Lecture 5: Parameter Estimation (Part II)

Statistical Methods for Data Science

Yinan Yu

Department of Computer Science and Engineering

November 18, 2024

Today

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

Today

What you will learn from this section

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: 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 $\rightarrow$ 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: 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
  - $P(B \mid A)$: 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 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* $f_{\Theta|data}(\theta \mid data)$, 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.
- 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

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

- Similarly, we can apply the negative log function

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

- 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

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

  - 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?

$$\underbrace{f_{\Theta|X_1\cdots X_N}(\theta \mid x_1, \cdots, x_N)}_{\text{posterior}} = \frac{f_{X_1\cdots X_N|\Theta}(x_1, \cdots, x_N \mid \theta)}{f_{X_1\cdots X_N}(x_1, \cdots, x_N)} \underbrace{f_\Theta(\theta)}_{\text{prior}} \tag{2}$$

- 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** $f_{X_1\cdots X_N|\Theta}(x_1, \cdots, x_N \mid \theta)$. For example, Beta distribution is the conjugate prior for the Bernoulli distribution, i.e.

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

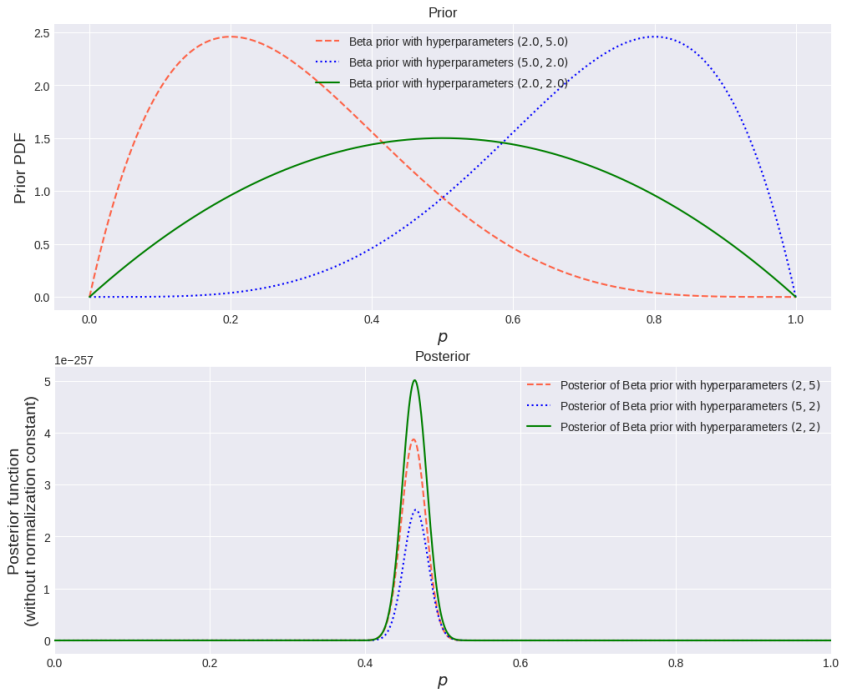
  - How to find conjugate priors for different distributions? - There's a look up table. Note that for each parameter you want to estimate from data, you need to choose a prior.

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 $y = g(x; \mathcal{O} \mid h)$, 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) 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: $f_{X|\Theta}(\boldsymbol{x} \mid \theta)$ (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.

Today

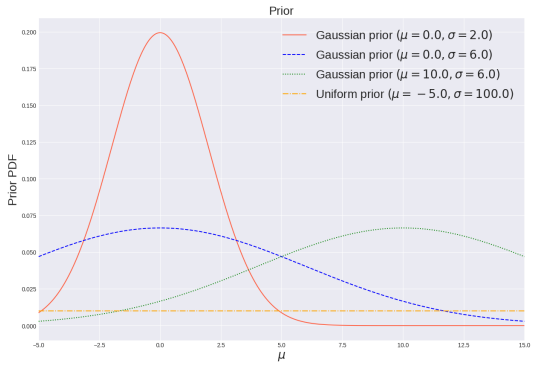
MLE vs MAP

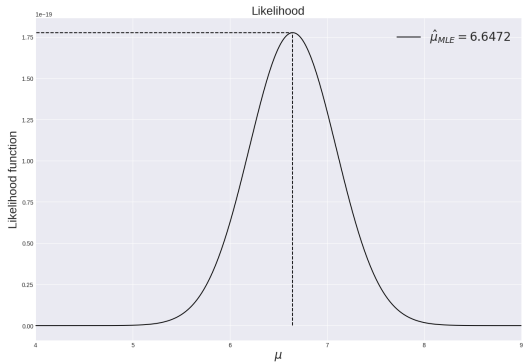
	maximum likelihood estimation	maximum a posteriori estimation
parameter θ	θ is not assumed random (Frequentist)	θ is assumed to be drawn from a random distribution (Bayesian)
hyperparameter	none	prior distribution and its parameters
parameter estimation $\hat{\theta}$	$\hat{\theta}_{MLE} = \arg\min_{\theta}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

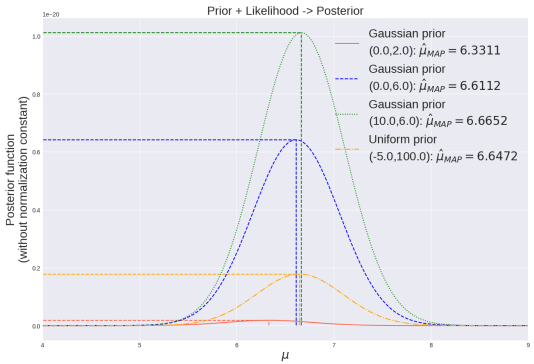
- The function that needs to be minimized is called the **objective function** or **loss**. You will see this a lot in machine learning.
- Objective function for MAP = objective function for MLE + $(-\log f_{\Theta}(\theta))$
  - $(-\log f_{\Theta}(\theta))$ is called the **regularization** or **penalty**.
  - 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:

- How to apply probabilistic modeling and these concepts in practice.

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...