Announcements

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

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 multimodal 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 $P(x)$
  - 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:
    $$P = \frac{\int p(x')dx'}{n}$$  \hspace{1cm} (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:
    $$P_k = \frac{n}{k} P^k (1-P)^{n-k}$$  \hspace{1cm} (2)
  - and the expected value for $k$ is:
    $$E(k) = nP$$  \hspace{1cm} (3)
ML estimation of $P = \theta$ is reached for $
abla P = k/n = P$.

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

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

$$\int_{R} p(x') \, dx' \cong p(x) V$$  \hspace{1cm} (4)

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

Combining equation (1), (3) and (4) yields:

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

**Density Estimation (cont.)**

- **Justification of equation (4)**

$$\int_{R} p(x') \, dx' \cong p(x) V$$  \hspace{1cm} (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.

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!

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

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

- **Condition for convergence**

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

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

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

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

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 $R_1, R_2, ..., R_n$: the first region contains one sample, the second two samples, and so on.

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

$$p(x) = \frac{k}{nV}$$  \hspace{1cm} (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 \( V_n = 1/n \) and show that

\[
p_n(x) \to p(x)
\]

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 \( k \)-nearest neighbor estimation method".

**Parzen Windows**

- Parzen-window approach to estimate densities assumes that the region \( R_n \) is a \( d \)-dimensional hypercube

\[
V_n = h_n^d \quad (h_n: \text{length of the edge of } R_n)
\]

Let \( \phi(u) \) be the following window function:

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

- \( \phi((x-x)/h_n) \) is equal to unity if \( x \) falls within the hypercube of volume \( V_n \) centered at \( x \) and equal to zero otherwise.

**Parzen Window Example**

- Draw samples from a Normal distribution, \( N(0,1) \)

Let \( \phi(u) = (1/(\sqrt{2\pi})) \exp(-u^2/2) \)

\[ h_n = h/\sqrt{n} \quad (n>1) \]

Thus:

\[
p_n(x) = \frac{1}{n} \sum_{i=1}^{n} \phi \left( \frac{x-x_i}{h_n} \right)
\]

is an average of normal densities centered at the samples \( x \).

**Numerical results:**

For \( n = 1 \) and \( h = 2 \)

\[
p_1(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \to N(0,1)
\]

For \( n = 10 \) and \( h = 0.1 \), the contributions of the individual samples are clearly observable!
Analogous results are also obtained in two dimensions as illustrated:

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

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 better show the structure in each graph. Note particularly that the $n = \infty$ estimates are the same and match the true density functions, regardless of window width. From: Richard O. Duda, Peter E. Hart, and David G. Stork. Pattern Classification. Copyright © 2001 by John Wiley & Sons, Inc.

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 better 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.

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.
• **Kₙ - 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 grow until it captures kᵢ samples \( (h = \sqrt{n}) \)
  • \( kᵢ \) are called the \( kᵢ \) 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 \( k &= \sqrt{n} \)

**Illustration**

For \( n = 1 \) and \( kᵢ = \sqrt{n} = 1 \); the estimate becomes:

\[
P_k(x) = kᵢ / n \cdot V
\]

Yikes! Well not so good as the probability goes to infinity at \( x₁ \) 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!

**Estimation of a-posteriori probabilities**

• **Goal:** estimate \( P(ωᵢ / x) \) from a set of \( n \) labeled samples

  • Let’s place a cell of volume \( V \) around \( x \) and capture \( k \) samples amongst \( k \) turned out to be labeled \( ωᵢ \) then:

  \[
p_k(x, ωᵢ) = kᵢ / n \cdot V
\]

  An estimate for \( p_k(ωᵢ / x) \) is:

  \[
p_ω(ωᵢ / x) = \frac{p_k(x, ωᵢ)}{\sum_{j=1}^{k} p_k(x, \omega_j)} = \frac{kᵢ}{k}
\]

• \( kᵢ/V \) is the fraction of the samples within the cell that are labeled \( ωᵢ \)

• 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 \( Vₙ \), we can directly get the a posteriori probabilities.
• The nearest neighbor rule

Let $D_n = \{ x_1, x_2, ..., x_n \}$ 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:

$P(\omega_i | x') = P(\omega_i | x)$

• 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 4.13:** 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.

**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 the $k = 5$ 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.