Parzen window method and classification

A slecture by Chiho Choi

Density estimation using Parzen window

Unlike parametric density estimation methods, non-parametric approaches locally estimate density function by a small number of neighboring samples [3] and therefore show less accurate estimation results. In spite of their accuracy, however, the performance of classifiers designed using these estimates is very satisfactory.

The basic idea for estimating unknown density function is based on the fact that the probability P that a vector $\mathbf{x}$ belongs to a region R [1]:

$$P = \int_R p(\mathbf{x}')d\mathbf{x}'.$$

It can be rewritten as

$$\int_R p(\mathbf{x}')d\mathbf{x}' \simeq p(\mathbf{x})V \simeq \frac{k}{n},$$

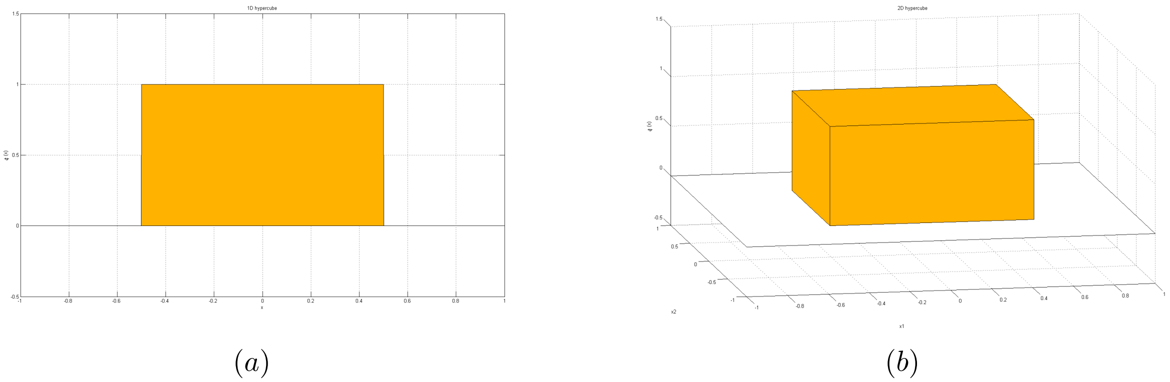
if we assume a small local region R, a large number of samples n, and k of n falling in R.
Suppose that the region R is a d-dimensional hypercube around $\mathbf{x}_i \in \mathbb{R}^n$ in the rest of this slecture, and let the volume V_n:

$$V_n = h_n^d$$

where h_n is the length of an edge. Then the window function for this hypercube can be defined by

$$\varphi(\mathbf{u}) = \begin{cases} 1, & |u_j| \leq \frac{1}{2} \quad j = 1, ..., d \\ 0, & \quad else. \end{cases}$$

and displayed in Figure 1(a) and 1(b) below such cases where $d = 1$ and $d = 2$, respectively.



(a) (b)

Figure 1: Given window function. (a) where $d = 1$, (b) where $d = 2$.

We simply shift this window function for $\mathbf{x}_i$ to determine if $\mathbf{x}_i$ belongs to the volume V_n, $\varphi(\frac{\mathbf{x}-\mathbf{x}_i}{h_n})$, and can compute the number of samples k_n falling in this volume using it:

$$k_n = \sum_{i=1}^{n} \varphi(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}).$$

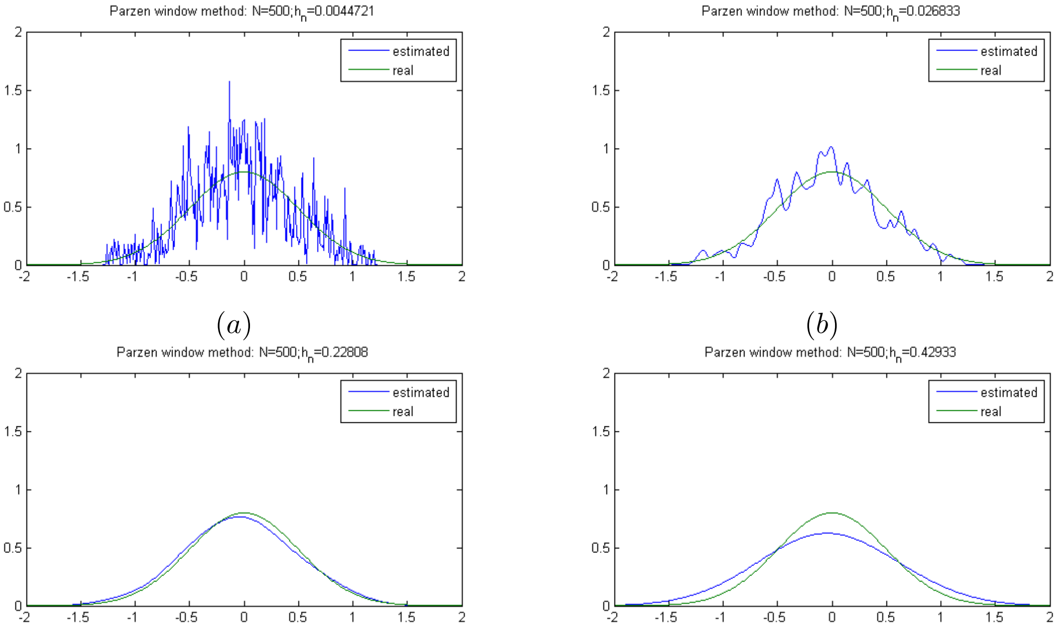
In Parzen window method, therefore, the estimate for density $p_n(\mathbf{x})$ is

$$p_n(\mathbf{x}) = \frac{k_n/n}{V_n} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{V_n}\varphi(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}).$$

In order to check how window length effects on $p_n(\mathbf{x})$, we define $\delta_n(\mathbf{x}-\mathbf{x}_i)$ by $\frac{1}{V_n}\varphi(\frac{\mathbf{x}-\mathbf{x}_i}{h_n})$ as an approximation of a unit impulse centered at $\mathbf{x}_i$ [2] and write $p_n(\mathbf{x})$ [1] by:

$$p_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \delta_n(\mathbf{x}-\mathbf{x}_i).$$

From this observation, we can infer the relationship between h_n and $p_n(\mathbf{x})$. If h_n is very large, the amplitude of δ_n is relatively small because $V_n = h_n^d$. Thus, $p_n(\mathbf{x})$ becomes a very smooth estimate for $p(\mathbf{x})$, see Figure 2(d). In contrast, if h_n is very small, the amplitude of δ_n is large. As a result, $p_n(\mathbf{x})$ is a noisy estimate for $p(\mathbf{x})$, see Figure 2(a). Therefore, when we have a limited number


Parzen window method: N=500;h_n=0.0044721

Parzen window method: N=500;h_n=0.026833

(a) (b)

Parzen window method: N=500;h_n=0.22808

Parzen window method: N=500;h_n=0.42933

(c) (d)

Figure 2: Density estimate using Parzen window where N = 500. (a) $h_n = 0.0044$, (b) $h_n = 0.0268$, (c) $h_n = 0.2280$, and (d) $h_n = 0.4293$.

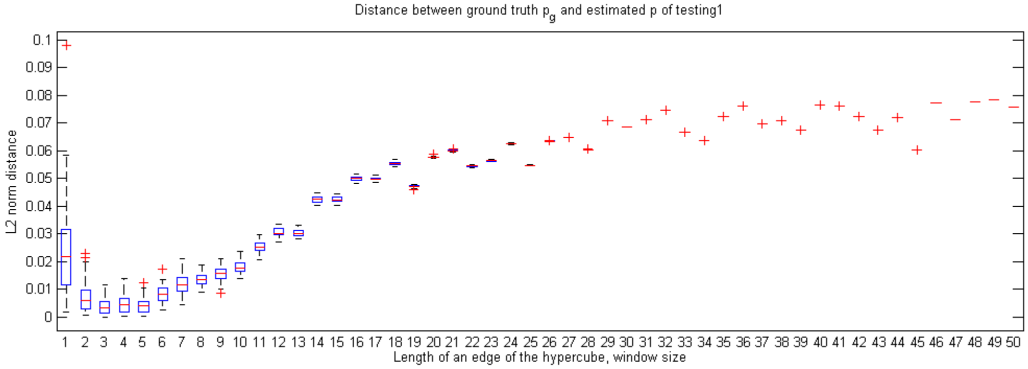


Figure 3: L2 norm distance between ground truth p and estimated p_n.

of samples, finding an optimal value of h_n is very important to accurately estimate $p(\mathbf{x})$. Let us assume that we have an unlimited number of samples. Then, V_n goes to zero as n increases, and hence $p_n(\mathbf{x})$ converges to $p(\mathbf{x})$ [1] no matter what h_n is.

In Figure 2, we can see the effect of the window length on the density estimate results. It demonstrates that a large or small h_n value causes inaccurate estimation for $p(\mathbf{x})$. Also, as shown in Figure 3, the optimal window length in this experiment can be obtained around 3rd-5th element (*i.e.*, $h_n = 0.15$-0.35) which shows minimum distance between ground truth p and estimated p_n. As stated earlier, the choice of h_n is very important, especially when the number of samples is small.

Convergence of the mean and variance

Since $\mathbf{x}$ is a vector of random samples $\mathbf{x}_1, ..., \mathbf{x}_n$, p_n has mean $E(p_n(\mathbf{x}))$ and variance $Var(p_n(\mathbf{x}))$. Thus, $p_n(\mathbf{x})$ converges to $p(\mathbf{x})$ [1], if

$$\lim_{n\to\infty} E(p_n(\mathbf{x})) = p(\mathbf{x}) \text{ and } \lim_{n\to\infty} Var(p_n(\mathbf{x})) = 0.$$

Let us prove the convergence of the mean and variance [2], respectively.

$$\begin{aligned} E(p_n(\mathbf{x})) &= \tfrac{1}{n}\sum_{i=1}^{n} E\left(\frac{1}{V_n}\varphi(\frac{\mathbf{x}-\mathbf{x}_i}{h_n})\right) \\ &= \tfrac{1}{n}\sum_{i=1}^{n} \int \frac{1}{V_n}\varphi(\frac{\mathbf{x}-\mathbf{x}_i}{h_n})p(\mathbf{x})d\mathbf{x} \\ &= \int \frac{1}{V_n}\varphi\left(-\frac{1}{h_n}(\mathbf{x}-\mathbf{x}_i)\right)p(\mathbf{x})d\mathbf{x} \\ &= \tfrac{1}{V_n}\varphi\left(-\tfrac{\mathbf{x}}{h_n} * p(\mathbf{x})\right) \\ \\ \text{if } (n\to\infty) \Leftrightarrow (h_n\to 0) &= \delta(-\mathbf{x}) * p(\mathbf{x}) \\ &= p(\mathbf{x}) \qquad \blacksquare \end{aligned}$$

$$\begin{array}{rcl} Var(p_n(\mathbf{x})) & = & Var\left(\sum_{i=1}^{n} \frac{1}{nV_n}\varphi(\frac{\mathbf{x}-\mathbf{x}_i}{h_n})\right) \\ & = & nVar\left(\frac{1}{nV_n}\varphi(\frac{\mathbf{x}-\mathbf{x}_i}{h_n})\right) \\ & = & n\left\{E\left(\frac{1}{n^2}\frac{1}{V_n^2}\varphi^2(\frac{\mathbf{x}-\mathbf{x}_i}{h_n})\right) - E\left(\frac{1}{nV_n}\varphi(\frac{\mathbf{x}-\mathbf{x}_i}{h_n})\right)^2\right\} \\ & \leq & nE\left(\frac{1}{n^2}\frac{1}{V_n^2}\varphi^2(\frac{\mathbf{x}-\mathbf{x}_i}{h_n})\right) \\ & = & \frac{1}{n}\int \frac{1}{V_n^2}\left(\varphi^2(\frac{\mathbf{x}-\mathbf{x}_i}{h_n})\right)p(\mathbf{x})d\mathbf{x} \\ & \leq & \frac{1}{nV_n}\sup\varphi \int \frac{1}{V_n}\left(\varphi(\frac{\mathbf{x}-\mathbf{x}_i}{h_n})\right)p(\mathbf{x})d\mathbf{x} \\ & = & \frac{1}{nV_n}\sup\varphi\left(p(\mathbf{x}) * \frac{1}{V_n}\varphi(\frac{\mathbf{x}}{h_n})\right) \\ & = & \frac{1}{nV_n}\sup\varphi\left(E(p_n(\mathbf{x})\right) \qquad \blacksquare \end{array}$$

In order to make this formula converges to zero as n goes to infinity, according to the lecture note [2], we can make $nV_n \to \infty$ and $V_n \to 0$ (*e.g.*, $V_n = \frac{1}{\sqrt{n}}$).

Classification using Parzen window method

A decision making using a classifier based on Parzen window estimation can be performed by simple majority voting method. Here, we check how it works. According to Professor Mireille Boutin [2], we pick the class such that $Prob(w_{i0}|x_0) \geq Prob(w_i|x_0) \forall i = 1, ..., c$ from Bayes' rule. In other words,

$$\begin{array}{rl} \Longleftrightarrow & \rho(x_0|w_{i0})Prob(w_{i0} \geq \rho(x_0|w_i)Prob(w_i) \\ \Longleftrightarrow & \rho(x_0, w_{i0}) \geq \rho(x_0, w_i) \\ \Longleftrightarrow & \sum_{\substack{i=1 \\ (\mathbf{x}_l \text{ in class } w_{i0})}}^{n} \varphi\left(\frac{\mathbf{x}_l - \mathbf{x}_0}{h_n}\right) \geq \sum_{\substack{i=1 \\ (\mathbf{x}_l \text{ in class } w_i)}}^{n} \varphi\left(\frac{\mathbf{x}_l - \mathbf{x}_0}{h_n}\right). \end{array}$$

We build a classifier using hypercube as a window function. Figure 4 illustrates the classification error rates with respect to the different number of samples while optimal window length obtained above is used. In this experiment, we can recognize that the error rate decreases as the number of samples in training dataset increases until it reaches about 190 samples. After that, the

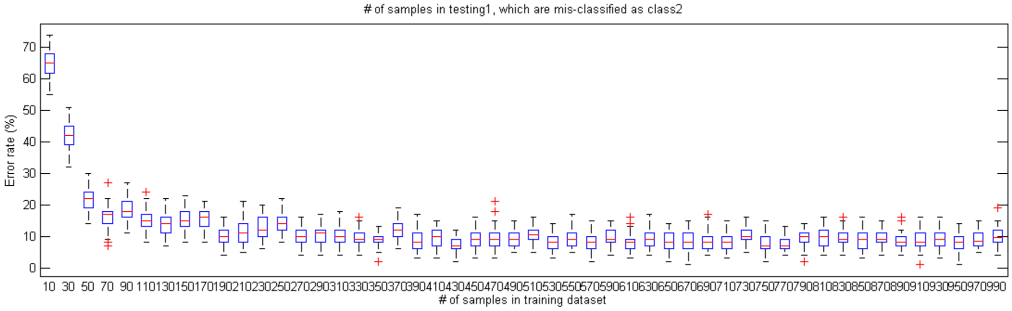


Figure 4: The classification error rates with respect to the number of samples from 10 to 1000.

classification error rate seems stable no matter what the sample size is. It demonstrates that the performance of classifiers designed using this method is satisfactory, even though density estimation is not accurate because density estimation performance is dependent on the samples size. As mentioned earlier, this is one of the advantages of non-parametric approaches.

Discussion

So far, we analyzed one of the non-parametric methods, Parzen Window density estimation, focused on how to estimate density, how to converge to actual density, and how to generate a classifier using it. As we proved both theoretically and experimentally in this slecture, Parzen window shows super ability for decision making without any assumptions about the distributions of given sample data. However, finding an appropriate window function which would show better performance is going to be a tedious work, that is one of the disadvantages of Parzen window method.

References

[1] Pattern classification. Richard O. Duda, Peter E. Hart, and David G. Stork.

[2] Lecture notes of ECE 662, Professor Mireille Boutin.

[3] Introduction to Statistical Pattern Recognition. K. Fukunaga.