An Application of Kernel Density Estimation in a Classification Problem

Nusrat Jahan
Department of Mathematics and Statistics
Mississippi State University,Mississippi State, MS 39762 USA
email:njahan@ra.msstate.edu

Abstract

A fundamental problem with sample data is that the underlying distribution is unknown. Therefore any modelling of the data or any kind of analysis has to be carried out based on the estimated distribution of the data. kernel density estimation technique is one of the widely used technique for density estimation. This is a non parametric technique. Therefore it does not impose any rigid distributional assumption on the data. When we have a data set containing data from different densities(or classes), the problem becomes identifying which data belongs to which class or density. Using kernel method, the maximum likelihood value for each data is computed for each class. The class that produces highest maximum likelihood, it is highly likely that the data point belong to that class than any other class.

1 Introduction

If we have a set of observed data points assumed to be a sample from an unknown probability density then there are primarily two different ways of estimating the underlying density [1].

(i) Parametric Approach : If it can be reasonably assumed that the observed sample is drawn from a known parametric family of distribution, then the values of corresponding parameters can be estimated from the observed data and substituted in the density function.

(ii) Nonparametric Approach : In this approach, no rigid assumption about the distribution of the observed data will be made. The underlying density will be estimated from the observed sample data.

There are various nonparametric approaches to density estimation [2]. A brief introduction of some those methods are provided here. Let $x_1, x_2, \ldots, x_n$ be independently and identically distributed random observations from unknown density, then there estimated density is denoted by $\hat{f}$.The methods discussed here are primaily for univariate data, but they can be easily generalized to multivariate case. Histogram : the most

widely used density estimator is the histogram.

$$\hat{f}(x) = \frac{1}{n} x \frac{(\text{no.of } X_i \text{ in same bin as x})}{(\text{width of bin containing x})}$$

The naive estimator :

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} w\left(\frac{x - X_i}{h}\right)$$

where h is a small number and w(x) is a weight function defined as,

$$w(x) = \begin{cases} \frac{1}{2} & \text{if } x \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

The kernel Estimator :

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

where h is the window width, that is the smoothing parameter or bandwidth and the kernel function K satisfies the condition,

$$\int_{-\infty}^{\infty} K(x)dx = 1.$$

The nearest neighbour method : If the density at t is $f(t)$, then for a sample of size n, exactly $(k-1)$ observations will fall in the interval $[t - d_k(t), t + d_k(t)]$, where $d_k(t)$ is the kth distance from t to the points of the sample. Therefore an estimate of the density at t can be derived using,

$$k - 1 = 2d_k(t)n\hat{f}(t)$$

so kth nearest neighbour density estimate is defined by,

$$\hat{f}(x) = \frac{(k-1)}{2nd_k(t).}$$

Variable Kernel method : The density estimate is constructed in the same way as it is done in kernel estimation method, except the scale parameter is allowed to vary from one data point to another.

Orthogonal Series Estimators : In this approach the density function is estimated by estimating the coefficients of its Fourier expansion.

Maximum Penalized Likelihood Estimators : This method places restrictions on the class of densities over which the likelihood is to be maximized.

For this study, kernel density estimation technique (for multivariate data) combined with maximum likelihood method has been used.

In this article, section 2 presents how kernel density estimation works. Section 3 discusses the results from the data sets. An overall summarization section ends this paper.

2 Kernel Density Estimation

The kernel estimator for univariate case (discussed in section 1) can easily be generalized for multivariate case [2]. In this case we assume, $x_1, x_2, \ldots, x_n$ be independently and identically distributed d-dimensional random vectors from unknown density, then there estimated density is denoted by,

$$\hat{f}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

where h is the window width (smoothing parameter), d is the dimension of the data vector x and the kernel function K(x) is now a function, defined for d-dimensional x which satisfies the condition,

$$\int_{-\infty}^{\infty} K(x)dx = 1.$$

Usually the kernel function is chosen to be symmetric unimodal probability density function. If we look at the definition of kernel estimator it is easily understood that the smoothness of the estimated density depends on the choice of h and the form of the estimated density depends on the form of the kernel function.

Properties of Kernel Estimate $\hat{f}$

(i) The kernel K is a density function and it is non-negative everywhere, therefore the estimated $\hat{f}$ will be a probability density function [3] .

(ii) The $\hat{f}$ will contain all the continuity and differentiability properties of the kernel K [3].

The kernel estimator becomes inefficient in case of long-tailed distribution. The reason for this is because the window width is fixed across the entire sample, sometimes spurious noise appear at the tail part of the estiamted density [2].

Couple of widely used kernel functions are given below [2].

Standard Multivariate Normal Density :

$$K(x) = (2\pi)^{d/2} exp(-\frac{1}{2}x^T x).$$

Multivariate Epanechnikov Kernel :

$$K_e(x) = \begin{cases} \frac{1}{2}c_d^{-1}(d+2)(1-x^TX) & \text{if } x^Tx \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

Where c_d is the volume of the unit d-demensional sphere. For the special case of $d = 2$, some useful kernels are available. These are

$$K(x) = \begin{cases} 3\pi^{-1}(1-x^Tx)^2 & \text{if } x^Tx \leq 1 \\ 0 & \text{otherwise;} \end{cases}$$

and

$$K(x) = \begin{cases} 4\pi^{-1}(1-x^Tx)^3 & \text{if } x^Tx \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

The choice of smoothing parameter, that is bandwidth is very crucial to density estimation. If it is too small, then the estimated density will be rough, local maxima and minima will dominate. On the other hand if h, the bandwith is too big, the density will be smooth, but some specific characteristics will be lost due to smoothness. So in choosing h one has to balance the desired smoothness with the desire to see the effect of actual data points. An optimal window width for smoothing of normally distributed data with unit variance has been suggested by [1],

$$h_{opt} = A(K)n^{-1/(d+4)}$$

where A(K) depends on the form of the kernel function. For a general data set with estiamted covariance matrix S, the numerical values for $A(K)$ can be as given below for different kernels.

Multivariate Normal K :

$$A(K) = 4/(2d+1)^{1/(d+4)}$$

Multivariate Epanechnikov K_e :

$$A(K) = \frac{8d(2d+1)(d+4)(2\sqrt{\pi})^d}{(2d+1)c_d}^{1/(d+4)}$$

3 Discussion

For this project, two data sets (evaluation sets) were provided to be classified into different classes (or densities). For each data set, a training set with known clssifications and a test set were provided. The purpose of the test set and training set was to train the specific classification algorithm that is to be used on evaluation sets. No information regarding to the context of the data were available.

In this paper I have used kernel density estimation technique to estimate the density or class for each evaluation set vectors, depending on the information obtained from training and test set vectors (whose classifications were known).

The application involved for each of the ith evaluation set vectors, the calculation of maximum likelihood value corresponding to each of k classes of the training set. Then all these maximum likelihood values were compared. The class that produced the highest ML value, has been identified as the class for the ith evaluation set vector. This process continued untill all the vectors of evaluation set were classified.

In this paper the multivariate normal kernel function is used. So the form of the estimated density function is,

$$\hat{f}(x) = \frac{1}{nh^d(2\pi)^{-d/2}} exp \sum_{i=1}^{n}\sum_{j=1}^{m} K\left(\frac{x-X_{ij}}{h}\right)^T \left(\frac{x-X_{ij}}{h}\right)$$

where x represent each vector of the evaluation set and X_{ij} is the jth training set vector for the ith class. And m is the number of vectors in ith class, n is the number of classes in training set. And ofcourse each vector is d (10) dimensional.

Group 1

The dimension of the data sets in this group is 10 with 11 classes or densities from where the data came. The training set has 528 vectors and the test set has 379 vectors. The bandwidth (h) used in this case is equal to .6737.

When the kernel density estimation algorithm has been used to identify the classes of test set

vectors errors have been found in 55% of the cases.

Group 2

The dimension of the data sets in this group is 39 with 5 classes or densities from where the data came. The training set has 925 vectors and the test set has 350 vectors.

When the kernel density estimation algorithm has been used to identify the classes of test set vectors, it identified only one class :class 1. error rate is approximately 80%.

4 Summary

Kernel density estimation technique has wide applications in the area of statistical data mining. The mathematical properties are well established for kernel densities. Though the results obtained here are not promising, still this method should be investigated further.

References

[1] Rao,Prakas.Nonpaarmetric Functional Estimation.Academic Press,London,1983.

[2] Silverman,B.W. Density Estimation for Statistics and Data Analysis. Chapman and Hall, London, 1986.

[3] Wegman,E.J. Nonparametric Probability Density estimation. Technometrics,14, pp.533-546,1972.