} = 7.27, \sigma = 2) \\
&= \arg\max_{\mu} L(\mu \mid \boldsymbol{x}) \text{ (for convenience)} \\
&= \arg\max_{\mu} \log L(\mu \mid \boldsymbol{x}) \text{ (does not change the location of the maximum)} \\
&\quad \text{Log function turns multiplications into sums, which makes everything much easier to compute} \\
&= \arg\max_{\mu} l(\mu \mid \boldsymbol{x}) \text{ (l is called the log likelihood)} \\
&= \arg\min_{\mu} -l(\mu \mid \boldsymbol{x}) \text{ (this is the negative log likelihood - mainly for standardization purposes)} \\
&= \arg\min_{\mu} -\log(\frac{1}{\sqrt{8\pi}})^{20} + \frac{(6.98-\mu)^2 + \cdots + (7.27-\mu)^2}{8} \text{ (plug in the expression of the PDFs)} \\
&\quad \text{Set the derivative to zero to find the optimal solution} \\
\iff & \frac{\partial}{\partial \mu}\left(-\log(\frac{1}{\sqrt{8\pi}})^{20} + \frac{(6.98-\mu)^2 + \cdots + (7.27-\mu)^2}{8}\right) = 0 \\
\iff & \sum_{i=1}^{20} x_i - 20\mu = 0 \\
\iff & \hat{\mu}_{MLE} = \frac{1}{20}\sum_{i=1}^{20} x_i = 6.647
\end{aligned}
$$

Compute $\hat{\mu}_{MLE}$ from the likelihood function (cont.)

Okay, here is the takeaway...

- The MLE of the mean value of a Gaussian distribution is the sample mean.
- It's an **optimization problem**. Read about the context here.
- Given the i.i.d. assumption, the MLE of parameters have closed-form solutions for many known distributions, e.g. Gaussian, uniform, Bernoulli, etc, e.g. google "mle Gaussian distribution".
- In general cases, the solution needs to be found using **iterative methods** and **approximation techniques**.
- The good news is 1) i.i.d. is usually a reasonable assumption; 2) many things in the world are naturally Gaussian!

Note: solving optimization problems is an important building block of data science and machine learning. For example, if you want to understand deep learning with backpropagation, you need to know at least why fiddling with the derivative leads to the optimal solution.

What if σ is also unknown?

- Now we have two unknown parameters μ and σ, the likelihood function becomes:

$$L(\mu, \sigma \mid \boldsymbol{x}) = \left(\frac{1}{\sqrt{2\sigma^2\pi}}\right)^{20} e^{-\frac{(6.98-\mu)^2+\cdots+(7.27-\mu)^2}{2\sigma^2}}$$

- Now we need to solve for both μ and σ:

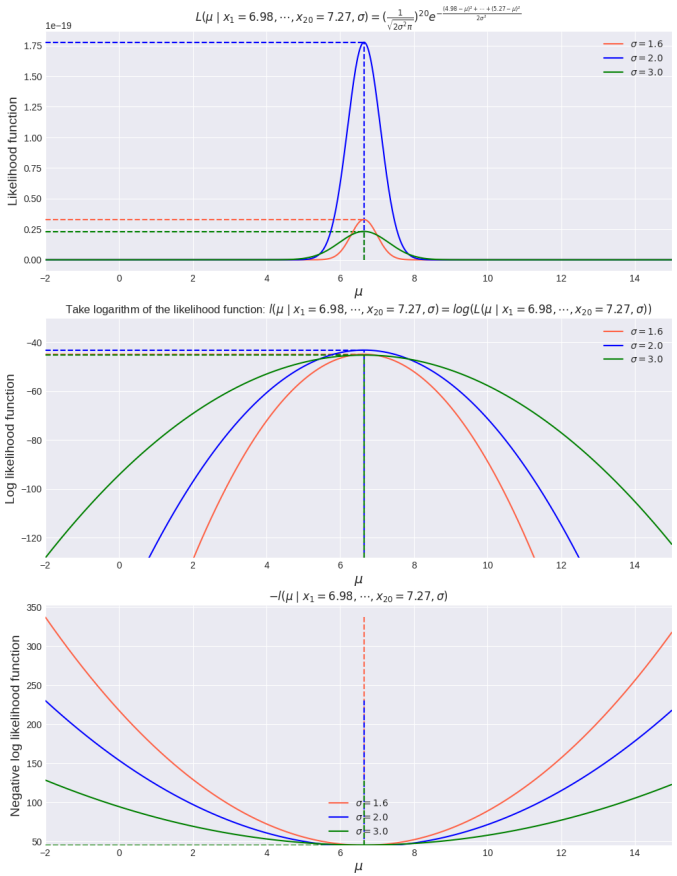
$$\hat{\mu}_{MLE} = \arg\max_{\mu} L(\mu, \sigma \mid \boldsymbol{x}) \iff \frac{\partial}{\partial \mu} L(\mu, \sigma \mid \boldsymbol{x}) = 0$$

$$\hat{\sigma}_{MLE} = \arg\max_{\sigma} L(\mu, \sigma \mid \boldsymbol{x}) \iff \frac{\partial}{\partial \sigma} L(\mu, \sigma \mid \boldsymbol{x}) = 0$$

Note: left as an exercise (data can be found in Jupyter Notebook).

- What happens to the partial derivative of σ with respect to μ?

$\hat{\mu}_{MLE}$ given different σ



NOTE: the maximum likelihood estimation of μ does not depend on σ

What about discrete distributions?

Likelihood function for discrete distributions:

- Likelihood function given one discrete random variable X with PMF $f_X(x)$:

$$L(\theta \mid x) = P(X = x) = f_X(x; \theta)$$

- Likelihood function given N i.i.d. discrete random variables with PMF $f(x)$:

$$L(\theta \mid x_1, \cdots, x_N) = \prod_{i=1}^{N} f(x_i; \theta)$$

Example: MLE for discrete distributions

- **Model**:
  - Assumption: the color of a duck is drawn from a Bernoulli distribution with PMF

$$f(k; p) = pk + (1 - p)(1 - k), k \in \{0, 1\}$$

- **Experiment**: we observe the colors from 20 ducks
- **Data**:

duck id	1	2	3	4	$\cdots$	19	20
color	red	blue	red	red	$\cdots$	red	blue

- **Random variable**: let $X_i = \begin{cases} 0, & \text{a duck is red} \\ 1, & \text{a duck is blue} \end{cases}$ be independent and identically distributed (i.i.d.) Bernoulli random variables.
- **Parameter of interest**: p
- **Estimation method**: maximum likelihood estimation
- **Compute** $\hat{p}_{MLE}$ by maximizing the likelihood function

Example: MLE for discrete distributions  

- Likelihood function:

$$\begin{aligned} L(p \mid \boldsymbol{x}) &= f(x_1; p) f(x_2; p) \cdots f(x_{20}; p) \text{ (using ";" instead of "|" because } p \text{ is unknown)} \\ &= (1-p) \times (1-0) \times p \times 1 \cdots p \times 1 \text{ (plug in Bernoulli PMF)} \\ &= p^{(\#\text{ of blue ducks})} (1-p)^{(\#\text{ of red ducks})} \\ &= p^6 (1-p)^{14} \end{aligned}$$

- Negative log likelihood function:

$$-l(p \mid \boldsymbol{x}) = -\log(p^6(1-p)^{14}) = -(6\log(p) + 14\log(1-p))$$

- Minimize the negative log likelihood function by setting the derivative to zero:

$$\begin{aligned} & \frac{\partial}{\partial p}(-l(p \mid \boldsymbol{x})) = 0 \\ \iff & \frac{\partial}{\partial p}(6\log(p) + 14\log(1-p)) = 0 \text{ (we need to use a tiny chain rule here)} \\ \iff & \frac{6}{p} - \frac{14}{1-p} = 0 \Rightarrow p = \frac{6}{20} \\ \Rightarrow & \hat{p}_{MLE} = \arg\min_p -l(p \mid \boldsymbol{x}) = \frac{6}{20} \end{aligned}$$

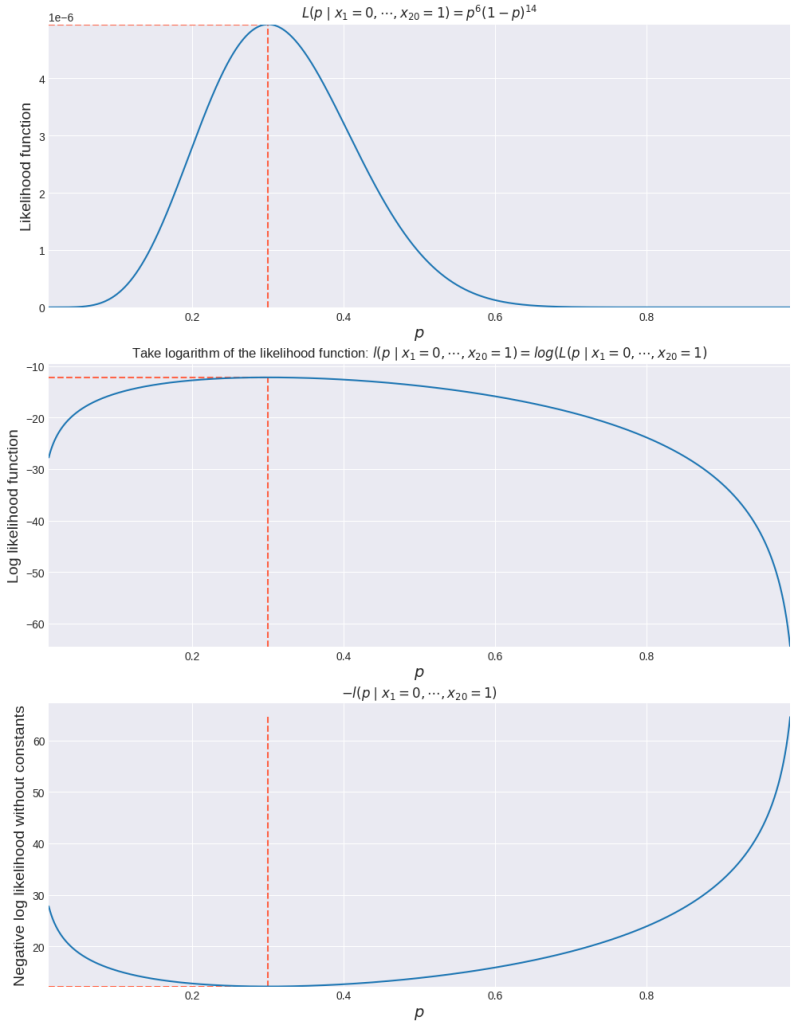
MLE for discrete distributions: two notes

Note 1: the MLE of the parameter p of the Bernoulli distribution is the sample mean of the observed data, i.e.,

$$\hat{p}_{MLE} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{\text{count}(x_i = 1)}{N}$$

MLE for discrete distributions: two notes

Note 2: the likelihood function of the parameter p in the Bernoulli distribution (discrete) is a *continuous* function.



Properties of MLE for interested readers

- Recall the point of learning how to obtain the MLE instead of just "taking the average or something" is so that we have a theoretical framework to analyze the behaviors of these estimates.
- There are three properties I want to mention here:
  - The maximum likelihood estimators are unbiased.
  - Functional invariance: given a transformation g (e.g. $t(x) = x^2$), $t(\hat{\theta}_{MLE}) = MLE(t(\theta))$. For example, if $\hat{\sigma}_{MLE}$ is the MLE for σ, then $\hat{\sigma}^2_{MLE}$ is the MLE for σ^2.
  - Under some regularity conditions (that are usually true in practice), MLE is asymptotically Gaussian. More precisely, for large N, $\hat{\theta}_{MLE}$ approximately follows a Gaussian distribution with mean θ and variance $\frac{1}{I_N(\theta)N}$, where $I_N(\theta)$ is called the **Fisher information**, which is defined as

$$I_N(\theta) = E_\Theta \left[-\frac{d^2}{d\theta^2} l(\theta \mid X_1, \cdots, X_N) \right]$$

    where $l(\theta \mid X_1, \cdots, X_N)$ is the log likelihood.

Note: in this context, "asymptotically" means "it is true when you have a lot of data"

Summary: steps to find the maximum likelihood estimation

Given a model $y = g(x; \mathcal{O} \mid h)$, where $\mathcal{O}$ is a set of parameters

- a) Describe the experiments
- b) Describe the data generated from the experiments
- c) Describe the random variables (typically with i.i.d. assumption)
- d) Choose a parameter of interest $\theta \in \mathcal{O}$
- e) Choose the maximum likelihood estimation as the estimation method:
  Given data $x_1, \cdots, x_N$, find the likelihood function

$$L(\theta \mid x_1, \cdots, x_N) = f_{X_1 \cdots X_N}(x_1, \cdots, x_N; \theta)$$

If the random variables X_i are assumed to be i.i.d., then we have:

$$L(\theta \mid x_1, \cdots, x_N) = \prod_{i=1}^{N} f(x_i; \theta)$$

where f is the PMF for discrete random variables and the PDF for continuous random variables.

Summary: steps to find the maximum likelihood estimation

- f) Compute $\hat{\theta}_{MLE}$ by maximizing the likelihood function:

$$\hat{\theta}_{MLE} = \arg\max_{\theta} L(\theta \mid x_1, \cdots, x_N) = \arg\max_{\theta} f_{X_1 \cdots X_N}(x_1, \cdots, x_N; \theta)$$

Given the i.i.d. assumption, we have

$$\hat{\theta}_{MLE} = \arg\max_{\theta} L(\theta \mid x_1, \cdots, x_N) = \arg\max_{\theta} \prod_{i=1}^{N} f(x_i; \theta)$$

or equivalently, minimizing the negative log likelihood function (easier to compute/avoiding underflow):

$$\hat{\theta}_{MLE} = \arg\min_{\theta} -\sum_{i=1}^{N} \log\left(f(x_i; \theta)\right)$$

  - Simple case, e.g. i.i.d. Gaussian, find the closed-form solution by:
    - Taking the partial derivative with respect to the parameter
    - Setting the derivative to zero
    - Solving for the parameter

    Note: the solutions for many distributions are already available in closed-form if you google them.
  - In general, the estimate needs to be found by iterative methods, e.g. gradient descent

Maximum a posteriori estimation (MAP)

What you will learn from this subsection

Terminology

- A posteriori and a priori
- Bayes' rule
- Parameters as random variables
- Prior distribution and posterior distribution of parameters
- Conjugate priors
- Maximum a posteriori estimation
- Frequentist and Bayesian methods

A posteriori

- Meaning (merriam-webster):

  A posteriori, Latin for "from the latter", is a term from logic, which usually refers to reasoning that works backward from an effect to its causes.

- A posteriori vs a priori:
  - *a priori* statements (aka the **prior**) are claims that come before experience
  - *a posteriori* statements (aka the **posterior**) are claims that come after experience
  - a priori statements + experience $\rightarrow$ a posteriori statements
  - experience: observation of data
- Example:
  - a priori statement: duck number 2 usually swims between 9PM and 10PM
  - experience: duck number 2 is observed wet at 9:30PM 
  - a posteriori statement:

    number 2 usually swims between 9PM and 10PM + number 2 is observed wet at 9:30PM $\rightarrow$ number 2 must have been swimming

  Note: this statement might not be true! For example, it could have rained and number 2 didn't make it home in time. 

Prior and posterior in mathematics

- Recall the Bayes' rule: given events A and B,

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

- Prior and posterior:
  - $P(A)$: the prior - your claim before you observe any data
  - $P(A \mid B)$: the posterior - what you conclude after you observe data
  - B: data
- Example:
  - A: duck number 2 usually swims between 9PM and 10PM
  - B: the observation that number 2 is wet at 9:30PM
  - $P(B)$: the probability of number 2 observed wet at 9:30PM
  - $P(B \mid A)$: the probability of number 2 observed wet at 9:30PM given that it has been swimming between 9PM and 10PM
  - $P(A)$: the probability of number 2 swimming between 9PM and 10PM - prior
  - $P(A \mid B)$: the probability of number 2 swimming between 9PM and 10PM given the observation that it looks wet at 9:30PM - posterior

Prior and posterior for random variables

- For random variables X and Y:

$$f_{X|Y}(x \mid y) = \frac{f_{Y|X}(y \mid x) f_X(x)}{f_Y(y)}$$

  - For continuous random variables, $f_\cdot$ is the PDF
  - For discrete random variables, $f_\cdot$ is the PMF
  - $f_{X|Y}(x \mid y)$, $f_{Y|X}(y \mid x)$ are the *conditional PDF or PMF*.
- Relation to parameter estimation: we want to estimate unknown parameter θ given some data

$$f_{\Theta|data}(\theta \mid data) = \frac{f_{data|\Theta}(data \mid \theta) f_\Theta(\theta)}{f_{data}(data)}$$

Prior and posterior in the context of parameter estimation

$$f_{\Theta|data}(\theta \mid data) = \frac{f_{data|\Theta}(data \mid \theta) f_{\Theta}(\theta)}{f_{data}(data)}$$

- Prior and posterior:
  - $f_{\Theta}(\theta)$: the prior - your assumption before observing any data
    - Here we **assume that θ is a random variable** with PDF/PMF $f_{\Theta}(\theta)$.
  - $f_{data|\Theta}(data \mid \theta)$ and $f_{data}(data)$: calculated from data
  - $f_{\Theta|data}(\theta \mid data)$: the posterior - what you conclude after you observe data
  - prior + data $\rightarrow$ posterior
- Note: $f_{data|\Theta}(data \mid \theta)$ is the likelihood function. Compared to MLE, θ here is a random variable so we use $|$ instead of ; to indicate that it is the conditional PDF/PMF.
- Similar to MLE (where we try to *maximize the likelihood function*), in MAP, we try to *maximize the posterior function* $f_{\Theta|data}(\theta \mid data)$, i.e., given $data = x_1, x_2, \cdots, x_N$,

$$\begin{aligned} \hat{\theta}_{MAP} &= \arg\max_{\theta} f_{\Theta|X_1 \cdots X_N}(\theta \mid x_1, \cdots, x_N) \\ &= \arg\max_{\theta} f_{X_1,\cdots,X_N|\Theta}(x_1, \cdots, x_N \mid \theta) f_{\Theta}(\theta) \end{aligned}$$

  Note: $f_{data}(data) = f_{X_1,\cdots,X_N}(x_1, \cdots, x_N)$ (called **normalization constant**) is not a function of θ and therefore does not contribute to the solution of the optimization problem.
- Compared to page 37, the differences between MAP and MLE: 1) θ is assumed random in MAP; 2) $\hat{\theta}_{MAP}$ is obtained by maximizing the posterior instead of the likelihood function. We will focus on presenting the steps that are different.

Maximum a posteriori estimation

- More formally, given i.i.d. $X_1, X_2, \cdots, X_N$ (same as in MLE),

$$\begin{aligned}\hat{\theta}_{MAP} &= \arg\max_{\theta} f_{\Theta|X_1 \cdots X_N}(\theta \mid x_1, \cdots, x_N) \\ &= \arg\max_{\theta} f_{X_1 \cdots X_N|\Theta}(x_1, \cdots, x_N \mid \theta) f_{\Theta}(\theta) \\ &= \arg\max_{\theta} \prod_{i=1}^{N} f(x_i \mid \theta) f_{\Theta}(\theta)\end{aligned}$$

- Now we come to a point, where we start to know how to compute things:
  - $f(x_i \mid \theta)$: the likelihood - we know how to compute them (same procedure as in MLE)
  - $f_{\Theta}(\theta)$: the prior - it's your assumption of the distribution of θ before observing any data (next slides)

Note: if you are confused by the conditional PDF/PMF, you can think of it as "plug in whatever on the right side of the bar | into the expression and roll with it" because whatever after the | is something that is "given".

Maximum a posteriori estimation (cont.)

- Recall, in MLE (cf. page 38)

$$\hat{\theta}_{MLE} = \arg\min_{\theta} \; -\sum_{i=1}^{N} \log\left(f(x_i;\theta)\right)$$

- Similarly, we can apply the negative log function (cf. page 27)

$$\hat{\theta}_{MAP} = \arg\min_{\theta} -\log \prod_{i=1}^{N} f(x_i \mid \theta) f_{\Theta}(\theta) = \arg\min_{\theta} -\left(\sum_{i=1}^{N} \log\left(f(x_i \mid \theta)\right) + \log f_{\Theta}(\theta)\right)$$

- The solution of $\hat{\theta}_{MAP}$ can be calculated either as a closed-form solution or with iterative techniques.

Choice of the prior $f_\Theta(\theta)$

- There are two things you need to choose to get a meaningful $f_\Theta(\theta)$:
  - 1) choose a family of probability distributions for Θ, which decides the functional form of $f_\Theta(\theta)$, e.g. if we choose Gaussian, then

$$f_\Theta(\theta) = \frac{1}{\sqrt{2\pi s^2}} e^{-\frac{(\theta-m)^2}{2s^2}} \tag{2}$$

  - 2) choose the parameters for $f_\Theta(\theta)$, e.g. m and s in Eq. (2)

Choice of the prior $f_\Theta(\theta)$

How to make these two choices?

$$\underbrace{f_{\Theta|X_1\cdots X_N}(\theta \mid x_1, \cdots, x_N)}_{\text{posterior}} = \frac{f_{X_1\cdots X_N|\Theta}(x_1, \cdots, x_N \mid \theta)}{f_{X_1\cdots X_N}(x_1, \cdots, x_N)} \underbrace{f_\Theta(\theta)}_{\text{prior}} \tag{3}$$

- 1) Choose a family of distributions for $f_\Theta(\theta)$:
  - Sometimes it's given by the problem setup!
  - If it's unknown, we typically use something called a **conjugate prior**.
    - Definition: if the resulting posterior for a given prior has the same functional form as the prior, i.e. they are from the same distribution family, then this prior $f_\Theta(\theta)$ is called the conjugate prior **for the likelihood function** $f_{X_1\cdots X_N|\Theta}(x_1, \cdots, x_N \mid \theta)$. For example, Beta distribution is the conjugate prior for the Bernoulli distribution, i.e.

$$\underbrace{Beta}_{prior} + \underbrace{Bernoulli}_{data\ (likelihood)} \rightarrow \underbrace{Beta}_{posterior}$$

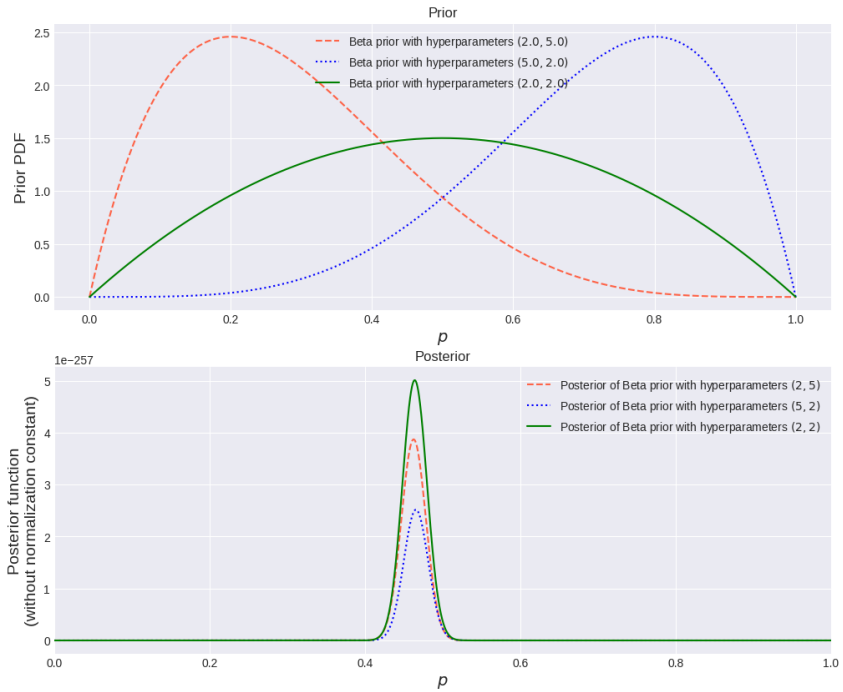
  - How to find conjugate priors for different distributions? - There's a look up table. Note that for each parameter you want to estimate from data, you need to choose a prior.

Parameter to be estimated	distribution (conjugate prior)
Normal μ	Normal(m, s)
Normal σ^2	Inverse Gamma (α, β)
Bernoulli p	Beta (a, b)

  and the list goes on... But not every distribution has a conjugate prior.
  - The next step is to choose the parameters of the priors, e.g. (m, s), (α, β), (a, b), etc.

Choice of the prior $f_\Theta(\theta)$ - conjugate prior (cont.)

- 2) Choose the parameters of the priors:
  - The parameters of the priors are **hyperparameters**.
  - Given enough data, the posterior will converge to the same (true) distribution even for different priors. This is illustrated in the following image. In this example, we estimate p for a Bernoulli distribution with data size 1000. We see that different hyperparameters give us very similar posteriors (figures plotted without the normalization constant for convenience).



Summary: steps to find the maximum a posteriori estimation

Given a model $y = g(x; \mathcal{O} \mid h)$, where $\mathcal{O}$ is a set of parameters

- a) Describe the experiments
- b) Describe the data generated from the experiments
- c) Describe the random variables (typically with i.i.d. assumption)
- d) Choose a parameter of interest $\theta \in \mathcal{O}$
- e) Choose the maximum a posteriori estimation as the estimation method
  - θ **is assumed to be drawn from a random distribution**
  - Choose a prior distribution for θ along with the hyperparameters: $f_\Theta(\theta)$
    - Prior might be known by the problem setup
    - If prior unknown, conjugate priors are typically chosen for various reasons
  - Find the likelihood function: $f_{X|\Theta}(\boldsymbol{x} \mid \theta)$ (same as in MLE)
  - Express the posterior distribution in terms of the prior and the likelihood function

$$f_{\Theta|X}(\theta \mid \boldsymbol{x}) = \frac{f_{X|\Theta}(\boldsymbol{x} \mid \theta) f_\Theta(\theta)}{f_X(\boldsymbol{x})}$$

- f) Compute $\hat{\theta}_{MAP}$ by maximizing the posterior function (or equivalently, minimizing the negative log posterior function without the normalization constant). The optimal solution can be found by a closed-form expression or using iterative techniques.

MLE vs MAP

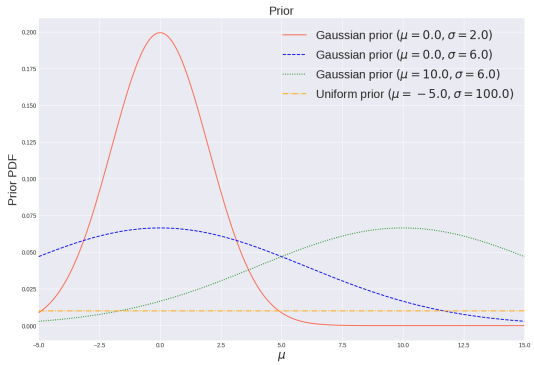
MLE vs MAP

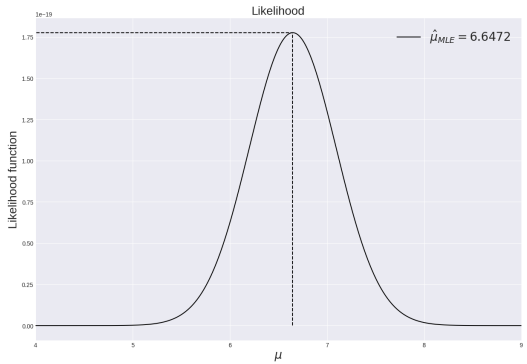
	maximum likelihood estimation	maximum a posteriori estimation
parameter θ	θ is not assumed random (Frequentist)	θ is assumed to be drawn from a random distribution (Bayesian)
hyperparameter	none	prior distribution and its parameters
parameter estimation $\hat{\theta}$	$\hat{\theta}_{MLE} = \arg\min_{\theta} \underbrace{-\log L(\theta \mid data)}_{\text{objective function for MLE}}$	$\hat{\theta}_{MAP} = \arg\min_{\theta} \underbrace{-\log L(\theta \mid data) - \log f_{\Theta}(\theta)}_{\text{objective function for MAP}}$

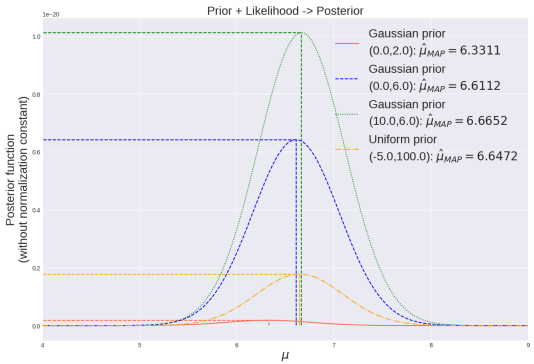
- The function that needs to be minimized is called the **objective function** or **loss**. You will see this a lot in machine learning.
- Objective function for MAP = objective function for MLE + $(-\log f_{\Theta}(\theta))$
  - $(-\log f_{\Theta}(\theta))$ is called the **regularization** or **penalty**.
  - Loosely speaking, when you use too few data points to estimate the unknown parameter, you might have an **overfitting** problem. That is, the estimate of the parameter might look good on the data you derive them from, but they generalize poorly for the prediction tasks on unseen data.
  - One solution is to put a regularization term on the variation of the parameters. It is also called **smoothing**.

MLE vs MAP (cont.)

- More on regularization:
  - Without any regularization, the estimated parameter can have a large variation - if you use a different data set to estimate the parameter, the estimate can end up with a completely different value! Loosely speaking, the regularization term keeps the value of the estimate from going crazy no matter what data set is used.
- MAP pulls the MLE towards its prior - the posterior is biased towards its prior compared to MLE
- MLE and MAP are equivalent if the prior distribution is a uniform distribution on an infinite interval. This is called a **non-informative prior**, because it means that the value of θ can be anything with equal chances.







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 (MLE) and maximum a posteriori estimation (MAP)

Not yet:

- Machine learning models for classification, regression, clustering, etc

Next:

- Classification, naive Bayes classifier

Before next lecture:

- Maximum a posteriori estimation, Bayes' rule, i.i.d. random variables, Bernoulli distribution

Until next time!