Machine Learning and NeuroEngineering

机器学习与神经工程

Lecture 9 – Bayesian Parameter Estimation

Quanying Liu （刘泉影）

SUSTech, BME department

Email: liuqy@sustech.edu.cn

BMEB212, Spring term, 2021

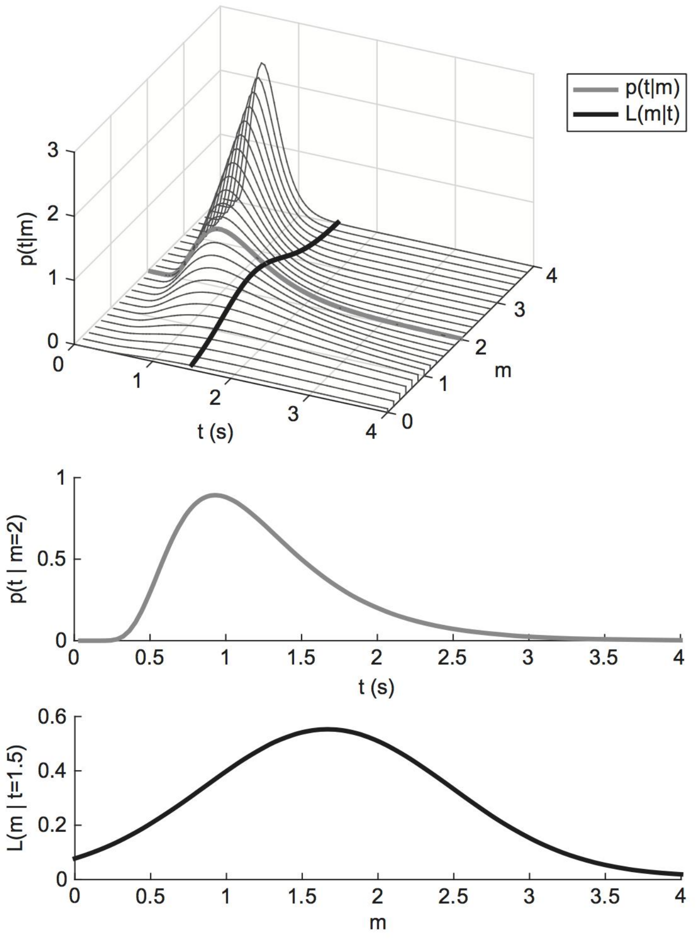
Lecture 9 – Bayesian Parameter Estimation

- What is Bayesian Inference?
  - Motivations
  - Bayes Theorem: prior, likelihood, evidence, posterior
- Analytic Methods for Obtaining Posteriors
  - The likelihood function
  - The Prior Distribution
  - The Posterior Distribution
  - An example: Estimating the bias of a coin
- Determining the Prior Distributions of Parameters
  - Non-Informative Priors
  - Reference Priors
- LSE vs MLE vs MAP
- Tutorial of MNE-python for EEG analysis (thanks to 林沛阳)

Motivations

- We have learned Maximum Likelihood Estimation (**MLE**) in Lecture 5.
- Recall **MLE**: Probability distribution $f(y|\boldsymbol{\theta}, M)$

- It permits to estimate parameter value by ***relative*** comparisons between different parameter values. But it is not suited for estimating ***absolute*** probabilities.
- It did **not** combine our prior knowledge of the parameters.



Motivations for Bayesian Parameter Estimation:

- We want to know is the likely range of the "true" parameter value that we can infer from our measurements (information about the probability distribution of the parameters)
- We want to incorporate our prior knowledge of the parameters into the consideration.

Bayes Theorem

Prior – what we know about $\boldsymbol{\theta}$ **BEFORE** collecting data y

Likelihood – propensity for observing a certain value of y given a certain value of $\boldsymbol{\theta}$

$$\boxed{p(\boldsymbol{\theta}|y) = \frac{p(\boldsymbol{\theta})p(y|\boldsymbol{\theta})}{p(y)} = \frac{p(\boldsymbol{\theta})p(y|\boldsymbol{\theta})}{\sum_{\boldsymbol{\theta}} p(\boldsymbol{\theta})p(y|\boldsymbol{\theta})}}$$

Posterior – what we know about y **AFTER** seeing x

Evidence – a **constant** to ensure that the left hand side is a valid distribution

$$\boxed{p(\boldsymbol{\theta}|y) \propto p(\boldsymbol{\theta})p(y|\boldsymbol{\theta})}$$

4 components

$$p(\boldsymbol{\theta}|y) = \frac{p(\boldsymbol{\theta})p(y|\boldsymbol{\theta})}{p(y)} = \frac{p(\boldsymbol{\theta})p(y|\boldsymbol{\theta})}{\sum_{\boldsymbol{\theta}} p(\boldsymbol{\theta})p(y|\boldsymbol{\theta})}$$

Prior $p(\boldsymbol{\theta})$	Our knowledge of our model's parameters **before** we collect data in an experiment. $\boldsymbol{\theta}$ is the probability of *head* for a coin an example of discrete prior $\boldsymbol{\theta}$ is $p(\boldsymbol{\theta} = 0.5) = 0.8$, $p(\boldsymbol{\theta} = 1) = 0.2$ an example of continuous prior $\boldsymbol{\theta}$ is Unif (0,1)
Likelihood $p(y|\boldsymbol{\theta})$ or $L(\boldsymbol{\theta}|y)$	Having collected data y in an experiment, we can now examine the probability of having obtained **a particular outcome** in light of the prior values of the parameters, $\boldsymbol{\theta}$. An example: Bin(y=5 | n=10, $\boldsymbol{\theta}$)
Evidence $p(y)$	The overall probability of the data, irrespective of the values of parameters. $\sum_{\boldsymbol{\theta}} p(\boldsymbol{\theta})p(y|\boldsymbol{\theta}) = p(\boldsymbol{\theta} = 0.5)\text{Bin}(y=5 \mid n=10, \boldsymbol{\theta} = 0.5) + p(\boldsymbol{\theta} = 1)\text{Bin}(y=5 \mid n=10, \boldsymbol{\theta} = 1)$
Posterior $p(\boldsymbol{\theta}|y)$	The posterior probability of $\boldsymbol{\theta}$, is a result of the application of Bayes theorem. It can answer all sorts of interesting questions: the mode/mean of $\boldsymbol{\theta}$

The likelihood function

- Bernoulli $f(x = k|\theta) = \theta^k (1-\theta)^{1-k}$
- Categorical $f(x = e_k|\boldsymbol{\theta}) = \theta_k$
- Binomial $f(x = k|n, \theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k}$
- Poisson $f(x = k|\theta) = e^{-\theta} \frac{\theta^k}{k!}$
- Gaussian $f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, $\boldsymbol{\theta} = [\mu, \sigma^2]$

The prior distribution

- The assumption of parameters $\boldsymbol{\theta}$ is captured by prior distribution. The parameters of the prior are called **hyper-parameters.**
- In Bayesian probability theory, if the posterior distributions p(θ | x) are in **the same probability distribution family** as the prior probability distribution p(θ), the prior and posterior are then called *conjugate distributions*, and the prior is called ***a conjugate prior*** for the likelihood function p(x | θ).
- Bernoulli and binomial likelihood has Beta conjugate prior

$$\mathrm{Beta}(\theta|a,b) \propto \theta^{a-1}(1-\theta)^{b-1}$$

$$\text{mode} = \frac{a-1}{a+b-2} \qquad \text{mean} = \frac{a}{a+b} \qquad \text{var} = \frac{ab}{(a+b)^2(a+b+1)}$$

```
curve(dbeta(x, 2, 4),ylim=c(0,6),ylab="Probability ↩
    Density",las=1)
curve(dbeta(x, 8, 16),add=TRUE,lty="dashed")
legend("topright",c("Johnnie","Jane"), ↩
    inset=.05,lty=c("solid","dashed"))
```

The Posterior distribution

- The posterior distribution is proportional to prior times likelihood

$$p(\boldsymbol{\theta}|y) \propto p(\boldsymbol{\theta})p(y|\boldsymbol{\theta})$$

- Beta prior $\qquad \mathrm{Beta}(\theta|a,b) \propto \theta^{a-1}(1-\theta)^{b-1}$

- Binomial likelihood $\mathrm{Bin}(y=k|n,\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$

- Posterior distribution

$$\begin{aligned} p(\theta|y) \propto p(\theta)p(y|\theta) &= \mathrm{Beta}(\theta|a,b)\mathrm{Bin}(y=k|n,\theta) \\ &\propto \theta^{a-1}(1-\theta)^{b-1}\,\theta^k(1-\theta)^{n-k} \\ &\propto \theta^{a+k-1}(1-\theta)^{\mathrm{b}+n-k-1} \\ &= \boldsymbol{\mathrm{Beta}(\theta|a+k,b+n-k)} \end{aligned}$$

Example: **Estimating the bias of a coin**

- Beta prior $\mathrm{Beta}(\theta|a,b) \propto \theta^{a-1}(1-\theta)^{b-1}$
- Binomial likelihood $\mathrm{Bin}(y=k|n,\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$
- Posterior distribution $p(\theta|y=k) = \mathrm{Beta}(\theta|a+k, b+n-k)$

- We do an experiment, and toss a coin **10** times
- We observe **6** heads, **4** tails.
- (θ is probability of head) What is the MLE of θ?

$$\underset{\theta}{\mathrm{argmax}}\,\mathrm{Bin}(y=6|10,\theta)$$

- What is the posterior distribution of θ?
  1) Assuming the prior distribution of θ as $\mathrm{Beta}(\theta|1,1)$
  2) Assuming the prior distribution of θ as $\mathrm{Beta}(\theta|12,12)$

Example: **Estimating the bias of a coin**

- Beta prior $\mathrm{Beta}(\theta|a,b) \propto \theta^{a-1}(1-\theta)^{b-1}$
- Binomial likelihood $\mathrm{Bin}(y=k|n,\theta) = \binom{n}{k}\theta^{k}(1-\theta)^{n-k}$
- Posterior distribution $p(\theta|y=k) = \mathrm{Beta}(\theta|a+k, b+n-k)$

- We do an experiment, and toss a coin **10** times
- We observe **6** heads, **4** tails.
- (θ is probability of head) What is the MLE of θ?

$$\underset{\theta}{\operatorname{argmax}}\, \mathrm{Bin}(y=6|10,\theta)$$

- What is the posterior distribution of θ?
  1) Assuming the prior distribution of θ as $\mathrm{Beta}(\theta|1,1)$, the posterior is $\mathrm{Beta}(\theta|7,5)$
  2) Assuming the prior distribution of θ as $\mathrm{Beta}(\theta|12,12)$, the posterior is $\mathrm{Beta}(\theta|19,16)$

Example: **Estimating the bias of a coin**

- Beta prior $\mathrm{Beta}(\theta|a,b) \propto \theta^{a-1}(1-\theta)^{b-1}$
- Binomial likelihood $\mathrm{Bin}(y=k|n,\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$
- Posterior distribution $p(\theta|y=k) = \mathrm{Beta}(\theta|a+k, b+n-k)$

- We do an experiment, and toss a coin **10** times
- We observe **6** heads, **4** tails.
- What is the posterior distribution of θ?
    1) Assuming the prior distribution of θ as $\mathrm{Beta}(\theta|1{,}1)$, the posterior is $\mathrm{Beta}(\theta|7{,}5)$
    2) Assuming the prior distribution of θ as $\mathrm{Beta}(\theta|12{,}12)$, the posterior is $\mathrm{Beta}(\theta|19{,}16)$
    3) Assuming the prior distribution of θ as $\mathrm{Beta}(\theta|1{,}4)$
    4) Assuming the prior distribution of θ as $\mathrm{Beta}(\theta|4{,}1)$
    5) Assuming the prior distribution of θ as $\mathrm{Beta}(\theta|100{,}100)$

Plotting the posterior

- Assuming the prior distribution of θ as $\text{Beta}(\theta|12,12)$
- We do experiments.
- We observe:
  1) 14 heads out of 26 tosses
  2) 113 heads out of 213 tosses
  3) 1130 heads our of 2130 tosses
- What is the posterior distribution of θ for each experiment?

- SEE CODE: Lecture9_plotBeta.r

- How Biased Is the Coin?
- **qbeta(c(0.025,0.975),1130,1000)** returns the upper and lower bounds of a 95% *credible* interval for θ

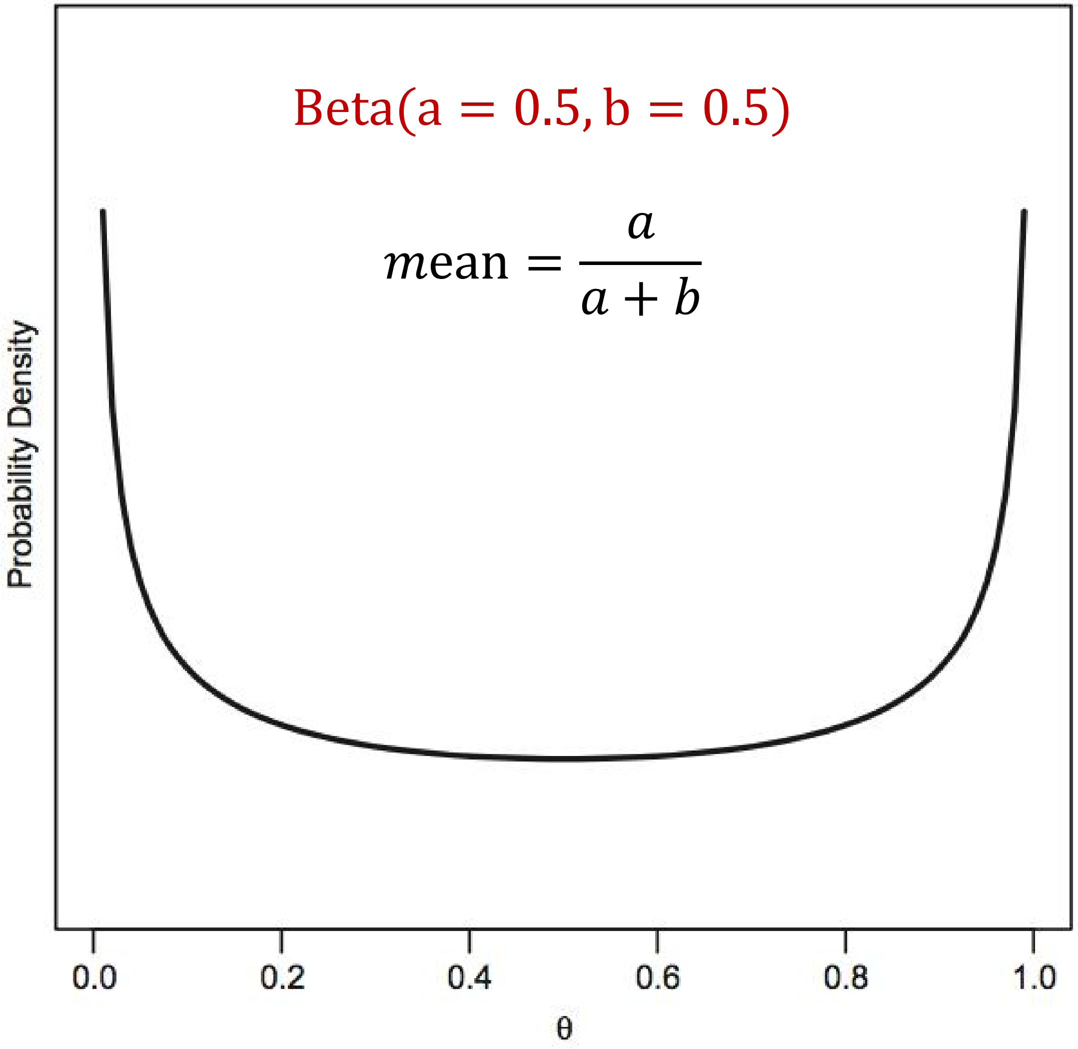
Determining the Prior Distribution

- The combining of *prior knowledge* with *new evidence* to revise one's knowledge is at the heart of Bayesian reasoning, statistics, and modeling.
- How to determine the prior distribution of parameters?

1) **Non-Informative Priors** (to be completely ignorant)

   - If we mean "all values of θ are equally likely", then a uniform prior over θ
   - if we mean that "all orders of magnitude of θ are equally likely", then a logarithmic prior – that is, a distribution that is uniform over log(θ) instead of over θ
   - If we mean the prior distribution is not invariant across different parameterizations of the same problem or model, then $\mathrm{Beta}(a = 0.5, b = 0.5)$

Jeffreys Prior (a non-informative prior)



See derivation in Zhu and Lu, 2004

This is a bit counter-intuitive.
It seems a strong prior expectation of the outcome being at 0 or 1.

Let's do some comparisons.
Jeffreys prior: beta(0.5, 0.5)
Uniform prior: beta(1,1)

(1) **k=0, n=10, MLE=0**
Beta(0.5, 10.5), mean=0.045
Beta(1, 11), mean=0.083

(2) **k=10, n=10, MLE=1.0**
Beta(10.5, 0.5), mean=0.95
Beta(2, 10), mean=0.92

Determining the Prior Distribution

1) **Jeffreys Prior** (to be completely ignorant)

- If we mean the prior distribution is not invariant across different parameterizations of the same problem or model, then $\mathrm{Beta}(a = 0.5, b = 0.5)$

2) **Reference Priors**

- An noninformative prior maximizes the dominance of the data over our prior knowledge
- Reference priors formalize this idea by seeking to maximize some measure of divergence between the posterior and the prior distribution in light of the data.
- The **greater** this divergence, the **less** the prior distribution has mattered.
- Difficulty: it is defined with respect to the data.

Berger et al 2009

LSE vs MLE vs MAP

- Least Square Estimation (LSE): argmin $\mathcal{L}(\boldsymbol{\theta})$
- Maximum Likelihood Estimation (MLE): argmax $\mathrm{P}(X|\boldsymbol{\theta})$
- Maximum A Posterior (MAP): argmax $\mathrm{P}(\boldsymbol{\theta}|X)$

Take **linear regression** as an example…

$$\mathrm{y} = w^T x + \varepsilon$$

LSE = MLE (when noise is Gaussian)

LSE + L2-norm regularization = MAP (when noise is Gaussian and prior is Gaussian)

Reading materials

Textbooks

- Chapter 6 (Bayesian Parameter Estimation)

Extra reading

- MLaPP– chapter 3 (Generative models for discrete data)

Video

1. 机器学习 - 白板推导 P2: 频率学派vs贝叶斯学派
2. 机器学习 - 白板推导 P9-P12: regression and ridge regression

https://www.bilibili.com/video/BV1aE411o7qd?p=2

Summary – Bayesian Parameter Estimation

- What is Bayesian Inference?
  - Motivations
  - Bayes Theorem: prior, likelihood, evidence, posterior
- Analytic Methods for Obtaining Posteriors
  - The likelihood function
  - The Prior Distribution
  - The Posterior Distribution
  - An example: Estimating the bias of a coin
- Determining the Prior Distributions of Parameters
  - Non-Informative Priors
  - Reference Priors
- LSE vs MLE vs MAP
- **Tutorial of MNE-python for EEG analysis (thanks to 林沛阳)**

Homework 2

EEG data processing with MNE-python

DDL: March 29, 2021

Tips:

- Watch the tutorial video
  https://www.bilibili.com/video/BV1YK411T7H8
- Read the official website of MNE-python
  https://mne.tools/stable/index.html

Requirements

1. Read the paper https://www.nature.com/articles/s41597-020-0535-2
2. Download the raw EEG data from [26] in the paper (https://doi.org/10.7910/DVN/RBN3XG)
   学生ID为奇数的同学下载sub-001；学生ID为偶数的同学下载sub-002
3. Plot the **time course** of raw EEG signals with 10-second window (as Figure 4 in the paper).
4. Data preprocessing, artifacts removal
5. Plot the **time course** of preprocessed EEG signals
6. Plot the **time-frequency maps** of the subject (as Figure 6 in the paper)
7. Plot the **topographical distribution** of power of the subject (as Figure 7 in the paper)
8. **Comparison** of power (in dB) changes with time (in s) during hand, elbow motor imagery, and resting state for electrode C3, and electrode C4 (as Figure 8 in the paper)