| | |
|---|---|
| **CSCE 970: Pattern Recognition** | **Spring Semester, 1999** |

Topic 2: Estimating Unknown pdf's

*Lecturer: Stephen Scott* *Scribe: Mengchu Cai*

# 1 Introduction

In earlier lectures we assumed that the probability density functions are known. However, this is not the most common case. In practice, the underlying pdf is unknown and has to be estimated from the examplar data available from each class, calling training samples. There are various ways to approach the problem. *Parametric methods* assume that we know the type of the pdf(e.g., Gaussian) but we do not know certain parameters, such as the mean value or the variance. *Nonparametric methods* make no such assumptions and classify a test object based on the training samples in the neighbourhood of the object in the feature space. Examples are the k-nearest neighbour and potential method[Butu 93].

# 2 Review of Discriminant Function

Recall the Bayes classification rule

$$\text{If } P(\omega_i|x) > P(\omega_j|x),\ \text{x is classified to } \omega_i$$
$$\text{If } P(\omega_i|x) < P(\omega_j|x),\ \text{x is classified to } \omega_j$$

Sometimes, it may be more convenient to work with an equivalent function than to work directly with probabilities. e.g $g_i(x) \equiv f(P(\omega_i|x)$, where $f(\cdot)$ is monotonically increasing. $g_i(x)$ is known as discriminant function. If we know the pdf, then we can use the following discriminant function:

$$g_i(x) = \ln(p(x|\omega_i)P(\omega_i))$$

And

$$\text{classify } x \text{ in } \omega_i \text{ if } g_i(x) > g_j(x)\, \forall j \neq i$$

# 3 Parametric Methods

Let us consider an M-class problem with vectors distributed according to $p(x|\omega_i), i = 1, 2, \ldots, M$. We assume that we know the type of the pdf but we do not know the certain parameters. Assume that the data from one class do not affect the parameter estimation of the others. Our goal is to estimate the unknown parameters using a set of known feature vectors in each class.

## 3.1 Maximum Likelihood Parameter Estimation

**Example** Suppose we know the pdf is a normal distribution with a known covariance matrix($\Sigma$) and unknown mean($\mu$). We can parameterize $\omega$'s pdf on mean $\mu$:

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

and use data $x_1, x_2, \ldots, x_N$ from $\omega$ to estimate $\mu$. Assuming statistical independence between samples we have

$$p(X;\mu) \equiv p(x_1, x_2, \ldots, x_N;\mu) = \prod_{k=1}^{N} p(x_k;\mu)$$

*Maximum likelihood (ML) method* estimates $\mu$ such that the likelihood function $p(X;\mu)$ is maximized.

Suppose the maximum value of the function is $\hat{\mu}_{ML}$. The gradient of the likelihood function with respect to $\mu$ must be zero at the point $\mu = \hat{\mu}_{ML}$.

Because of the monotonicity of the logarithmic function, we take logarithm and set gradient $= 0$:

$$\frac{\partial}{\partial \mu}\left(-\frac{N}{2}\ln\left((2\pi)^l|\sigma|\right) - \frac{1}{2}\sum_{k=1}^{N}(x_k-\mu)^T\sigma^{-1}(x_k-\mu)\right) = 0$$

Assuming statistical independent of $x_k$'s, $\Sigma_{ji}^{-1} = 0$ for $i \neq j$, so

$$\frac{\partial L(\mu)}{\partial \mu} = \begin{bmatrix} \frac{\partial L}{\partial \mu_1} \\ \vdots \\ \frac{\partial L}{\partial \mu_l} \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial \mu_1}\left(-\frac{1}{2}\sum_{k=1}^{N}\sum_{j=1}^{l}(x_{kj}-\mu_j)^2\Sigma_{jj}^{-1}\right) \\ \vdots \\ \frac{\partial}{\partial \mu_l}\left(-\frac{1}{2}\sum_{k=1}^{N}\sum_{j=1}^{l}(x_{kj}-\mu_j)^2\Sigma_{jj}^{-1}\right) \end{bmatrix} = \sum_{k=1}^{N}\Sigma^{-1}(x_k-\mu) = 0$$

and we get,

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

**Generalization** The technique can be generalized for other distribution and parameters. Let $x_1, x_2, \ldots, x_N$ be samples drawn from pdf $p(x;\theta)$ where $\theta$ is the unknown parameter. We form the joint pdf $p(X;\theta)$, where $X = \{x_1, x_2, \ldots, x_N\}$. Assuming statistical independence between the different samples we have

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

By ML method, let

$$L(\theta) = \ln \prod_{k=1}^{N} p(x_k;\theta)$$

and set gradient of $L(\theta) = 0$ and get the following equation,

$$\frac{\partial L(\theta)}{\partial \theta} = \prod_{k=1}^{N} \frac{1}{p(x_k;\theta)}\frac{\partial p(x_k;\theta)}{\partial \theta} = 0$$

Then we solve the above equation for each class independently.

**Properties** Maximum likelihood method has some nice properties as $N \to \infty$:

- The ML estimate converges in the mean to the true value, which means that

  $$\lim_{n\to\infty} E[\hat{\theta}_{ML}] = \theta_0$$

  and is also called asymptotically unbiased.

- The ML estimate converges in probability, which means that

$$\lim_{N\to\infty} prob\{||\hat{\theta}_{ML} - \theta_0|| \le \epsilon\} = 1$$

and is also called asymptotically consistent.

- The estimate converges in the mean square

$$\lim_{N\to\infty} E[||\hat{\theta}_{ML} - \theta_0||^2] = 0$$

- The ML is asymptotically efficient; it achieves the CramerRao lower bound.

- The pdf of the ML estimate as $N \to \infty$ approaches the Gaussian distribution with mean $\theta_0$.

### 3.2 Maximum A Posteriori Probability Estimator

Let us consider the example in previous subsection again. Now we assume that the unknown mean vector $\mu$ is normally distributed with $\Sigma = \sigma_\mu^2 I$, mean $= \mu_0$, thus

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

The *maximum a posteriori probability*(MAP) estimate $\hat{\mu}_{MAP}$ is defined at the point where $p(\mu|X)$ becomes maximum. From Bayes theorem we have

$$p(\mu|X) = \frac{p(\mu)p(X|\mu)}{p(X)}$$

and assume that the samples are independent, so maximizing $p(\mu|X)$ is same as maximizing the following

$$p(\mu)p(X|\mu) = \prod_{k=1}^{N} p(x_k|\mu)p(\mu)$$

Again, we take logarithm and set gradient=0:

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

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

If $\frac{\sigma_\mu^2}{\sigma^2} \gg 1$, that is $p(\mu)$ is almost uniform, or $N \to \infty$, then

$$\mu_{MAP} \approx \mu_{ML} = \frac{1}{N}\sum_{k=1}^{N} x_k$$

**Generalization** Suppose that the unknown parameter is $\theta$ and we know its distributed function $p(\theta)$. Let $X = \{x_1, x_2, \ldots, x_N\}$ , then MAP estimate $\hat{\theta}_{MAP}$ is defined at the point where $p(\theta|X)$ becomes maximum.

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

The difference between ML and MAP estimates lies in the involvement of $p(\theta)$ in MAP estimate. If $p(\theta)$ is almost uniform, both estimates yield the same result. The following figure illustrate the relation between MAP and ML.

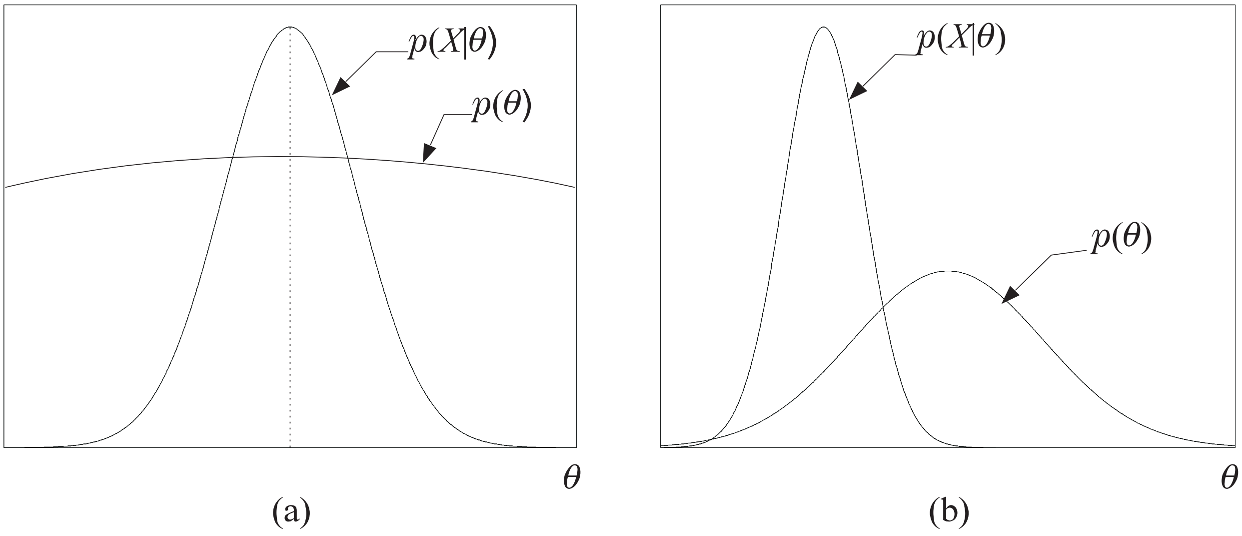


Figure 1: ML and MAP estimates of $\theta$ will be approximately the same in(a) and different in (b)

# 4 Nonparametric Methods

## 4.1 Histogram-based Approximation of pdf

Histogram based technique partition space into "bins" and count the number of samples per bin. Suppose the total number of samples is $N$, and $k_N$ is the number of samples located in a bin, we use the frequency ratio to approximate the corresponding probability.

$$P \approx k_N/N$$

This approximation converges to the true $P$ as $N \to \infty$. Then we assumed the pdf value is constant throughout the bin. So the pdf value is P divided by the volume of the bin. In one-dimentional case,

$$\hat{p}(x) \equiv \hat{p}(\hat{x}) \approx \frac{1}{h}\frac{k_N}{N}, |x - \hat{x}| \leq \frac{h}{2}$$

where $\hat{x}$ is the mid point of the bin and h is the length of the bin. Figure 2 shows two examples of a pdf and its approximation by histogram method.

In the $l$-dimensional case, the space is divided into hypercubes with length of side h and volume $h^l$. Given x, to compute the approximation of pdf $p(x)$, we should

- Count number of training vectors in hypercube H centered at x. To do this, we define function $\phi(x)$ so that

$$\phi(x) = \begin{cases} 1 & \text{if } |x_j| \leq 1/2, j = 1, 2, \ldots, l \\ 0 & \text{otherwise} \end{cases}$$

  then the number of training vectors in H equals to $\sum_{i=1}^{N} \phi\left(\frac{x_i - x}{h}\right)$.

- Divide by $N$ to estimate the probability of getting point in H.

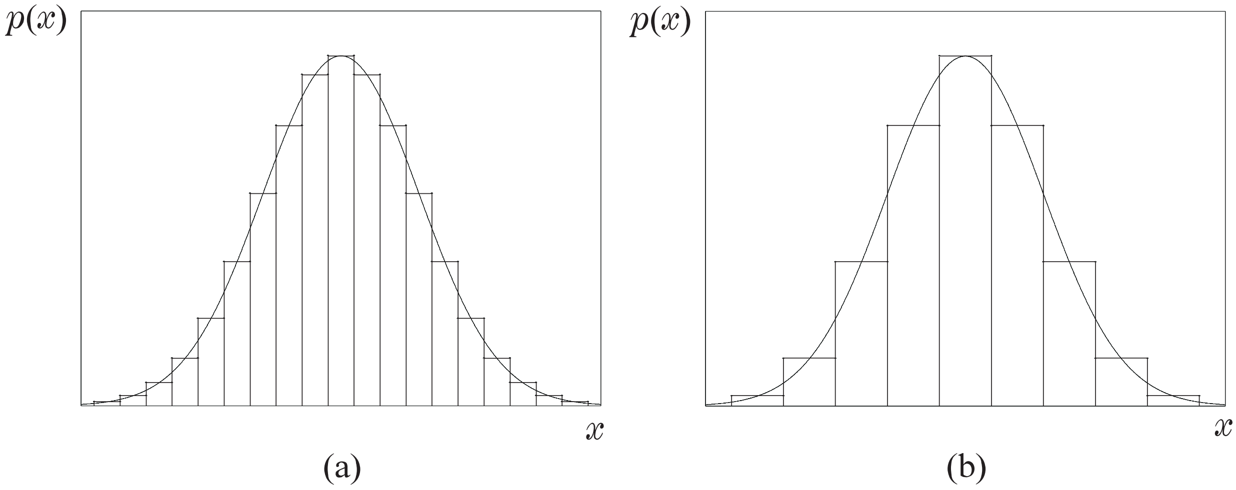


Figure 2: pdf approximation by histogram method with (a)small and (b)large bin size

- Divide by volume of H to approximate the pdf value at x.

$$\hat{p}(x) = \frac{1}{h^l}\left(\frac{1}{N}\sum_{i=1}^{N}\phi\left(\frac{x_i - x}{h}\right)\right)$$

### 4.2 Parzen Window

The problem about the simple histogram method is that it tries to approximate a continuous function $p(x)$ via an expansion of discontinuous function $\phi(\cdot)$. The solution is using smooth functions in place of $\phi(\cdot)$. The smooth function $\phi(x)$ should satisfies:

$$\phi(x) \geq 0 \ and$$

$$\int_x \phi(x)dx = 1$$

For example, $\phi(x) = (1/(2\pi)^{l/2})exp(-x^T x/2)$. Such smooth functions are known as *kernels or Parzen windows.*

Remarks:

- For fixed h, the variance decreases as N increases, e.g Fig4(a) and (b).
- For fixed N, the variance decreases as h increases, e.g Fig4(a) and fig5(a).
- In practice, a compromise between $h$ and $N$ must be made.

### 4.3 k Nearest Density Estimation

In the Parzen estimation, the volume around the point x was considered fixed and the number of samples $k_N$ falling inside the volume was considered to vary from point to point. Here we will fix the number points and adjust the volume that contains the k points.

$$\hat{p}(x) = \frac{k}{NV(x)}$$

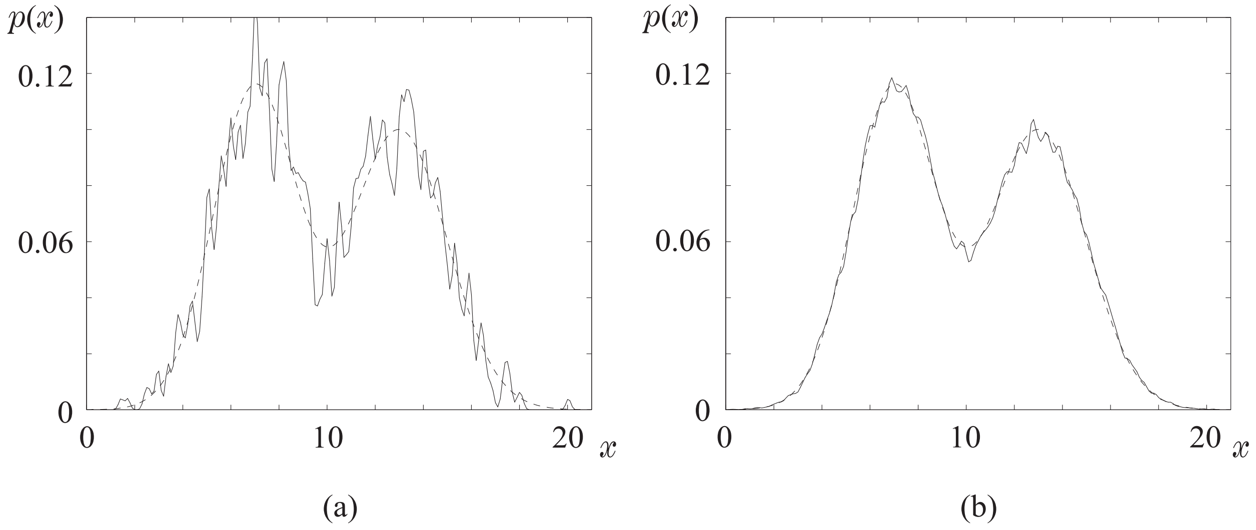


Figure 3: Approximation of pdf(dot line) via Parzen windows, using Gaussian kernels with $h = 0.1$ and (a) 1000 training samples and (b)20,000 training samples

this is an asymptotically unbiased and consistent estimate of the true pdf and is called *k Nearest Neighbour(kNN) density estimate.*

*Nearest Neighbour Rule* Given an unlabeled feature vector and a distance measure, we classify x according to a majority vote of its k nearest neighbours. If $k = 1$, it is known as 1-NN and assigns x to the class of its nearest neighbour. As $N \to \infty$, the classification error probability of 1-NN($P_{NN}$) is at most twicw optimal Bayes error($P_B$). For k-NN, the error probability $P_{kNN}$ is bounded by

$$P_B \leq P_{kNN} \leq P_B + \frac{1}{\sqrt{ke}}$$

## References


[Butu 93] Buturovic L.J. "Improve k-nearest neighbour density and error estimates," Pattern Recognition, Vol. 26(4), pp.611-616, 1993.

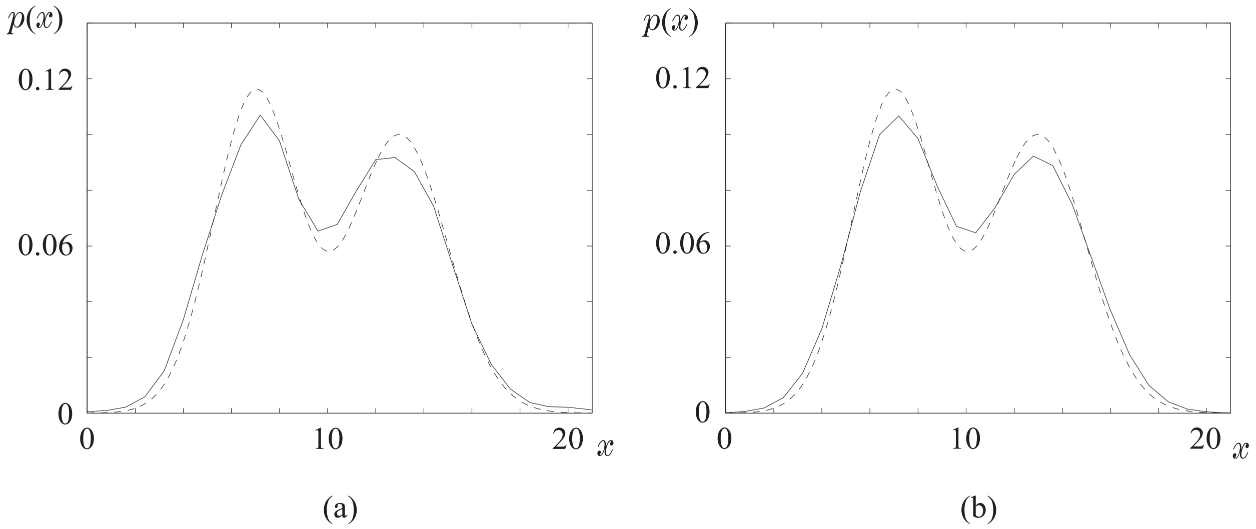


Figure 4: Approximation of pdf(dot line) via Parzen windows, using Gaussian kernels with $h = 0.8$ and (a) 1000 training samples and (b)20,000 training samples