Ali Rahimpour

Student Number: 810192071

Department of
Electrical Engineering and Computer Science
University of Tehran

Assignment Number Two

Pattern Recognition course

Course Lecturer: Dr.Babak Nadjar Araabi

26 October 2013

Contents

Abstract

There are two basic series for estimation PDF with sampling observation : Parametric and Non-parametric Estimation. Parametric estimation include Moment Method Maximum Likelihood ,Maximum A posteriori. Non-parametric one include Histogram that is a simple one , Parzen method and K-NN (K nearest neighbour). Also,there is a complicated parametric method that have the advantages of non-parametric methods.

In this report, we use each of this methods to solve problems and we can drew a conclusion that each of them has characteristics determining the advantages and disadvantages and in different conditions.

Introduction

As we mentioned above,there are two basic series for estimation PDF with sampling observation:Parametric and Non-parametric Estimation. Parametric estimation include Moment Method, Maximum Likelihood , Maximum A posteriori. Non-parametric one include Histogram that is a simple one , Parzen method and K-NN (K nearest neighbour). Also,there is a complicated parametric method that have the advantages of non-parametric methods.

In parametric method, moment method estimate moments with considering the samples.This method is useful when obtained equations could be resolvable.There is several constraints to do that.after that,the second one,maximum likelihood is objective. As a result, the parameters are adjusted however what we observe have maximum probability .For solving optimization,ML should be too difficult then several methods are presented such as Expectation Maximization (EM). Then the third one is Maximum A Posteriori that should be suitable for estimation of parameters.If we do not have any prior observation on θ ,we can replace consistent distribution for $f(\theta)$ that is equal to ML.

In Non-parametric method,histogram method is a simplest and there are some constraint such as sample number and number of bins and then this is useful for two or three random variables.We have more complicated methods that have the idea of histogram method but in more general way.Second one,Parzen method,that in this method, V is constant and controller and K is variable.then ,we consider a hypercube and then count the number of samples.Third one is K-NN (K nearest neighbour) in this method ,K is constant and controller and V is variable.In this method we should expand hyper cube till involve all of K samples.

Subject: Year Month Date

Ali Rahimpour ٨١٠١٩٢٠٧١

Homework 2

1. In Maximum likelihood Method with having $\mathcal{X} = \{\underline{x}_k\}_1^N$, first of all, we assume that this dataset bring to gether from a distribution with unknown parameters and then we choose these parameters as they have most probability. Another supposal is that samples are i.i.d.

Then we have:

$$\hat{\underline{\theta}}_{ML} = \arg\max_{\underline{\theta}} f(X;\underline{\theta}) \xrightarrow{i.i.d} \hat{\underline{\theta}}_{ML} = \arg\max_{\underline{\theta}} \prod_{k=1}^{N} f(\underline{x}_k;\underline{\theta})$$

we should find absolute maximum of this function:

a) As we mentioned before:

$$\hat{\theta}_{ML} = \arg\max_{\theta} \prod_{k=1}^{N} \theta e^{-\theta x_k} = \arg\max_{\theta} \theta^N e^{-\theta(x_1+\dots+x_N)} = \arg\max_{\theta} F(\theta)$$

To obtain absolute maximum we should test boundary and cross point into the function, therefore with obtaining derivation & do below we can find these points:

$$\frac{dF}{d\theta} = N\theta^{N-1} e^{-\theta(x_1+\dots+x_N)} - (x_1+x_2+\dots+x_N)\theta^N e^{-\theta(x_1+\dots+x_N)} = 0$$

$$\Rightarrow \theta^{N-1}\left(Ne^{-\theta(x_1+\dots+x_N)} - (x_1+\dots+x_N)\theta e^{-\theta(x_1+\dots+x_N)}\right) = 0 \Rightarrow \begin{cases} \theta = 0 \quad \theta > 0 \text{ غ ق ق} \\ \boxed{\theta = \dfrac{N}{x_1+\dots+x_N}} \end{cases}$$

As that's so clear the absolute maximum is $\longrightarrow$ $\theta = \dfrac{N}{x_1+\dots+x_N}$

b)

$$\hat{\theta}_{ML} = \arg\max_{\theta} \prod_{k=1}^{N} \frac{x_k}{\theta^2} e^{-\frac{x_k^2}{2\theta^2}} = \arg\max_{\theta} \frac{x_1 * \dots * x_K}{\theta^{2N}} e^{-\frac{(x_1^2+\dots+x_N^2)}{2\theta^2}}$$

$$= \arg\max_{\theta} F(\theta)$$

$$\frac{dF}{d\theta} = -2N * \frac{x_1 \cdot x_2 \cdots x_N}{\theta^{2N+1}} e^{\frac{-(x_1^2+\dots+x_N^2)}{2\theta^2}} + \frac{2(x_1^2+\dots+x_N^2)}{2\theta^3} \cdot \frac{x_1 \cdots x_K}{\theta^{2N}} e^{\frac{-(x_1^2+\dots+x_N^2)}{2\theta^2}}$$

$$\Rightarrow \frac{x_1 \cdots x_k}{\theta^{2N+1}} e^{\frac{-(x_1^2+\dots+x_N^2)}{2\theta^2}} \cdot \left(-2N + \frac{x_1^2+\dots+x_N^2}{\theta^2}\right) = 0$$

$$\Rightarrow \boxed{\theta = \sqrt{\frac{x_1^2+\dots+x_N^2}{2N}}, \ \theta > 0}$$

PAPCO

Year Month Date

c. $\hat{\theta}_{ML} = \arg\max_{\theta} \prod_{k=1}^{N} \sqrt{\theta}\, x_k^{\sqrt{\theta}-1} = \arg\max_{\theta} \sqrt{\theta^N}\,(x_1 x_2 \cdots x_N)^{\sqrt{\theta}-1} = \arg\max_{\theta} f(\theta)$

$$\frac{dF}{d\theta} = \frac{N}{2}\theta^{\frac{N}{2}-1}(x_1\cdots x_N)^{\sqrt{\theta}-1} + \ln(x_1\cdots x_N)\frac{1}{2\sqrt{\theta}}\theta^{\frac{N}{2}}(x_1\cdots x_N)^{\sqrt{\theta}-1} = 0$$

$$\Rightarrow \frac{1}{2}\theta^{\frac{N}{2}-1}(x_1\cdots x_N)^{\sqrt{\theta}-1}\left(N + \ln(x_1\cdots x_N)\theta^{\frac{1}{2}}\right) = 0 \Rightarrow \boxed{\hat{\theta}_{ML} = \left(\frac{-N}{\ln(x_1\cdots x_N)}\right)^2}$$

---

2)

a) As we mentioned in previous question, we have:

$$\hat{\underline{\theta}}_{ML} = \arg\max_{\underline{\theta}} f(X;\underline{\theta}) \xrightarrow{iid} \hat{\underline{\theta}}_{ML} = \arg\max_{\underline{\theta}} \prod_{k=1}^{N} f(\underline{x_k};\underline{\theta})$$

$$\Rightarrow \hat{\theta}_{ML} = \arg\max_{\theta} \begin{cases} \prod_{k=1}^{N} \frac{1}{\theta} & 0 \le x_k \le \theta, \forall k = 1,\dots,N \\ 0 & \text{o.w} \end{cases}$$

As this is obvious, θ must be bigger than all of x_i then the function would be nonzero, in result we can write:

$$\hat{\theta}_{ML} = \arg\max_{\theta} \frac{1}{\theta^N} = \arg\max_{\theta} f(\theta) \qquad \max(x_k) \le \theta < \infty$$

$\frac{dF}{d\theta} = \frac{-N}{\theta^{N+1}}$ $\Rightarrow$ as a result, this can't be $\frac{dF}{d\theta} = 0$ then

Maximum amount occurs in boundaries $\Rightarrow$ Max(0)

b) note that:

$$\prod_{k=1}^{N} f(\underline{x_k};\underline{\theta}) = f(\theta) = \begin{cases} 0 & 0 \le \theta < 0.6 \\ \frac{1}{\theta^5} & 0.6 \le \theta < 1 \end{cases}$$

that is shown in figure 2.b.

As that's obvious in this figure, Maximum likelihood amount is equal to zero for amount smaller than 0.6. About 4 another points we know they are positive that's sufficient to know that Because it should be bigger than others and that is bigger than biggest of them. & knowing that do not affect on determination of θ.

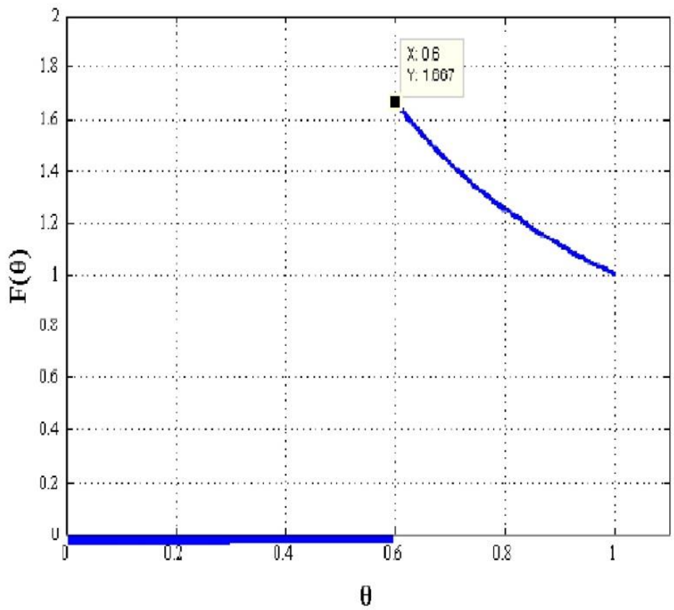

Figure 2.b.Figure of F(θ) .Illustrating of Maximum Likelihood amount that is equal to zero for positive amount smaller than 0.6 and for amount bigger than 0.6 is equal to θ^{-5}

As we mentioned in previous question, we have:

$$\hat{\theta}_{ML} = \arg\max_{\theta} f(x;\theta) \Rightarrow$$

As we know: $p(x|\theta) = \begin{cases} \theta^2 x e^{-(\theta x)} & x > 0 \\ 0 & x \leq 0 \end{cases}$

$$\hat{\theta}_{ML} = \arg\max_{\theta} \prod_{k=1}^{N} f(x_k;\theta)$$

$$\Rightarrow \hat{\theta}_{ML} = \arg\max_{\theta} \prod_{k=1}^{N} \theta^2 x_k e^{-(\theta x_k)}$$

$$= \arg\max_{\theta} \theta^{2N}(x_1 \cdot x_2 \cdots x_N)\, e^{-\theta(x_1 + x_2 + \cdots + x_N)} = \arg\max_{\theta} f(\theta)$$

$$\frac{df}{d\theta} = 2N\theta^{2N-1}(x_1 \cdot x_2 \cdots x_N)\, e^{-\theta(x_1 + \cdots + x_N)} + \theta^{2N}(x_1 \cdot x_2 \cdots x_N)(-1)(x_1 + \cdots + x_N) \cdots e^{-(\theta)(x_1 + \cdots + x_N)} = 0$$

$$\Rightarrow \boxed{\hat{\theta}_{ML} = \frac{2N}{x_1 + x_2 + \cdots x_N}}$$

Subject:
Year: Month: Date:

4- For Gaussian distribution we can write:

$$f(x_k;\theta|j)=\frac{1}{(2\pi\sigma_j^2)^{l/2}}e^{-\frac{\|x_k-\mu_j\|^2}{2\sigma_j^2}}$$

$$Q(\theta;\hat{\theta}(t))=\sum_{k=1}^{N}\sum_{j=1}^{J}P(j;\hat{\theta}(t)|x_k)\left\{-\frac{\|x_k-\mu_j\|^2}{2\sigma^2}-\frac{l}{2}\ln(2\pi\sigma_j^2)+\ln P_j\right\}$$

we should maximize upper function, then:

$$\frac{\partial Q}{\partial\mu_i}=\sum_{k=1}^{N}P(i;\hat{\theta}(t)|x_k)\left\{\frac{x_k-\mu_i}{\sigma_i^2}\right\}=0 \quad (1),\qquad \frac{\partial Q}{\partial\sigma_i}=\sum_{k=1}^{N}P(i;\hat{\theta}(t)|x_k)\left\{\frac{\|x_k-\mu_i\|^2}{\sigma_i^3}-\frac{l}{\sigma_i}\right\}=0 \quad (2)$$

$$\frac{\partial Q}{\partial P_i}=\sum_{k=1}^{N}P(i;\hat{\theta}(t)|x_k)\cdot\frac{1}{P_i}=0 \quad (3)$$

From (1) $\Rightarrow$ $-\frac{1}{\sigma_i}\cdot\mu_i\sum_{k=1}^{N}P(i;\hat{\theta}(t)|x_k)=-\frac{1}{\sigma_i^2}\sum_{k=1}^{N}x_kP(i;\hat{\theta}(t)|x_k)$

$$\mu_i(t+1)=\frac{\sum_{k=1}^{N}x_kP(i;\hat{\theta}(t)|x_k)}{\sum_{k=1}^{N}P(i;\hat{\theta}(t)|x_k)}$$

From (2) $\Rightarrow$ $\frac{1}{\sigma_i^3}\sum_{k=1}^{N}\{\|x_k-\mu_i\|+l\sigma_i^2\}P(i;\hat{\theta}(t)|x_k)=0$

$$\sigma_i^2=\frac{\sum_{k=1}^{N}\|x_k-\mu_i\|P(i;\hat{\theta}(t)|x_k)}{l\cdot\sum_{k=1}^{N}P(i;\hat{\theta}(t)|x_k)}$$

From 3, we can conclude that P_i should be infinitive, also we know that probability should be less than 1 & it means that we don't suppose $\sum_{j=1}^{J}P_j=1$, so this problem change to optimization. This problem is solved with Lagrangh method. Then new function:

$$L=Q(\theta,\hat{\theta}(t))+\lambda\left(\sum_{j=1}^{J}P_j-1\right)$$

We can observe that this function does not affect on calculation of σ_i^2 & μ_i. But for λ & P:

$$\frac{\partial Q}{\partial P_i}=\sum_{k=1}^{N}P(i;\hat{\theta}(t)|x_k)\frac{1}{P_i}+\lambda=0 \quad (4)$$

$$\frac{\partial Q}{\partial\lambda}=\sum_{j=1}^{J}P_j-1=0 \quad (5)$$

PAPCO

Subject:
Year. Month. Date.

$$\Rightarrow \sum_{j=1}^{J} P_j = -\frac{1}{\lambda} \sum_{j=1}^{J} \sum_{k=1}^{N} P(j; \hat{\Theta}(t) | x_k) = 1 = -\frac{1}{\lambda} \sum_{k=1}^{N} \sum_{j=1}^{J} P(j; \hat{\Theta}(t) | x_k)$$

we have : $\sum_{j=1}^{J} P(j; \hat{\Theta}(t) | x_k) = 1 \Rightarrow \lambda = -\sum_{k=1}^{N} 1 = -N$

$$P_i = \frac{1}{N} \sum_{k=1}^{N} P(i; \hat{\Theta}(t) | x_k)$$

b) $P(j; \hat{\Theta}(t) | x_k) = \dfrac{f(x_k; \hat{\Theta}(t) | j) P_j(t)}{h_x(x_k; \hat{\Theta}(t))} = \dfrac{f(x_k; \hat{\Theta}_j(t)) P_j(t)}{h_x(x_k; \hat{\Theta}(t))}$

$h_x(x_k; \hat{\Theta}(t))$ is initial distribution that we assume in Mixture Model method :

$$h_x(x_k; \hat{\Theta}(t)) = \sum_{i=1}^{J} f(x_k; \hat{\Theta}(t) | i) P_i(t) = \sum_{i=1}^{J} f(x_k; \hat{\Theta}_i(t)) P_i(t)$$

$$P(j; \hat{\Theta}(t) | x_k) = \frac{f(x_k; \hat{\Theta}_j(t)) P_j(t)}{\sum_{i=1}^{J} f(x_k; \hat{\Theta}_i(t)) P_i(t)}$$

$f(x_k; \hat{\Theta}_j(t))$ is such a distribution that we want to find unknown parameters that is gussian distrebution we had in previous sections for other variables such as $\hat{\Theta}_i(t)$ that is σ_i & μ_i, we suppose an inital amount for first Itteration & for next steps we use obtained amount from previous section. For P_i ; it is similar.

c) We know that $\Theta = [\theta^T, (P_1, \dots, P_J)^T]^T$ then we can obtain ~~from~~ for P that $P_i = \frac{1}{N} \sum_{k=1}^{N} P(i; \hat{\Theta}(t) | x_k)$ therefore this amount depends on N (number of data), x & θ that θ is unknown distributive parameter that is dependent on μ_i & σ_i on previous step.

5. We implement this problem in MATLAB Software. (that there is M-file attached in folder.). We report results here.

a) We assume below as a initial Conditions

$$\sigma^2 = j,\quad \mu_j = j,\quad P_i = \frac{1}{J},\quad J = 3$$

Figure (Table 5.a is shown bellow)

As we can see, amounts are near to real ones.

b) Table 5.b related to this part is shown below.

In this state we apply Estimation on underparameter therefore we havent a good adabtiblity. We can see from this amounts that first distribution try to estimate approximation 1 & 2 and try to solve both of distributions.

c. Table 5.c related to this part is shown below.

In that state we apply Estimation on over parameter & first & second distribution try to estimate first distribution & other two distributions estimate real two & three distribution.

$P_1 + P_2 \simeq 0.25 \qquad \dfrac{P_1\mu_1 + P_2\mu_2}{P_1 + P_2} = 1 \Rightarrow$ that's a good estimation, too.

Table 5.a.With assuming factors in below there are some final characteristics

$$J = 3; P_i = \frac{1}{J}; \mu_j = j; \sigma^2 = j;$$

	1	2	3
P_i	0.2537	0.4905	0.2557
μ_i	1.0061	1.5067	2.0204
σ_i	0.0102	0.0094	0.0441

Table 5.b. With assuming factors related to question 5.b there are some final characteristics

	1	2
P_i	0.8625	0.1375
μ_i	1.4111	2.1377
σ_i	0.0973	0.0274

Table 5.c. With assuming factors related to question 5.c there are some final characteristics

	1	2	3	4
P_i	0.1178	0.1367	0.4890	0.2565
μ_i	1.0166	1.0001	1.5065	2.0205
σ_i	0.0041	0.0162	0.0098	0.0447

Subject:
Year. Month. Date. ( )

6.

a) For Parzen method we follow procedure as below:

1. we assume a Hypercube. $V_n = h_N^n$
2. we consider center of Hypercube on point x.
3. we ~~counter~~ the points into the Hypercube: k_N
4. $\hat{f}(x) = \frac{k_N}{h_N^n * N} = \frac{1}{h_N^n * N} \sum_{j=1}^{N} \varphi\left(\frac{x - x_j}{h_N}\right)$

Then, we have for this problem:

i. $V_N = 2^2$, Because the features are two dimentional. We have such this amount for Both class.

For class ω_2 we have $k_N = 4$ and for another one is $k_N = 2$

For the first one (o) we have $N=5$ & for another one have $N=6$

then:

$$\begin{cases} f((0.5,0)^T|\omega_1) = \frac{2}{2^2 . 6} = \frac{1}{12} \\ f((0.5,0)^T|\omega_2) = \frac{4}{2^2 . 5} = \frac{1}{5} \end{cases}$$

$$f((0.5,0)^T|\omega_2) > f((0.5,0)^T|\omega_1) \Rightarrow \begin{bmatrix} 0.5 \\ 0 \end{bmatrix} \in \omega_2$$

ii) $V_N = \pi(1)^2$, Because the features are two dimentional we have this amount for two class.

For class ω_2 we have $k_N = 2$ and for another one have $k_N = 0$.

Also, For first one (.) we have $N=5$ & for another one have $N=6$.

$$\left.\begin{aligned} f((0.5,0)^T|\omega_1) &= \frac{0}{\pi . 6} = 0 \\ f((0.5,0)^T|\omega_2) &= \frac{4}{\pi . 5} = \frac{4}{5\pi} \end{aligned}\right\} \Rightarrow f((0.5,0)^T|\omega_2) > f((0.5,0)^T|\omega_1) \Rightarrow \begin{bmatrix} 0.5 \\ 0 \end{bmatrix} \in \omega_2$$

b) K_Nearest neighbor

we do similar to previous method. But with a basic difference, in this method we consider samples on a special amount k rather than considering ~~samples~~ ~~a hypercube & then counter amount of samples. The rest of steps~~

Year. Month. Date.

is similar to previous method. $\hat{f}(x)$ is equal to:

$$\hat{f}(x) = \frac{k_N}{h_N^n \cdot N}$$

Then we have:

i) With considering problem; $k_N = 3$

For class ω_2, we have $k_N = \pi \cdot (1.18)^2$ & for class ω_1 we have $k_N = \pi (1.5)^2$.
For firstone $N=5$ & for another one $N=6$. Then we have:

$$\begin{cases} f((0.5,0)^T|\omega_1) = \dfrac{3-1}{1.18^2 \pi (6)} = 0.0472 \\ f((0.5,0)^T|\omega_2) = \dfrac{3-1}{(1.5)^2 \pi (5)} = 0.0914 \end{cases} \Rightarrow f((0.5,0)^T|\omega_2) > f((0.5,0)^T|\omega_1) \Rightarrow \begin{bmatrix} 0.5 \\ 0 \end{bmatrix} \in \omega_2$$

ii) with considering problem: $k_N = 3$

For class ω_2, we have $k_N = 2^2$ and for class ω_1 have $k_N = 3^2$. For firstone $N=5$ and for another one $N=6$. Then we have:

$$\begin{cases} f((0.5,0)^T|\omega_1) = \dfrac{3-1}{3^2 \cdot 6} = \dfrac{1}{27} \\ f((0.5,0)^T|\omega_2) = \dfrac{3-1}{2^2 \cdot 5} = \dfrac{1}{10} \end{cases} \Rightarrow f((0.5,0)^T|\omega_2) > f((0.5,0)^T|\omega_1) \Rightarrow \begin{bmatrix} 0.5 \\ 0 \end{bmatrix} \in \omega_2$$

PAPCO

Conclusion

As far as we are concerned with the concept of estimation, we can draw the conclusion that each of this series is usefull for special application .To explain more clearly from parametric and Non-parametric method and knowing some meanings and explainations to estimating function, there is a special way to analytically and statistically optimize our classification.

When we have less number of N ,K-NN is more precise and Parzen method is usefull for larger amount. In general ,Non-parametric methods need more amount of data in comparison with parametric one.