8001652 Introduction to Pattern Recognition. Lecture 8: Density Estimation and Parzen Windows

Jussi Tohka

jussi.tohka@tut.fi

Institute of Signal Processing

Tampere University of Technology

Introduction

- Previously, we have studied pattern classification and supervised learning under the assumption that the parametric forms of underlying (class conditional) density functions were known.
- However, in practical pattern recognition this assumption is in suspect.
- During this lecture and the next one, we will consider *nonparametric* techniques for learning and classification. These can be used with arbitrary distributions and without knowing the parametric form of the underlying (class conditional) densities.

Types of nonparametric methods

1. Estimation of density functions $p(\mathbf{x}|\omega_i)$ using sample patterns.
2. Estimation of a posteriori probabilities $P(\omega_i|\mathbf{x})$ directly based on sample patterns or prototypes.

Preliminaries - combinations

- Consider the problem of selecting k labeled balls without replacement from an urn containing n balls.
- In how many different ways may we select those k balls?
- The answer is $C(n,k) = \left(\begin{array}{c} n \\ k \end{array} \right) = \frac{n!}{k!(n-k)!}$
- $n! = n \cdot (n-1) \cdots 2 \cdot 1$, (convention $0! = 1$)

Preliminaries - combinations

- More formally: If a set contains n elements, then there exist

$$C(n,k) = \frac{n!}{k!(n-k)!}$$

combinations (subsets) containing k elements.

- *Proof*: Consider first selecting a subset in the case the order of selection is important. For the first element a_1 in a subset there exist n possible choices. For the second elements a_2, there are $n-1$ possible choices and so on. Finally, for the kth element, there are $n-k+1$ choices left. Putting this all together, there exist $n \cdot (n-1) \cdots (n-k+1)$ ordered subsets. Now, since an unordered set of k elements corresponds to $k!$ ordered subsets, the number of possible subsets is $\frac{n \cdot (n-1) \cdots (n-k+1)}{k!} = \frac{n!}{k!(n-k)!}$.

Density estimation

- Consider the problem of estimating the density $p(\mathbf{x})$ given a set of training samples $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$.
- The basic idea of the most fundamental techniques is to divide the feature space in small regions $R_1, \ldots, R_k$, compute the number of training samples in that region and derive the density based on these counts.

Density estimation

- Let us consider the technique further: The probability that a training sample $\mathbf{x}_i$ will fall in a region R is

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

- Remember now we do not know the density $p(\mathbf{x})$, it is to be estimated.
- From Eq. (1) we see that P is smoothed or averaged version of the density function $p(\mathbf{x})$.
- We can estimate an averaged $p(\mathbf{x})$ by estimating P.

Density estimation

- Suppose now that we have i.i.d. training samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ distributed according to $p(\mathbf{x})$.
- The probability of k of these samples falling into region R is given by the binomial density

(2) $$P_k = \begin{pmatrix} n \\ k \end{pmatrix} P^k (1-P)^{n-k}$$

- And the expected value for k is $E[k] = \sum_{k=0}^{n} k P_k = nP$.(Based on binomial theorem.)
- Estimating $E[k]$ by the observed k leads to estimating $P = k/n$.

Density estimation

- Let us continue by writing $\int_R p(\mathbf{x})d\mathbf{x} \simeq p(\mathbf{x}^*)V$, where V is the volume enclosed by R and $\mathbf{x}^*$ is a point within R.
- Combining
  1. $P = k/n$
  2. $P = \int_R p(\mathbf{x})d\mathbf{x}$
  3. $\int_R p(\mathbf{x})d\mathbf{x} \simeq p(\mathbf{x}^*)V$

  we then arrive at the obvious estimate

(3) $$p(\mathbf{x}) = \frac{k}{nV}$$

within R.

Problems

- If we fix the volume V and take more and more training samples, the ratio k/n will converge as desired, but then we have only obtained an estimate of the space-averaged value of the true density function p.
- If we would like an estimate of p we would have to let V approach zero. However, when working with fixed n, this would lead to useless estimate $p(\mathbf{x}) \simeq 0$.
- From the practical viewpoint, we can only say that the number of training samples is always limited, and V cannot be arbitrarily small.

Theory: Unlimited amount of samples

- In theory, we may ask what happens if we have an unlimited number of training samples.
- To estimate the density $p(\mathbf{x})$ at $\mathbf{x}$, we create sequence of regions $R_1, R_2, \ldots$ containing $\mathbf{x}$ - the first region to be used with one training sample, the second with two and so on.
- Moreover, let V_n denote the volume of R_n, k_n is the number of samples falling into R_n, and nth estimate of the density

$$p_n(\mathbf{x}) = \frac{k_n}{nV_n}.$$

Theory: Unlimited amount of samples

- If we would like to $\lim_{n\to\infty} p_n(\mathbf{x}) = p(\mathbf{x})$ (in probability), three conditions are required:
  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 ways of obtaining the sequence of regions that satisfy these constraints. 1) Shrink an initial region by specifying V_n as some function of n. (Parzen windows) 2) Specify k_n as some function of n and let V_n grow until it encloses k_n neighbors of $\mathbf{x}$ (k_n nearest neighbors).

Parzen windows - An example

- Assume temporarily that region R_n is a d-dimensional hypercube.
- If h_n is the length of the side of the hypercube, its volume is given by $V_n = h_n^d$.
- We can obtain an analytic expression for k_n - the number of samples falling into the hypercube - by defining the following *window function*:

(4) $$\varphi(\mathbf{u}) = \begin{cases} 1 & |u_j| \leq 1/2 \\ 0 & \text{otherwise} \end{cases}$$

- That is φ has the value one inside and the value zeros outside the unit hypercube centered at origin.

Parzen windows - An example

- It follows that $\varphi((\mathbf{x} - \mathbf{x}_i)/h_n) = 1$ if $\mathbf{x}_i$ falls in the hypercube of volume V_n centered at $\mathbf{x}$. And $\varphi((\mathbf{x} - \mathbf{x}_i)/h_n) = 0$ otherwise.
- The number of samples in this hypercube is therefore given by

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

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

Parzen windows

- If we consider also other window functions, we obtain a more general approach to density estimation.
- The Parzen-window density estimate using n training samples and the window function φ is defined by

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

- The estimate $p_n(\mathbf{x})$ is an average of (window) functions of $\mathbf{x}$ and the training samples $\mathbf{x}_i$.
- In essence, the window function is then used for interpolation - each training sample contributing to the estimate in accordance with its distance from $\mathbf{x}$.

Window functions

- We want $p_n(\mathbf{x})$ to be legitimate density, i.e. 1) $p_n(\mathbf{x}) \geq 0$ for all $\mathbf{x}$ and 2) $\int p_n(\mathbf{x}) d\mathbf{x} = 1$.
- If we maintain the relation $h_n^d = V_n$, this is guaranteed if the window function is a legitimate density:
  1. $\varphi(\mathbf{x}) \geq 0$ for all $\mathbf{x}$
  2. $\int \varphi(\mathbf{x}) d\mathbf{x} = 1$

Window width

- How should we select the window width h_n?
- If h_n is too large, the density estimate $p_n(\mathbf{x})$ will be very smooth and 'out-of-focus'.
- If h_n is too small, the estimate $p_n(\mathbf{x})$ will be just superposition of n sharp pulses centered at training samples, i.e. an erratic noisy estimate of the true density.
- In practise, we have to seek some acceptable compromise since the number of training samples is always limited and moreover we may not be able to affect the number of available training samples.

Convergence of Parzen window estimates

- With an unlimited number of training samples it is possible to let V_n approach zero, and have $p_n(\mathbf{x})$ converge to $p(\mathbf{x})$.
- By convergence we mean that for all $\mathbf{x}$
  1. $\lim_{n\to\infty} E[p_n(\mathbf{x})] = p(\mathbf{x})$,
  2. $\lim_{n\to\infty} Var[p_n(\mathbf{x})] = 0$.
- That means we want to obtain correct estimates on the average, and the variance within these estimates should be negligible as the number of training samples approaches infinity. Expectations are taken with respect to the sequence (of length n) of training samples.

Convergence of Parzen window estimates

- In order to guarantee the convergence, we must place conditions on the unknown density, the window function $\varphi(\mathbf{x})$, and the window width h_n:
- The density function must be continuous.
- The window function must be bounded.
- The values of the window function must be very small at infinity.
- $V_n \to 0$ when $n \to \infty$.
- $nV_n \to \infty$ when $n \to \infty$.