Pattern Recognition

Lecture 08(a). Parametric methods: Maximum *a Posteriori* Probability Estimation (MAP)

Dr. Shanshan ZHAO

School of AI and Advanced Computing
Xi'an Jiaotong-Liverpool University

Academic Year 2023-2024

Notations

- X : The dataset observed
- 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 $p(\mathsf{x})$
- N : Number of samples
- $p(\theta|X)$ or $p(X|\theta)$: we consider θ and X as two random variables, this is to denote the dependence between variables
- $p(\mathsf{x}_k;\theta)$: The semicolon means that it is the pdf with respect to x_k, the parameter of it is θ .

Revisit MLE

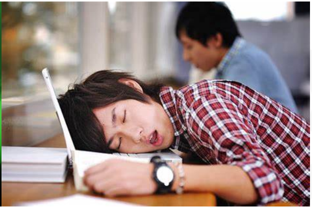
data: 'you didn't do the homework'

possible parameters:

- dog ate your homework
- abducted by aliens
- too lazy

(a)

(b)



(c)

All parameters explain the data!

Maximum *a Posteriori* Probability Estimation (MAP)

- In ML estimation, we consider θ as an unknown parameter. In MAP estimation, we consider θ as a random variable.
- Our starting point in MAP is $p(\theta|X)$.

From our familiar Bayes theorem we have

$$p(\theta)p(X|\theta) = p(X)p(\theta|X) \tag{1}$$

$$p(\theta|X) = \frac{p(\theta)p(X|\theta)}{p(X)} \tag{2}$$

The **maximum a posteriori probability (MAP)** estimate $\hat{\theta}_{MAP}$ is defined at the point where $p(\theta|X)$ becomes maximum,

$$\hat{\theta}_{MAP} : \frac{\partial}{\partial} p(\theta|X) = 0 \tag{3}$$

Note that the difference between the ML an MAP estimates lies in the involvement of $p(\theta)$ in the latter case.
If $p(\theta)$ obeys the uniform distribution, both estimates yield identical results.

MAP *example 1.*

- Posterior Probability

$$p(\theta|X) \propto p(X|\theta)p(\theta) \equiv \ln p(X|\theta) + \ln p(\theta) \tag{4}$$

- Maximum a Posterior Estimation

$$\max_{\theta}[\ln p(X|\theta) + \ln p(\theta)] \tag{5}$$

- No homework example

lazy student	dog ate it	abducted by alien
0.9998	0.000199	0.000001

MAP *example 2.*

Consider the example 1 on page 17 of Lecture 07:

revisit the example

Let $x_1, x_2, ..., x_N$ be vectors stemmed from a normal distribution with known convariance matrix and unknown mean, that is

$$p(x_k|\mu) = \frac{1}{(2\pi)^{l/2}|\Sigma|^{1/2}} exp(-\frac{1}{2}(x_k - \mu)^T \Sigma^{-1}(x_k - \mu)) \qquad (6)$$

where l is the dimension of vector x_k $(k = 1, ..., N)$.

MAP *example 2.*

Here, let's assume that the unknown mean vector μ is know to be normally distributed as

$$p(\mu) = \frac{1}{(2\pi)^{l/2}\sigma_\mu^l} exp(-\frac{1}{2}\frac{||\mu - \mu_0||^2}{\sigma_\mu^2}) \tag{7}$$

($\Sigma_\mu = \sigma_\mu^2$I, I is the Identity Matrix)

There is a distribution of μ here for the reason that μ is continuous variable, we use its pdf to represent its prior knowledge. In comparison, the prior knowledge in the **homework** example, the prior knowledge is the probability of each reason, since the variable (*reasons* why didn't do homework) is a discrete variable.

MAP *example 2.**

The MAP estimate is given by the solution of

$$\frac{\partial}{\partial \mu} \ln \left(\prod_{k=1}^{N} p(x_k|\mu) p(\mu) \right) = 0 \tag{8}$$

for $\Sigma = \sigma^2 \mathrm{I}$,

$$\sum_{k=1}^{N} \frac{1}{\sigma^2}(x_k - \hat{\mu}) - \frac{1}{\sigma_\mu^2}(\hat{\mu} - \mu_0) = 0 \tag{9}$$

$$\hat{\mu}_{MAP} = \frac{\mu_0 + \frac{\sigma_\mu^2}{\sigma^2} \sum_{k=1}^{N} x_k}{1 + \frac{\sigma_\mu^2}{\sigma^2} N} \tag{10}$$

MAP *example 2.**

$$\hat{\mu}_{MAP} = \frac{\mu_0 + \frac{\sigma_\mu^2}{\sigma^2}\sum_{k=1}^{N} x_k}{1 + \frac{\sigma_\mu^2}{\sigma^2} N} \tag{11}$$

We observe that if $\frac{\sigma_\mu^2}{\sigma^2} \gg 1$, that is the variance σ_μ^2 is very large and the corresponding Gaussian is very with very little variation over the range of interest, then

$$\hat{\mu}_{MAP} \approx \hat{\mu}_{ML} = \frac{1}{N}\sum_{k=1}^{N} x_k \tag{12}$$

Furthermore, for the case $N \to \infty$, regardless of the values of the variances, the MAP estimates to the ML one. This is a more general result.

Thank You !

$\mathscr{Q}$ & $\mathscr{A}$