THU-70250043, Pattern Recognition (Spring 2016) **Homework: 2**

Parameter Estimation Method

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

Student:

1.

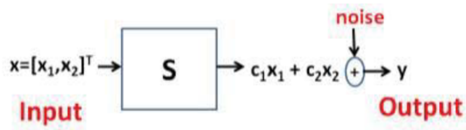

Figure 1: System S

Figure 1 shows a system S which takes two inputs x_1, x_2(which are deterministic) and outputs a linear combination of those two inputs, $c_1x_1 + c_2x_2$, introduces an additive error ϵ which is a random variable following some distribution. Thus the output y that you observe is given by equation (1). Assume that you have $n > 2$ instances $< x_{j1}, x_{j2}, y_j >_{j=1,...,n}$

$$y = c_1x_1 + c_2x_2 + \epsilon \qquad (1)$$

In other words having n equations in your hand is equivalent to having n equations of the following form:$y_j = c_1x_{j1} + c_2x_{j2} + \epsilon_j, j = 1, ..., n$ The goal is to estimate c_1, c_2 from those measurements by maximizing conditional log-likelihood given the input, under different assumptions for the noise. Specifically:

1) Assume that the ϵ_i for $i = 1, ..., n$ are iid Gaussian random variables with zero mean and variance σ^2.

(a) Find the conditional distribution of each y_i given the inputs

(b) Compute the log-likelihood of y given the inputs

(c) Maximize the likelihood above to get c_{ls}

2) Assume that the ϵ_i for $i = 1, ..., n$ are independent Gaussian random variables with zero mean and variance $Var(\epsilon_i) = \sigma_i$.

(a) Find the conditional distribution of each y_i given the inputs

(b) Compute the log-likelihood of y given the inputs

(c) Maximize the likelihood above to get c_{wls}

3) Assume that the ϵ_i for $i = 1, ..., n$ has density $f_{\epsilon_i}(x) = f(x) = \frac{1}{2b}exp(-\frac{|x|}{b})$. In other words our noise is iid following Laplace distribution with location parameter $\mu = 0$ and scale parameter b.

(a) Find the conditional distribution of each y_i given the inputs

(b) Compute the log-likelihood of y given the inputs

(c) Comment on why this model leads to more robust solution.

2. Consider a normal $p(x) \sim N(\mu, \sigma^2)$ and Parzen-window function $\phi(x) \sim N(0,1)$ Show that the Parzen-window estimate

$$p_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} \phi(\frac{x - x_i}{h_n})$$

has the following properties:

(a)$\overline{p}_n(x) \sim N(\mu, \sigma^2 + h_n^2)$

(b)$Var[p_n(x)] \simeq \frac{1}{2nh_n\sqrt{\pi}} p(x)$

(c)$p(x) - \overline{p}_n(x) \simeq \frac{1}{2}(\frac{h_n}{\sigma})^2[1 - (\frac{x-\mu}{\sigma})^2]p(x)$ for small h_n

(Note: if $h_n = \frac{h_1}{\sqrt{n}}$, this show that the error due to bias goes to zero as $1/n$, whereas the standard deviation of the noise only goes to zero as $\sqrt[4]{n}$.)

3. One measure of the difference between two distributions in the same space is the Kullback-Leibler divergence of Kullback-Leibler "distance":

$$D_{KL}(p_1(x), p_2(x)) = \int p_1(x) ln \frac{p_1(x)}{p_2(x)} dx$$

(This "distance" does not obey the requisite symmetry and triangle inequalities for a metric.)Suppose we seek to approximate an arbitrary distribution $p_2(x)$ by a normal $p_1(x) \sim N(\mu, \Sigma)$. Show that the values that lead to the smallest Kullback-Leibler divergence are the obvious ones:

$$\mu = \epsilon_2[x]$$
$$\Sigma = \epsilon_2[(x - \mu)(x - \mu)^T]$$

where the expectation ϵ_2 taken is over the density $p_2(x)$.

4. (Programming) Assume $p(x) \sim 0.1N(-1,1) + 0.9N(1,1)$. 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 $P(x) = \begin{cases} \frac{1}{a}, -\frac{1}{2}a \le x \le \frac{1}{2}a \\ 0, otherwise. \end{cases}$. 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.