Pattern Recognition

Lecture 9. Generative Methods I: Parametric methods: Maximum *a Posteriori* Probability Estimation (MAP) Practice

Dr. Shanshan ZHAO

shanshan.zhao@xjtlu.edu.cn

School of AI and Advanced Computing

Xi'an Jiaotong-Liverpool University

Academic Year 2021-2022

Table of Contents

Notations

- X : The dataset observed
- $\mathbf{x}$: the random variable, i.e., the feature vector
- x : the univariant , or a random variable in the feature vector
- θ : the parameters unknown in $p(\mathbf{x})$
- N : Number of samples
- $p(\theta|X)$ or $p(X|\theta)$: we consider θ and X as two random variables, this is to denote conditional probability
- $p(\mathbf{x}_k; \theta)$: The semicolon means that it is the pdf with respect to $\mathbf{x}_k$, ($\mathbf{x}_k$ is the argument of function p), the parameter of it is θ .

Random variable *VS* Parameter

- Both Random variable and Parameter vary with some conditions.
- A 'variable' is something you measure when collecting data
- A 'parameter' is the link between variables

$p(x;\theta)$ *VS* $p(x|\theta)$

- $p(x;\theta)$: It is to denote a function p, the argument is x, the parameter of function is θ
- $p(x|\theta)$: It is to represent a conditional probability
- $L(\theta|D)$: The vertical bar might also be used when describing the likelihood
- Basically, vertical bar is to demonstrate the conditional relationship between two variables; semicolon to distinguish the argument and the parameter.

Conditional probability *VS* Likelihood *VS* Likelihood function

- Likelihood not a probability, but is **proportional to a probability**.
- The likelihood of a hypothesis (H) given some data (D) is proportional to the probability of obtaining D given that H is true, multiplied by an arbitrary positive constant (K).
  In other words, $L(H|D) = K \times P(D|H)$.
  - **L(H|D):** likelihood
  - **P(D|H):** conditional probability
  - **p(D;H) or L(D;H) or L(D):** (likelihood) function p with respect to D. In other words, D is the argument of function p. H is the pameter of p.
- Since a likelihood isn't actually a probability it doesn't obey various rules of probability. For example, likelihood need not sum to 1.

https://alexanderetz.com/2015/04/15/understanding-bayes-a-look-at-the-likelihood/

ML *VS* MAP estimate

- **ML estimate**

  In ML, we use the likelihood function

$$L(\theta|X) = P(X;\theta) = \prod_{k=1}^{N} p(x_k;\theta) \tag{1}$$

  It is proportional to the conditional probability (or density) $P(X|\theta)$
  ML estimates θ so that the likelihood function takes its maximum value, that is,

$$\hat{\theta}_{ML} = arg\max_{\theta} \prod_{k=1}^{N} p(x_k;\theta) \equiv \max_{\theta} p(X|\theta) \tag{2}$$

- **MAP estimate**

$$\hat{\theta}_{MAP} = \max_{\theta}[p(X|\theta)p(\theta)] \tag{3}$$

  which is equivalent to

$$\hat{\theta}_{MAP} = \max_{\theta}[\ln p(X|\theta) + \ln p(\theta)] \tag{4}$$

Frequentist *VS* Bayesian

- https://www.youtube.com/watch?v=r76oDIvwETI
- https://www.youtube.com/watch?v=GEFxFVESQXc&t=299s
- https://www.youtube.com/watch?v=7-Ud4nyHO_Q

ML *VS* MAP estimate

- Maximum likelihood is a special case of Maximum A Posterior estimation. To be specific, MLE is what you get when you do MAP estimation using a uniform prior.
- Both methods come about when we want to answer a question of the form: “What is the probability of scenario Y given some data, X, i.e. $P(Y|X)$.

A question of this form is commonly answered using Bayes’ Law.

$$\underbrace{P(Y|X)}_{\text{posterior}} = \frac{\overbrace{P(X|Y)}^{\text{likelihood}}\overbrace{P(Y)}^{\text{prior}}}{\underbrace{P(X)}_{\text{probability of seeing the data}}}.$$

ML *VS* MAP estimate

Boring but useful

- **MLE** If we're doing Maximum Likelihood Estimation, we do not consider prior information (another way of saying "we have a uniform prior") . In this case, the above equation reduces to

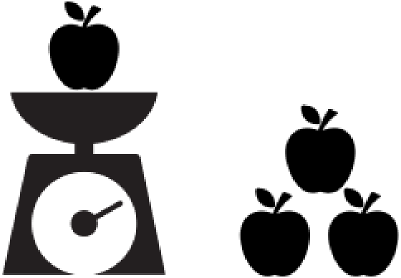
$$P(Y|X) \propto P(X|Y) \tag{5}$$

  In this scenario, we can fit a statistical model to correctly predict the posterior, P(Y|X), by maximizing the likelihood, P(X|Y). Hence "Maximum Likelihood Estimation."

- **MAP** If we know something about the probability of Y, we can incorporate it into the equation in the form of the prior, P(Y). In This case, Bayes' laws has it's original form.
  We then find the posterior by taking into account the likelihood and our prior belief about Y. Hence "Maximum A Posterior".

ML *VS* MAP estimate *example*

Let's say you have a barrel of apples that are all different sizes. You pick an apple at random, and you want to know its weight. Unfortunately, all you have is a broken scale.



ML *VS* MAP estimate *example*

(a)

- For the sake of this example, lets say you know the scale returns the weight of the object with an error of +/- a standard deviation of 10g. We can describe this mathematically as:

$$measurement = weight + error \tag{6}$$

$$error \sim \mathcal{N}(0, 10g) \tag{7}$$

- Let's also say we can weigh the apple as many times as we want, so we'll weigh it 100 times.
- Notice that here the 'weight' is the 'parameter' that we are going to estimate.
- The 'measurement' corresponds to the data 'x'.

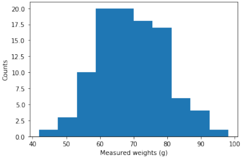
We can view it in this way

$$error = measurement - weight = x - \mu \tag{8}$$

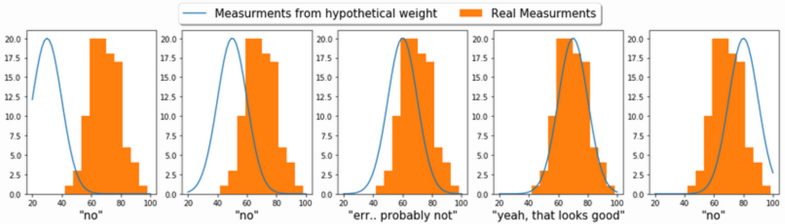
$$p(x; \mu) = \mathcal{N}(x - \mu, 10g) \tag{9}$$

ML *VS* MAP estimate *example*

We can look at our measurements by plotting them with a histogram



An intuitive way to show how to find the value of the 'weight' that can fit the data best.



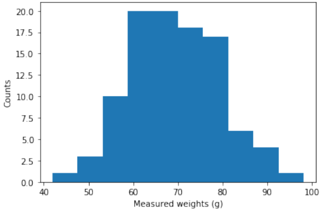
ML *VS* MAP estimate *example*

We know that the ML estimation of a Gaussian is the average of the samples

$$\mu = \frac{1}{N}\sum_{i}^{N} x_i \tag{10}$$

$$SE = \frac{\sigma}{\sqrt{N}} = 10/\sqrt{100} = 1 \tag{11}$$

where, SE is the standard error of the samples in statistics. The weight of the apple is (69.62 +/- 1.) g



ML *VS* MAP estimate *example*

(b)

Now lets say we don't know the error of the scale. We know that its additive random normal, but we don't know what the standard deviation is

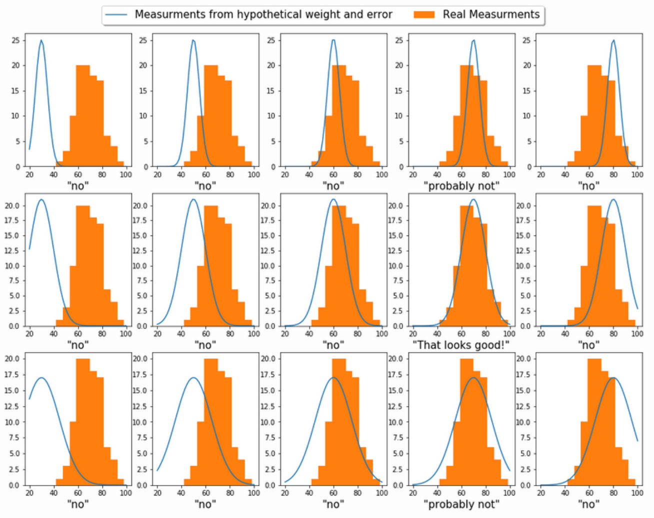
$$\textit{measurement} = \textit{weight} + \textit{error} \tag{12}$$

$$\textit{error} \sim \mathcal{N}(0, \sigma) \tag{13}$$

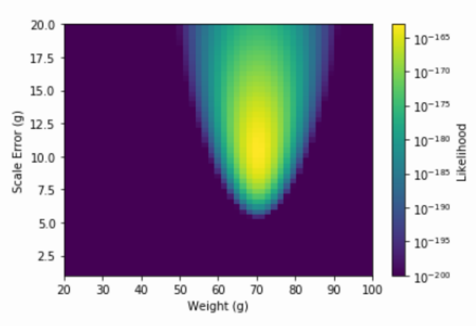
we want to find the mostly likely weight of the apple and the most likely error of the scale

$$P(\mu, \sigma | X) \propto P(X | \mu, \sigma) \tag{14}$$

ML *VS* MAP estimate *example*



ML *VS* MAP estimate *example*

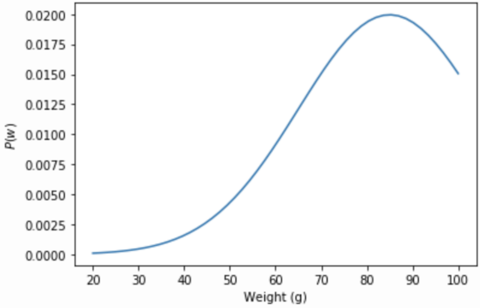


The maximum point will then give us both our value for the apple's weight and the error in the scale.
The weight of the apple is (69.39 +/- .97) g
(you may get a different value or figure in the exercise)

ML *VS* MAP estimate *example*

(c) We have prior on the weight:

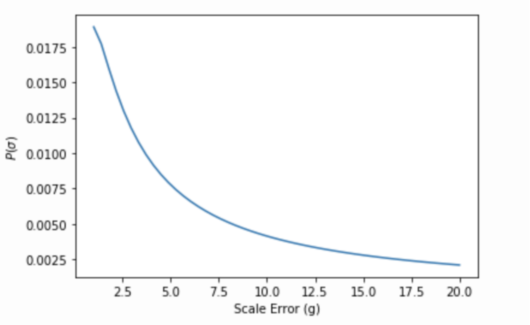
$$P(\mu) = \mathcal{N}(85, 40) \tag{15}$$



ML *VS* MAP estimate *example*

We have prior on the error:

$$P(\sigma) = Inv[Gamma(.05)] \tag{16}$$



ML *VS* MAP estimate *example*

$$P(\mu, \sigma|X) \propto P(X|\mu, \sigma)P(\mu, \sigma) \tag{17}$$

$$P(\mu, \sigma) = P(\mu)P(\sigma) \tag{18}$$

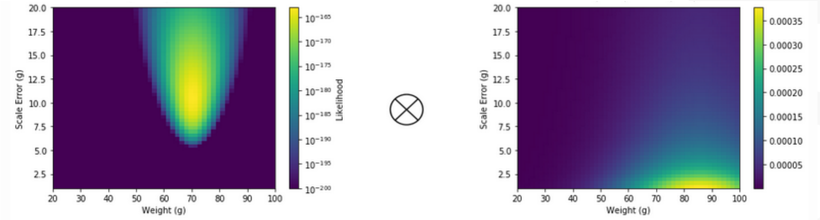
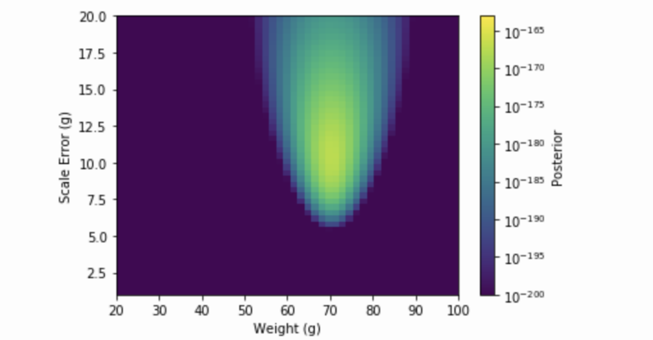


Figure: left: $P(X|\mu, \sigma)$; right: $P(\mu)P(\sigma)$

ML *VS* MAP estimate *example*

The weight of the apple is (69.39 +/- .97) g
(you may get a different value or figure in the exercise)



Reference I

Thank You !

$\mathcal{Q}$ & $\mathcal{A}$