THU-70250043, Pattern Recognition (Spring 2018) **Homework: 2**

Parameter Estimation Method

Lecturer: Changshui Zhang `zcs@mail.tsinghua.edu.cn`

Student:

MLE and MAP

Maximum Likelihood Estimation (MLE) and Maximum A Posterior (MAP) are two basic principles for learning parametric distributions. In this problem you will derive the MLE and the MAP estimates for some widely-used distributions.

Before stating the problems, we first give a brief review of MLE and MAP. Suppose we consider a family of distributions (c.d.f or p.m.f.) $F := \{f(x|\theta) : \theta \in \Theta\}$, where x denotes the random vector, θ denotes a vector of parameters, and Θ denotes the set of all possible values of θ. Given a set $\{x_1, x_2, ..., x_n\}$ of sample points independently drawn from some $f^* \in F$, or equivalently some $f(x|\theta^*)$ such that $\theta^* \in \Theta$, we want to obtain an estimate of the value of θ^*. Recall that in the case of an independently and identically distributed(i.i.d.) sample the log-likelihood function is in the following form

$$l(\theta) = \sum_{i=1}^{n} \log f(x_i|\theta), \tag{1}$$

which is a function of θ under some fixed sample $\{x_1, x_2, ..., x_n\}$. The MLE estimate $\hat{\theta}_{mle}$ is then defined as follows:

- $\hat{\theta}_{mle} \in \Theta$,
- $\forall \theta \in \Theta$, $l(\theta) \leq l(\hat{\theta}_{mle})$.

If we have access to some prior distribution $P(\theta)$ over Θ, be it from past experiences or domain knowledge or simply belief, we can think about the posterior distribution over Θ:

$$q(\theta) := \frac{\left(\prod_{i=1}^{n} f(x_t|\theta)\right) p(\theta)}{z(x_1, x_2, ..., x_n)}, \tag{2}$$

where

$$z(x_1, x_2, ..., x_n) := \int_{\Theta} \left(\prod_{i=1}^{n} f(x_t|\theta) \right) p(\theta) d\theta. \tag{3}$$

The MAP estimate $\hat{\theta}_{map}$ is then defined as follows:

- $\hat{\theta}_{map} \in \Theta$,
- $\forall \theta \in \Theta$, $q(\theta) \leq q(\hat{\theta}_{map})$, or equivalently,

$$l(\theta) + \log p(\theta) \leq l(\hat{\theta}_{map}) + \log p(\hat{\theta}_{map}). \tag{4}$$

1. MLE for the uniform distribution

Consider a uniform distribution centered on 0 with width $2a$. The density function is given by

$$p(x) = \frac{1}{2a} I(x \in [-a, a]) \tag{5}$$

a. Given a data set $x_1, x_2, \cdots, x_n$, what is the maximum likelihood estimate of a(call it $\hat{a}$)

b. What probability would the model assign to a new data point x_{n+1} using $\hat{a}$

c. Do you see any problem with the above approach? Briefly suggest a better approach

2. Consider a training data of N i.i.d. (independently and identically distribute) observations, $\boldsymbol{X} = \{x_1, x_2, ..., x_N\}$ with corresponding N target values $\boldsymbol{T} = \{t_1, t_2, ..., t_N\}$.

We want to fit these observations into some model

$$t = y(x, \boldsymbol{w}) + \epsilon \tag{6}$$

where $\boldsymbol{w}$ is the model parameters and ϵ is some error term.

2.1 To find $\boldsymbol{w}$, we can minimize the sum of square error

$$E(\boldsymbol{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \boldsymbol{w}) - t_n\}^2 \tag{7}$$

Now suppose we believe that the distribution of error term ϵ is gaussian

$$p(\epsilon|\beta) = \mathcal{N}(\epsilon|0, \beta^{-1}) \tag{8}$$

where $\beta = \frac{1}{\sigma^2}$ is the inverse of variance. Using the property of gaussian distribution, we have

$$p(t|x, \boldsymbol{w}, \beta) = \mathcal{N}(t|y(x, \boldsymbol{w}), \beta^{-1}) \tag{9}$$

Under this assumption, the likelihood function is given by

$$p(\boldsymbol{T}|\boldsymbol{X}, \boldsymbol{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|y(x_n, \boldsymbol{w}), \beta^{-1}) \tag{10}$$

Show that the problem of finding the maximum likelihood (ML) solution for $\boldsymbol{w}$ is equivalent to the problem of minimizing the sum of square error (7).

2.2 In order to avoid overfitting, we often add a weight decay term to (7)

$$E(\boldsymbol{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \boldsymbol{w}) - t_n\}^2 + \frac{\lambda}{2} ||\boldsymbol{w}||^2 \tag{11}$$

On the other hand, we believe that $\boldsymbol{w}$ has a prior distribution of

$$p(\boldsymbol{w}|\alpha) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \alpha^{-1}\boldsymbol{I}) \tag{12}$$

Using Bayes?theorem, the posterior distribution for $\boldsymbol{w}$ is proportional to the product of the prior distribution and the likelihood function

$$p(\boldsymbol{w}|\boldsymbol{X},\boldsymbol{T},\alpha,\beta) \propto p(\boldsymbol{T}|\boldsymbol{X},\boldsymbol{w},\beta)p(\boldsymbol{w}|\alpha) \tag{13}$$

Show that the problem of finding the maximum of the posterior (MAP) solution for $\boldsymbol{w}$ is equivalent to the problem of minimizing (11).

Programming

3. Assume $p(x) \sim 0.2N(-1,1)+0.8N(1,1)$. Draw n samples from $p(x)$, for example, $n = 5, 10, 50, 100, \cdots, 1000, \cdots, 10000$. Use Parzen-window method to estimate $p_n(x) \approx p(x)$ (Hint: use randn() function in matlab to draw samples)

(a) Try window-function $P(x) = \begin{cases} \frac{1}{a}, -\frac{1}{2}a \leq x \leq \frac{1}{2}a \\ 0, otherwise. \end{cases}$. Estimate $p(x)$ with different window width a.

(b)Derive how to compute $\epsilon(p_n) = \int [p_n(x) - p(x)]^2 dx$ numerically.

(c)Demonstrate the expectation and variance of $\epsilon(p_n)$ w.r.t different n and a .

(d)With n given, how to choose optimal a from above the empirical experiences?

(e)Substitute $h(x)$ in (a) with Gaussian window. Repeat (a)-(e).

(g)Try different window functions and parameters as many as you can. Which window function/parameter is the best one? Demonstrate it numerically.