

Lecture 5: Parameter Estimation (Part II)

Statistical Methods for Data Science

Yinan Yu

Department of Computer Science and Engineering

November 13, 2023

Today

1 Parameter estimation

- Maximum likelihood estimation (MLE)
- Maximum a posteriori estimation (MAP)
- MLE vs MAP

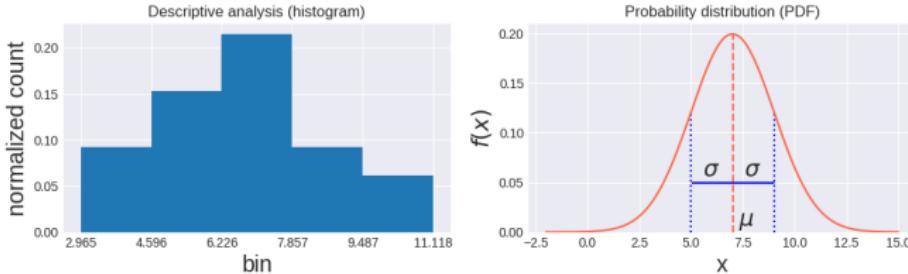
2 Summary

Learning outcome

- Be able to explain these concepts: prior, posterior, Bayes' rule
- Be able to describe the steps of MLE and MAP
- Be able to explain the connections and differences between MLE and MAP

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 continue with question 3.

Today

1 Parameter estimation

- Maximum likelihood estimation (MLE)
- Maximum a posteriori estimation (MAP)
- MLE vs MAP

2 Summary

Recap - 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.

Maximum likelihood estimation (MLE)

Recap - parameter estimation 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**

Conclusion: $\hat{\mu}_{MLE} = 4$

Recap - parameter estimation given more observations

The likelihood function

$$L(\mu | x_1 = 6.98, \dots, x_{20} = 7.27, \sigma = 2) = f(x_1 = 6.98, \dots, x_{20} = 7.27; \mu | \sigma = 2)$$

- **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	...	19	20
weight	6.98	5.43	2.97	7.07	...	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, \dots, X_{20} are independent - [new!] assumption:

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

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

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

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

$$\begin{aligned} & L(\mu | x_1 = 6.98, x_2 = 5.43, \dots, x_{20} = 7.27, \sigma = 2) \\ &= f(x_1 = 6.98, x_2 = 5.43, \dots, x_{20} = 7.27; \mu | \sigma = 2) - \text{joint PDF} \\ &= f(6.98; \mu | \sigma = 2)f(5.43; \mu | \sigma = 2) \cdots f(7.27; \mu | \sigma = 2) - \text{i.i.d. assumption} \\ &= \frac{1}{\sqrt{8\pi}} e^{-\frac{(6.98-\mu)^2}{8}} \frac{1}{\sqrt{8\pi}} e^{-\frac{(5.43-\mu)^2}{8}} \cdots \frac{1}{\sqrt{8\pi}} e^{-\frac{(7.27-\mu)^2}{8}} - \text{plug in data} \\ &= \left(\frac{1}{\sqrt{8\pi}}\right)^{20} e^{-\frac{(6.98-\mu)^2 + \dots + (7.27-\mu)^2}{8}} - \text{something concrete} \end{aligned}$$

- MLE can be computed as:

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

- From page 6, 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

$$\begin{aligned}\hat{\mu}_{MLE} &= \arg \max_{\mu} L(\mu | x_1 = 6.98, x_2 = 5.43, \dots, x_{20} = 7.27, \sigma = 2) \\ &= \arg \max_{\mu} L(\mu | x) \text{ (for convenience)} \\ &= \arg \max_{\mu} \log L(\mu | x) \text{ (does not change the location of the maximum)}$$

Log function turns multiplications into sums, which makes everything much easier to compute

$$\begin{aligned}&= \arg \max_{\mu} I(\mu | x) \text{ (I is called the **log likelihood**)} \\ &= \arg \min_{\mu} -I(\mu | x) \text{ (this is the **negative log likelihood** - mainly for standardization purposes)} \\ &= \arg \min_{\mu} -\log\left(\frac{1}{\sqrt{8\pi}}\right)^{20} + \frac{(6.98 - \mu)^2 + \dots + (7.27 - \mu)^2}{8} \text{ (plug in the expression of the PDFs)}$$

Set the derivative to zero to find the optimal solution

$$\begin{aligned}\iff &\frac{\partial}{\partial \mu} \left(-\log\left(\frac{1}{\sqrt{8\pi}}\right)^{20} + \frac{(6.98 - \mu)^2 + \dots + (7.27 - \mu)^2}{8} \right) = 0 \\ \iff &\sum_{i=1}^{20} x_i - 20\mu = 0 \\ \iff &\hat{\mu}_{MLE} = \frac{1}{20} \sum_{i=1}^{20} x_i = 6.647\end{aligned}$$

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".
- For general cases, the solution typically does not have closed-form and 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 optimality.

What if σ is also unknown?

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

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

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

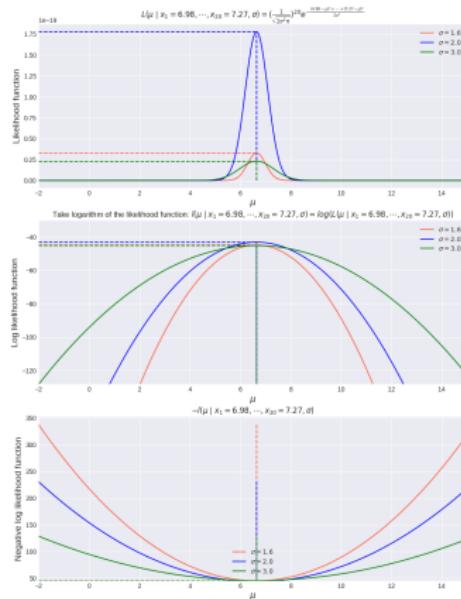
$$\hat{\mu}_{MLE} = \arg \max_{\mu} L(\mu, \sigma | \mathbf{x}) \iff \frac{\partial}{\partial \mu} L(\mu, \sigma | \mathbf{x}) = 0$$

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

Note: left as an exercise.

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

$\hat{\mu}_{MLE}$ 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 $f(x)$:

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

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	...	19	20
color	red	blue	red	red	...	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:

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

- Negative log likelihood function:

$$-I(p | \mathbf{x}) = -\log(p^6(1-p)^{14}) = -(6\log(p) + 14\log(1-p))$$

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

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

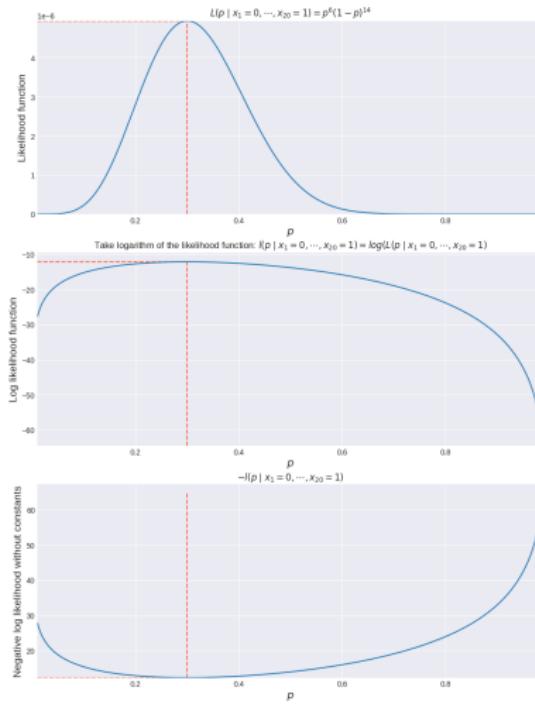
MLE for discrete distributions: two notes

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

- The point of learning how to obtain the MLE instead of just "taking the average or something" is to have a theoretical framework to analyze the behaviors of these estimates.
- There are two example properties:
 - The maximum likelihood estimators are unbiased.
 - 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

$$I_N(\theta) = E_{\Theta} \left[-\frac{d^2}{d\theta^2} I(\theta | X_1, \dots, X_N) \right]$$

where $I(\theta | X_1, \dots, X_N)$ 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 $y = g(x; \mathcal{O} | h)$, 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) Identify parameters of interest $\theta \in \mathcal{O}$
- e) Choose the maximum likelihood estimation as the estimation method:

Given data x_1, \dots, x_N , find the likelihood function

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

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

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

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:

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

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

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

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

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

- 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 → *a posteriori* statements
- experience: observation of data

- Example:

- *a priori* statement: you claim that duck number 2 enjoys swimming between 9PM and 10PM



- experience: duck number 2 is observed wet at 9:30PM

- *a posteriori* statement:

number 2 enjoys swimming between 9PM and 10PM + number 2 is observed wet at 9:30PM → 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 ,

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

- Prior and posterior:
 - $P(A)$: the prior - your claim before you observe any data
 - $P(A | B)$: the posterior - what you conclude after you observe data
 - B : events observed from data
- Example:
 - A : duck number 2 enjoys swimming 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
 - $P(B | A)$: 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
 - $P(A | B)$: 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 :

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

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

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

Prior and posterior in the context of parameter estimation

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

- 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 $f_{\Theta}(\theta)$.
 - $f_{data|\Theta}(data | \theta)$ and $f_{data}(data)$: calculated from data
 - $f_{\Theta|data}(\theta | data)$: the posterior - what you conclude after you observe data
 - prior + data \rightarrow posterior
- Note: $f_{data|\Theta}(data | \theta)$ is the likelihood function. Compared to MLE, θ here is modeled as a random variable in MAP 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* $f_{\Theta|data}(\theta | data)$, i.e., given $data = x_1, x_2, \dots, x_N$,

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

Note: $f_{data}(data) = f_{x_1, \dots, x_N}(x_1, \dots, x_N)$ (called **normalization constant**) is not a function of θ and therefore does not contribute to the solution of the optimization problem.

- Compared to page 21, 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, \dots, X_N (same as in MLE),

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

- Now we come to a point, where we start to know how to compute things:
 - $f(x_i | \theta)$: 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 22)

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

- Similarly, we can apply the negative log function (cf. page 11)

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

- 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

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

- 2) choose the parameters for $f_{\Theta}(\theta)$, e.g. m and s in Eq. (1)

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

How to make these two choices?

$$\underbrace{f_{\Theta|x_1 \dots x_N}(\theta | x_1, \dots, x_N)}_{\text{posterior}} = \frac{f_{x_1 \dots x_N|\Theta}(x_1, \dots, x_N | \theta)}{f_{x_1 \dots x_N}(x_1, \dots, x_N)} \underbrace{f_{\Theta}(\theta)}_{\text{prior}} \quad (2)$$

- 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** $f_{x_1 \dots x_N|\Theta}(x_1, \dots, x_N | \theta)$. For example, Beta distribution is the conjugate prior for the Bernoulli distribution, i.e.

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

- 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)

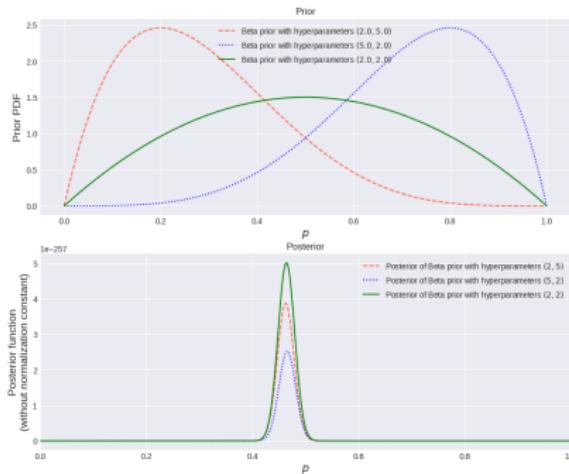
and the list goes on... But not every distribution has a conjugate prior.

- 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, MAP returns the same estimate 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).



Summary: steps to find the maximum a posteriori estimation

Given a model $y = g(x; \mathcal{O} | h)$, 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) Identify parameters 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: $f_{X|\Theta}(x | \theta)$ (same as in MLE)
 - Express the posterior distribution in terms of the prior and the likelihood function

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

- 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

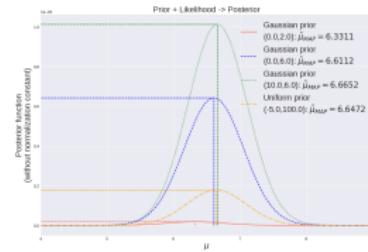
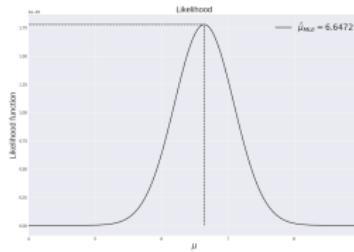
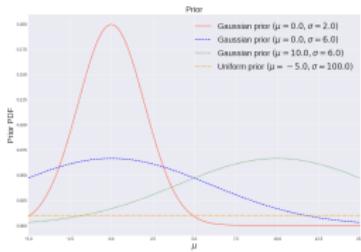
MLE vs MAP

	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 $\hat{\theta}$	$\hat{\theta}_{MLE} = \arg \min_{\theta} \underbrace{-\log L(\theta data)}_{\text{objective function for MLE}}$	$\hat{\theta}_{MAP} = \arg \min_{\theta} \underbrace{-\log L(\theta data) - \log f_{\Theta}(\theta)}_{\text{objective function for MAP}}$

- 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

1 Parameter estimation

2 Summary

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:

- How to apply probabilistic modeling and these concepts?

Next:

- Classification, naive Bayes classifier

Before next lecture:

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

This level of cuteness is a bit too much...

