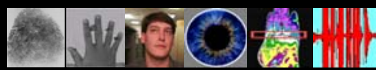
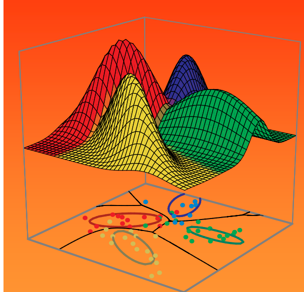


Non-Parametric Density Estimation Techniques

Biometrics
CSE 190-a
Lecture 7

Announcements

- HW1 assigned
- Most of this lecture was on the blackboard. These slides cover the same material as presented in DHS



Pattern Classification

All materials in these slides were taken from *Pattern Classification (2nd ed) by R. O. Duda, P. E. Hart and D. G. Stork, John Wiley & Sons, 2000*
with the permission of the authors and the publisher

Chapter 4 (Part 1): Non-Parametric Classification (Sections 4.1-4.3)

- Introduction
- Density Estimation
- Parzen Windows

Introduction

- All Parametric densities are unimodal (have a single local maximum), whereas many practical problems involve multi-modal densities
- 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 / ω_j)
  - Bypass probability and go directly to a-posteriori probability estimation

Density Estimation

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

$$P = \int_{\Re} p(x')dx' \qquad (1)$$

- 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 R is then:

$$P_k = \binom{n}{k} P^k (1-P)^{n-k} \qquad (2)$$

and the expected value for k is:

$$E(k) = nP \qquad (3)$$

6

ML estimation of $P = \theta$

$\underset{\theta}{Max}(P_k \mid \theta)$ is reached for $\hat{\theta} = \frac{k}{n} \cong P$

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 *R* is so small that p does not vary significantly within it, we can write:

$$\int_{\Re} p(x')dx' \cong p(x)V \qquad (4)$$

where is a point within *R* and V the volume enclosed by *R*.

7

Combining equation (1) , (3) and (4) yields: $p(x) \cong \frac{k/n}{V}$

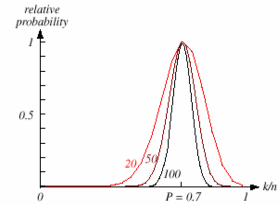


FIGURE 4.1. The relative probability an estimate given by Eq. 4 will yield a particular value for the probability density, here where the true probability was chosen to be 0.7. Each curve is labeled by the total number of patterns *n* sampled, and is scaled to give the same maximum (at the true probability). 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.

8

Density Estimation (cont.)

- Justification of equation (4)

$$\int_{\Re} p(x')dx' \cong p(x)V \qquad (4)$$

We assume that *p(x)* is continuous and that region *R* is so small that p does not vary significantly within *R*. Since *p(x)* = constant, it is not a part of the sum.

9

$$\int_{\Re} p(x')dx' = p(x')\int_{\Re} dx' = p(x')\int_{\Re} 1_{\Re}(x)dx' = p(x')\mu(\Re)$$

Where: $\mu(R)$ is: a surface in the Euclidean space R^2
a volume in the Euclidean space R^3
a hypervolume in the Euclidean space R^n

Since $p(x) \cong p(x')$ = constant, therefore in the Euclidean space R^3:

$$\int_{\Re} p(x')dx' \cong p(x).V$$

$$\text{and} \quad p(x) \cong \frac{k}{nV}$$

10

- Condition for convergence

The fraction *k/(nV)* is a space averaged value of *p(x)*. *p(x)* is obtained only if V approaches zero.

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

This is the case where no samples are included in *R*: it is an uninteresting case!

$$\lim_{V \to 0, k \neq 0} p(x) = \infty$$

In this case, the estimate diverges: it is an uninteresting case!

11

- 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)*:

$$p_n(x) = (k_n/n)/V_n \qquad (7)$$

Three necessary conditions should apply if we want $p_n(x)$ to converge to $p(x)$:

$$1)\ \lim_{n\to\infty} V_n = 0$$
$$2)\ \lim_{n\to\infty} k_n = \infty$$
$$3)\ \lim_{n\to\infty} k_n / n = 0$$

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

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

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

This is called "the Parzen-window estimation method"

(b) Specify k_n as some function of n, such as $k_n = \sqrt{n}$; the volume V_n is grown until it encloses k_n neighbors of x. This is called "the k_n-nearest neighbor estimation method"

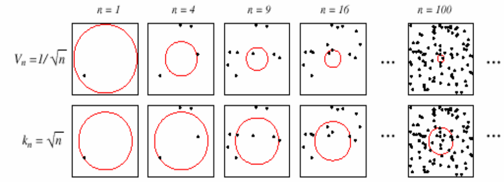


FIGURE 4.2. There are two leading methods for estimating the density at a point, here at the center of each square. The one shown in the top row is to start with a large volume centered on the test point and shrink it according to a function such as $V_n = 1/\sqrt{n}$. The other method, shown in the 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.

Parzen Windows

- Parzen-window approach to estimate densities assume that the region R_n is a d-dimensional hypercube

$$V_n = h_n^d \ (h_n : \text{length of the edge of } \Re_n)$$

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

$$\varphi(u) = \begin{cases} 1 & |u_j| \le \frac{1}{2} \quad j = 1,\dots,d \\ 0 & \text{otherwise} \end{cases}$$

- $\varphi((x-x_i)/h_n)$ 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:

$$k_n = \sum_{i=1}^{i=n} \varphi\left(\frac{x - x_i}{h_n}\right)$$

By substituting k_n in equation (7), we obtain the following estimate:

$$p_n(x) = \frac{1}{n}\sum_{i=1}^{i=n} \frac{1}{V_n}\varphi\left(\frac{x - x_i}{h_n}\right)$$

$P_n(x)$ estimates $p(x)$ as an average of functions of x and the samples (x_i) $(i = 1,\dots,n)$. These functions φ can be general!

- Parzen Window Example
  - Draw samples from a Normal distribution, $N(0,1)$
  - Let $\varphi(u) = (1/\sqrt{(2\pi)}) \exp(-u^2/2)$
    $h_n = h_1/\sqrt{n}$ $(n>1)$

Thus:

$$p_n(x) = \frac{1}{n}\sum_{i=1}^{i=n} \frac{1}{h_n}\varphi\left(\frac{x - x_i}{h_n}\right)$$

is an average of normal densities centered at the samples x_i.

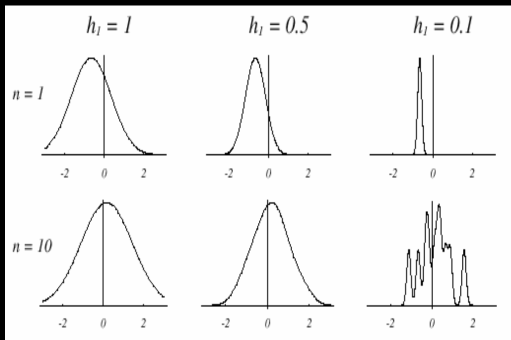
- Numerical results:

For $n = 1$ and $h_1=1$

$$p_1(x) = \varphi(x - x_1) = \frac{1}{\sqrt{2\pi}} e^{-1/2}(x - x_1)^2 \rightarrow N(x_1, 1)$$

For $n = 10$ and $h = 0.1$, the contributions of the individual samples are clearly observable !

18



19

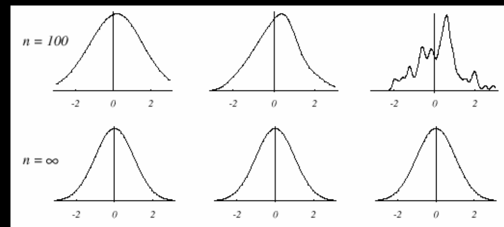
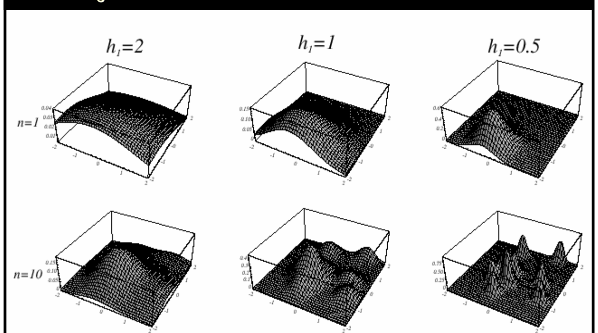


FIGURE 4.5. Parzen-window estimates of a univariate normal density using different window widths and numbers of samples. The vertical axes have been scaled to best show the structure in each graph. Note particularly that the $n = \infty$ estimates are the same (and match the true density function), regardless of window width. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

20

Analogous results are also obtained in two dimensions as illustrated:



21

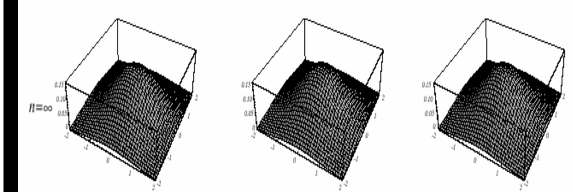
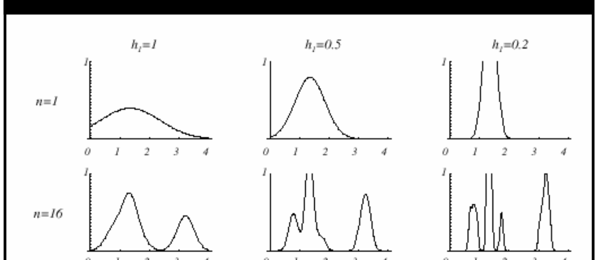


FIGURE 4.6. Parzen-window estimates of a bivariate normal density using different window widths and numbers of samples. The vertical axes have been scaled to best show the structure in each graph. Note particularly that the $n = \infty$ estimates are the same (and match the true distribution), regardless of window width. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

22

- Case where $p(x) = \lambda_1.U(a,b) + \lambda_2.T(c,d)$ (unknown density) (mixture of a uniform and a triangle density)



23

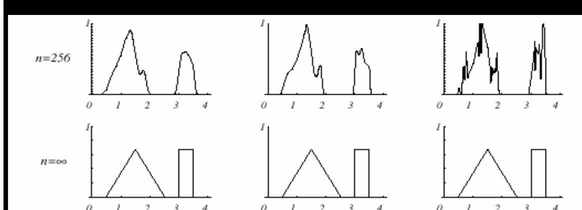


FIGURE 4.7. Parzen-window estimates of a bimodal distribution using different window widths and numbers of samples. Note particularly that the $n = \infty$ estimates are the same (and match the true distribution), regardless of window width. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

24

- K_n - Nearest neighbor estimation
  - Goal: a solution for the problem of the unknown "best" window function
    - Let the cell volume be a function of the training data
    - Center a cell about x and let it grows until it captures k_n samples ($k_n = f(n)$)
    - k_n are called the k_n nearest-neighbors of x

  2 possibilities can occur:
    - Density is high near x; therefore the cell will be small which provides a good resolution
    - Density is low; therefore the cell will grow large and stop until higher density regions are reached

  We can obtain a family of estimates by setting $k_n=\sqrt{n}$

25

Illustration

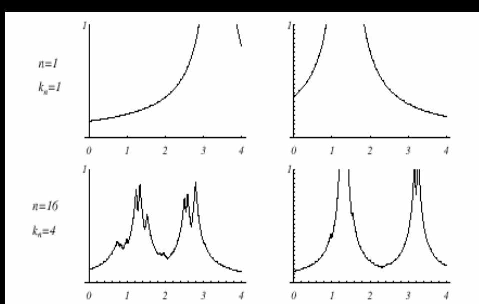
For $n= 1$ and $k_n = \sqrt{n} = 1$; the estimate becomes:

$P_n(x) = k_n / n.V_n = 1 / V_1 = 1 / 2|x-|$

Yikes! Well not so good as the probability goes to infinity at x_1 but at least we do not have holes in the density!

Things get better as n gets bigger! And we still don't have holes in the density even for higher dimensions!

26



27

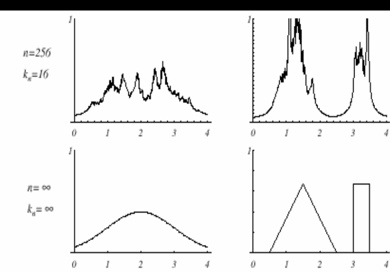


FIGURE 4.12. Several k-nearest-neighbor estimates of two unidimensional densities: a Gaussian and a bimodal distribution. Notice how the finite n estimates can be quite "spiky." From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

28

- Estimation of a-posteriori probabilities
  - Goal: estimate $P(\omega_i / x)$ from a set of n labeled samples
    - Let's place a cell of volume V around x and capture k samples
    - k_i samples amongst k turned out to be labeled ω_i then:

      $p_n(x, \omega_i) = k_i / n.V$

    An estimate for $p_n(\omega_i / x)$ is:

$$p_n(\omega_i \mid x) = \frac{p_n(x, \omega_i)}{\sum_{j=1}^{j=c} p_n(x, \omega_j)} = \frac{k_i}{k}$$

29

- k_i/k is the fraction of the samples within the cell that are labeled ω_i
- For minimum error rate, the most frequently represented category within the cell is selected
- If k is large and the cell sufficiently small, the performance will approach the best possible
- So whether we use Parzen windows (or K-th nearest neighbors to determine our window size V_n , we can directly get the a posteriori probabilities.

- **The nearest –neighbor rule**
  - Let $D_n = \{x_1, x_2, \ldots, x_n\}$ be a set of n labeled prototypes
  - Let $x' \in D_n$ be the closest prototype to a test point x then the nearest-neighbor rule for classifying x is to assign it the label associated with x'
  - The nearest-neighbor rule leads to an error rate greater than the minimum possible: the Bayes rate
  - If the number of prototype is large (unlimited), the error rate of the nearest-neighbor classifier is never worse than twice the Bayes rate (it can be demonstrated!)
  - If $n \to \infty$, it is always possible to find x' sufficiently close so that:

$$P(\omega_i \mid x') \cong P(\omega_i \mid x)$$

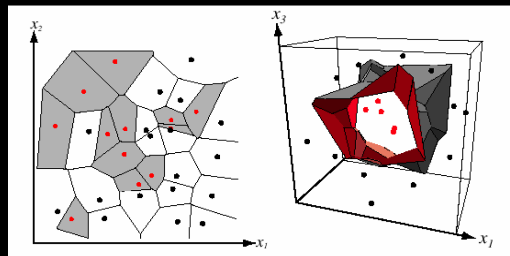


FIGURE 4.13. In two dimensions, the nearest-neighbor algorithm leads to a partitioning of the input space into Voronoi cells, each labeled by the category of the training point it contains. In three dimensions, the cells are three-dimensional, and the decision boundary resembles the surface of a crystal. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

- The k – nearest-neighbor rule
  - Goal: Classify x by assigning it the label most frequently represented among the k nearest samples and use a voting scheme

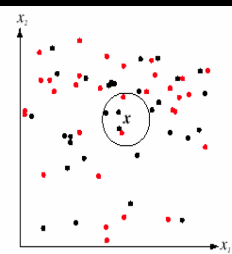


FIGURE 4.15. The k-nearest-neighbor query starts at the test point **x** and grows a spherical region until it encloses k training samples, and it labels the test point by a majority vote of these samples. In this $k = 5$ case, the test point **x** would be labeled the category of the black points. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.