Non-parametric Density Estimation: Introduction

• Useful parametric densities are limited in the shape they take on-- they may not fit your data well.
• Nonparametric procedures can be used with arbitrary distributions and without the assumption that the forms of the underlying densities are known

• There are two types of nonparametric methods:
  – Estimating $P(x \mid \mathcal{D}_j)$
  – Bypass probability and go directly to a-posteriori probability estimation
Density Estimation via Binning

– Basic idea:
  Probability that a vector x will fall in region is:

\[ P = \int p(x) \, dx \quad (1) \]

– \( P \) is a smoothed (or averaged) version of the density function \( p(x) \) if we have a sample of size \( n \); therefore, the probability that \( k \) points fall in is then:

\[ P_k = \binom{n}{k} P^k (1 - P)^{n-k} \quad (2) \]

and the expected value for \( k \) is:

\[ E(k) = nP \quad (3) \]
Histogram
ML estimation of $\hat{q} = P$

$\arg\max_{\square} P_k(\square)$ is reached for

$\hat{q} = \frac{k}{n} \square P$

Therefore, the ratio $k/n$ is a good estimate for the probability $P$ and hence for the density function $p$.

$p(x)$ is continuous and that the region $\mathcal{R}$ is so small that $p$ does not vary significantly within it, we can write:

$$\square p(x)dx = \bar{p}(x)V \square p(x')V$$

(4)

Where $x'$ is a point within $\mathcal{R}$ and $V$ the volume enclosed by $\mathcal{R}$. 
Combining equation (1), (3) and (4) yields: \[ p(x) \propto \frac{k}{n} \frac{1}{V} \]

**FIGURE 4.1.** The relative probability an estimate given by Eq. 4 will yield a particular value for the probability density, here where the true probability was chosen to be 0.7. Each curve is labeled by the total number of patterns \( n \) sampled, and is scaled to give the same maximum (at the true probability). The form of each curve is binomial, as given by Eq. 2. For large \( n \), such binomials peak strongly at the true probability. In the limit \( n \to \infty \), the curve approaches a delta function, and we are guaranteed that our estimate will give the true probability. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.
\[ \square p(x) \, dx = p(x') \square dx = p(x') \square 1 \, (x) \, dx = p(x') \square \left( \ \right) \]

Where: \( \square(R) \) is:
- an area in the Euclidean space \( \mathbb{R}^2 \)
- a volume in the Euclidean space \( \mathbb{R}^3 \)
- a hypervolume in the Euclidean space \( \mathbb{R}^n \)

Since \( p(x) \square p(x') = \text{constant} \), therefore in the Euclidean space \( \mathbb{R}^3 \):

\[ \square p(x) \, dx \square p(x') \cdot V \]

and \( p(x) \square \frac{k}{nV} \)
- Condition for convergence

The fraction \( k/(nV) \) is a space averaged value of \( p(x) \). \( p(x) \) is obtained only if \( V \) approaches zero. This is the case where no samples are included in \( R \): it is an uninteresting case!

\[
\lim_{V \to 0, n \text{ fixed}} p(x) = 0 \quad (\text{if } n = \text{ fixed})
\]

In this case, the estimate diverges: it is an uninteresting case!
The volume $V$ needs to approach 0 anyway if we want to use this estimation

- Practically, $V$ cannot be allowed to become small since the number of samples is always limited
- One will have to accept a certain amount of variance in the ratio $k/n$
- Theoretically, if an unlimited number of samples is available, we can circumvent this difficulty

To estimate the density of $x$, we form a sequence of regions $R_1, R_2, \ldots$ containing $x$: the first region contains one sample, the second two samples and so on.

Let $V_n$ be the volume of $R_n$, $k_n$ the number of samples falling in $R_n$ and $p_n(x)$ be the $n^{th}$ estimate for $p(x)$:

$$p_n(x) = (k_n/n)/V_n \quad (7)$$
Three necessary conditions should apply if we want $p_n(x)$ to converge to $p(x)$:

1) $\lim_{n \to \infty} V_n = 0$

2) $\lim_{n \to \infty} k_n = \infty$

3) $\lim_{n \to \infty} \frac{k_n}{n} = 0$

There are two different ways of obtaining sequences of regions that satisfy these conditions:

(a) Shrink an initial region where $V_n = \frac{1}{n}$ and show that

$$p_n(x) \xrightarrow{n \to \infty} p(x)$$

This is called “the Parzen-window estimation method”

(b) Specify $k_n$ as some function of $n$, such as $k_n = n$; the volume $V_n$ is grown until it encloses $k_n$ neighbors of $x$. This is called “the $k_n$-nearest neighbor estimation method”
Fig. 4.2. There are two leading methods for estimating the density at a point, here at the center of each square. The one shown in the top row is to start with a large volume centered on the test point and shrink it according to a function such as $V_n = 1 / \sqrt{n}$. The other method, shown in the bottom row, is to decrease the volume in a data-dependent way, for instance letting the volume enclose some number $k_n = \sqrt{n}$ of sample points. The sequences in both cases represent random variables that generally converge and allow the true density at the test point to be calculated. From: Richard O. Duda, Peter E. Hart, and David G. Stork, Pattern Classification. Copyright © 2001 by John Wiley & Sons, Inc.
Parzen Windows

– Parzen-window approach to estimate densities assume that the region $\mathcal{R}_n$ is a d-dimensional hypercube

$$V_n = h_n^d \ (h_n : length \ of \ the \ edge \ of \ \mathcal{R}_n )$$

Let $\Box(u)$ be the following window function:

$$\Box(u) = \begin{cases} 1 & |u_j| \leq \frac{1}{2} \ j = 1, \ldots, d \\ 0 & \text{otherwise} \end{cases}$$

– $\Box((x-x_i)/h_n)$ is equal to unity if $x_i$ falls within the hypercube of volume $V_n$ centered at $x$ and equal to zero otherwise.
- The number of samples in this hypercube is:

\[ k_n = \prod_{i=1}^{n} \frac{h_n}{x - x_i} \]

Which yields the probability estimate:

\[ p_n(x) = \frac{1}{n} \prod_{i=1}^{n} \frac{1}{V_n} \frac{h_n}{x - x_i} \]

\( p_n(x) \) estimates \( p(x) \) as an average of functions of \( x \) and the samples \( (x_i) \) \( (i = 1, \ldots, n) \). These functions \( \square \) can be general!
Illustration

The behavior of the Parzen-window method

Case where \( p(x) \rightarrow N(0,1) \)

Let \( \tilde{p}(u) = \frac{1}{\sqrt{2\pi}} \exp(-u^2/2) \) and \( h_n = h_1/n \) \((n>1)\) 

Thus:

\[
p_n(x) = \frac{1}{n} \prod_{i=1}^{n} \frac{1}{h_n} \frac{x}{h_n} x_i
\]

is an average of normal densities centered at the samples \( x_i \).
– Numerical results:

For \( n = 1 \) and \( h_1 = 1 \)

\[
p_1(x) = \mathbb{E}(x \mathbb{I} x_1) = \frac{1}{\sqrt{2\pi}} e^{-1/2} (x \mathbb{I} x_1)^2 \mathbb{I} \ N(x_1, 1)
\]

For \( n = 10 \) and \( h = 0.1 \), the contributions of the individual samples are clearly observable!
FIGURE 4.5. Parzen-window estimates of a univariate normal density using different window widths and numbers of samples. The vertical axes have been scaled to best show the structure in each graph. Note particularly that the $n = \infty$ estimates are the same (and match the true density function), regardless of window width. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.
Analogous results are also obtained in two dimensions as illustrated:
FIGURE 4.6. Parzen-window estimates of a bivariate normal density using different window widths and numbers of samples. The vertical axes have been scaled to best show the structure in each graph. Note particularly that the $n = \infty$ estimates are the same (and match the true distribution), regardless of window width. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.
Case where \( p(x) = \square_1 U(a,b) + \square_2 T(c,d) \) (unknown density) (mixture of a uniform and a triangle density)
FIGURE 4.7. Parzen-window estimates of a bimodal distribution using different window widths and numbers of samples. Note particularly that the $n = \infty$ estimates are the same (and match the true distribution), regardless of window width. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.
– Classification example

In classifiers based on Parzen-window estimation:

1. We estimate the densities for each category and classify a test point by the label corresponding to the maximum posterior.

2. The decision region for a Parzen-window classifier depends upon the choice of window function as illustrated in the following figure.
FIGURE 4.8. The decision boundaries in a two-dimensional Parzen-window dichotomizer depend on the window width $h$. At the left a small $h$ leads to boundaries that are more complicated than for large $h$ on same data set, shown at the right. Apparently, for these data a small $h$ would be appropriate for the upper region, while a large $h$ would be appropriate for the lower region; no single window width is ideal overall. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification.* Copyright © 2001 by John Wiley & Sons, Inc.
Nearest Neighbor Approach

• Problem with Parzen (kernel):
  • Unknown “best” window function

• Nearest Neighbor Approach:
  • let the cell volume be a function of the training data, by centering a cell about each point \( x \) and increasing the volume until \( k_n \) samples are contained, where \( k_n \) depends on \( n \).
  • **These samples are the** \( k_n \) **nearest-neighbors of** \( x \).

\[
p_n(x) = \frac{k_n/n}{V_n}
\]
FIGURE 4.2. There are two leading methods for estimating the density at a point, here at the center of each square. The one shown in the top row is to start with a large volume centered on the test point and shrink it according to a function such as $V_n = 1/\sqrt{n}$. The other method, shown in the bottom row, is to decrease the volume in a data-dependent way, for instance letting the volume enclose some number $k_n = \sqrt{n}$ of sample points. The sequences in both cases represent random variables that generally converge and allow the true density at the test point to be calculated. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.
Figure 4.10: Eight points in one dimension and the $k$-nearest-neighbor density estimates, for $k = 3$ and 5. Note especially that the discontinuities in the slopes in the estimates generally occur away from the positions of the points themselves.
• The $k$-nearest-neighbor estimate of a two-dimensional density for $k = 5$. 
Estimation of a posteriori Prob

\[ p_n(x, \omega_i) = \frac{k_i}{n}, \]

\[ P_n(\omega_i|x) = \frac{p_n(x, \omega_i)}{\sum_{j=1}^{c} p_n(x, \omega_j)} = \frac{k_i}{k}. \]

Thus, the estimate is just the fraction of the samples in a cell from the \( i^{th} \) class.
\[ P(\omega_m|x) = \max_i P(\omega_i|x), \]

Figure 4.13: In two dimensions, the nearest-neighbor algorithm leads to a partitioning of the input space into Voronoi cells, each labelled by the category of the training point it contains. In three dimensions, the cells are three-dimensional, and the decision boundary resembles the surface of a crystal.