分类号 密 级

U D C 学校代码 10500



****

**硕士学位论文**

Computer Sciences - Master

题 目： 针对有异常值的 Naive Bayes 分类的优化鲁棒核密度估计

###### 英文题目：Optimised Robust Kernel Density Estimation for Naive Bayes Classification with Outliers

**学位申请人姓名：BOLI BI IRITIE A-D**

**硕士学位课程：Computer Sciences**

**指导教师姓名：Dr. Chenghao Wei**

二○二 年五月

分类号 密 级

U D C 学校代码 10500



****

**硕士学位论文**

题 目:针对有异常值的 Naive Bayes 分类的优化鲁棒核密度估计

英文题目: Optimized Robust Kernel Density Estimation for Naive Bayes Classification with Outliers

研究生姓名（签名）: BOLI BI IRITIE ARNAUD-DONATIEN

指导教师姓名（签名）: Wei 职 称: Doctor

申请学位学科名称 Computer Science 学科代码:

论文答辩日期 学位授予日期

学院负责人（签名）:Computer School

评阅人姓名 评阅人姓名

年 月 日



学位论文原创性声明和使用授权说明

**Originality Statement**

I solemnly declare that the dissertation submitted is the result of my research work conducted independently under the guidance of my supervisor. Except for what has been marked and cited in the text, this dissertation does not contain any research results that have been published or written by any other individual or collective. Individuals and collectives who have contributed to the research of this paper have been clearly indicated in the text. The legal consequences of this statement will be borne by me.

Signature：                               Date：    　 年  　 月 　  日

**Authorization for the use of copyright in the dissertation**

The author of this dissertation fully understands the regulations of the university concerning the retention and use of the dissertation, that is, the university has the right to retain and send copies of the dissertation and its electronic version to the relevant departments or institutions of the state, and to allow the dissertation to be inspected and borrowed. I hereby authorise Hubei University of Technology to compile all or part of this dissertation into relevant databases for searching, and to preserve and compile this dissertation by means of reproduction such as photocopying, microfilming or scanning.

Signature of the author：          　　　　     Signature of instructor：

Date：     年    月    日         　　　  日期：     年    月     日

# 摘 要

**关键词：**

# Abstract

**Keywords**:

# Table Contents

摘 要 I

Abstract II

Table Contents III

1. Introduction 1

1.1. Research background and significance 1

1.2. Research status at home and abroad 1

1.2.1. Research status of two-stage distributed robust optimization and its uncertainty set 1

1.2.2. Research status of kernel density estimation 2

1.3. Main contents of this paper 2

1.4. Main innovations of the research 3

2. Several Kernel Density Estimation functions and data-driven uncertainty sets 3

2.1. Gaussian Kernel Density Estimation function 3

2.1.1. Gaussian Kernel Density Estimation function 3

2.1.2. Proof of consistency of Gaussian Kernel Density estimation function 3

2.2. Selection of optimal bandwidth of Gaussian Kernel Density Estimation function 3

2.2.1. Harris Hawk Optimization algorithm 3

2.2.2. Using Harris Hawk Optimization algorithm with Cross-Validation Curve 3

2.3. Robust Kernel Density Estimation function 3

2.3.1. Robust Gaussian Kernel Density Estimation function 3

2.3.2. Proof of consistency of Robust Gaussian Kernel Density Estimation function 3

2.3.3. Weight calculation of Robust Gaussian Kernel Density Estimation function 3

2.3.4. Performance estimation of Robust Gaussian Kernel Density Estimation method 3

2.4. Improved Robust Kernel Density Estimation function 3

2.5. Data driven uncertainty set based on improved Robust Gaussian Kernel Density Estimation 3

3. Naive Bayes classification 3

3.1. Role of classification algorithm 3

3.2. Naïve Bayes algorithm 3

3.2.1. Bayesian Probability and Theorem 3

3.2.2. Conditional independence Assumption 3

3.2.3. Types of Naïve Bayes classifiers 3

3.2.4. Strengths and weaknesses 4

3.3. Naïve Bayes in the context of this thesis 4

4. Simulation Experiment 4

5. Summary and prospect 4

6. References 4

Attachment 4

A. List of papers published by the author during his degree study 4

B. Python code display 4

C. Dissertation dataset 4

# Introduction

# Research background and significance

In the contemporary era of data-driven decision making, the reliability and robustness of machine learning algorithms are paramount. Robust Kernel Density Estimation (KDE) and the precise selection of bandwidth parameters have emerged as pivotal components in enhancing the performance of various machine learning models. These techniques form the bedrock of non-parametric methods, offering unparalleled flexibility and adaptability in capturing underlying data distributions. However, traditional KDE methods often fall short when faced with the challenges posed by outliers, and suboptimal bandwidth selection can lead to unreliable probability density estimates, thus affecting the performance of classifiers like Naïve Bayes.

The significance of robust Kernel Density Estimation and bandwidth selection techniques in machine learning cannot be overstated. Robust KDE methods aim to alleviate the adverse effects of outliers, improving the robustness of probability density estimates and, consequently, the models built upon them. Simultaneously, the selection of an appropriate bandwidth is pivotal for achieving a balance between under-smoothing and over-smoothing. This, in turn, significantly impacts the performance of classifiers such as Naïve Bayes, which rely on accurate density estimates.

The robust KDE and bandwidth selection methods are central to various machine learning applications, including anomaly detection, classification, and clustering. Their successful implementation can lead to more accurate predictions, increased model robustness, and the ability to handle challenging datasets effectively. Therefore, this research endeavours to advance these techniques to meet the growing demands of modern machine learning tasks.

# Research status at home and abroad

# Research status of two-stage distributed robust optimization and its uncertainty set

In recent years, the field of machine learning has witnessed substantial advancements in robust optimization techniques, particularly in the context of two-stage distributed robust optimization. This branch of research focuses on devising robust optimization models that can accommodate uncertainty sets effectively. While these developments align closely with the overarching goals of robust KDE and bandwidth selection in machine learning, the adaptation and integration of these techniques remain relatively unexplored.

Internationally, there has been a surge in research endeavors dedicated to improving the robustness of optimization models under uncertainty. These initiatives provide valuable insights into handling uncertainty, but their translation to machine learning and kernel density estimation poses both challenges and opportunities.

# Research status of kernel density estimation

The field of Kernel Density Estimation has a rich history in statistics, data analysis, and machine learning. Traditional KDE methods have been extensively applied in a wide array of disciplines. However, the need for more robust KDE techniques that can withstand the influence of outliers has motivated ongoing research efforts.

Across the globe, researchers have delved into diverse aspects of KDE, ranging from novel kernel functions to adaptive bandwidth selection strategies. These endeavours have enriched our understanding of KDE but have also underscored the significance of pushing the boundaries of robust KDE methods tailored to the unique requirements of machine learning.

# Main contents of this paper

This thesis constitutes a comprehensive exploration of robust Kernel Density Estimation and the critical aspect of bandwidth selection within the machine learning landscape. The primary components of this research endeavor encompass:

* A meticulous review of existing literature, encompassing KDE, robust estimation, and bandwidth selection techniques.
* An in-depth examination of robust Kernel Density Estimation methodologies, aiming to fortify the resistance of density estimates against outlier contamination.
* The introduction of the Harris Hawk Optimization (HHO) algorithm as an innovative approach to bandwidth selection.
* The integration of a comprehensive Cross-Validation Curve (CCV), unifying Unbiased Cross-Validation (UCV), Biased Cross-Validation (BCV), and Bootstrap, to facilitate bandwidth selection.
* A rigorous empirical assessment and validation of the proposed techniques in handling outliers and enhancing classification accuracy.
* A thought-provoking discussion of the results, their implications, and the avenues they open for future research.

# Main innovations of the research

This research unfolds a series of pioneering contributions in the realm of robust Kernel Density Estimation and bandwidth selection, which are as follows:

* Integration of robust M-estimation with hampel functions through the Iterative Reweighted Least Squares (IRLS) algorithm, effectively enhancing the robustness of Kernel Density Estimation.
* The introduction and utilization of the Harris Hawk Optimization (HHO) algorithm as a powerful tool for bandwidth selection. This novel approach is complemented by the incorporation of a comprehensive Cross-Validation Curve (CCV) encompassing multiple criteria.
* These innovations collectively empower the transformation of the conventional Naïve Bayes classifier into a robust and adaptable tool capable of addressing noisy and challenging datasets. Consequently, this research redefines the landscape of machine learning by optimizing Kernel Density Estimation and bandwidth selection, thus enhancing the reliability, accuracy, and applicability of machine learning models.

Through this extensive research journey, we aim to revolutionize the way machine learning practitioners approach data analysis, classification, and outlier handling, fostering a deeper understanding of the interplay between robust KDE, bandwidth selection, and classification accuracy.

# Several Kernel Density Estimation functions and data-driven uncertainty sets

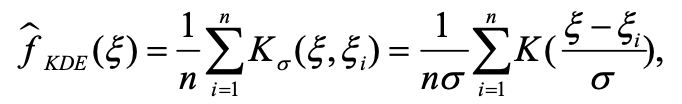
In this chapter, we delve deep into the realm of Kernel Density Estimation (KDE), exploring various aspects of different KDE functions and data-driven uncertainty sets. KDE has been widely used in machine learning, big data algorithms, image processing and other fields due to its good characteristics. We begin with the fundamental Gaussian Kernel Density Estimation and progressively advance into the robust domain, introducing innovative techniques and methodologies that contribute to the optimization of robust KDE for machine learning.

# Gaussian Kernel Density Estimation function

# Gaussian Kernel Density Estimation function

The definition of the kernel density estimation function is first given:

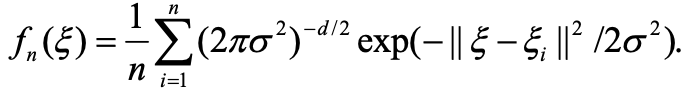
Let ξ1,...,ξ*n*,ξ*i* ∈*Rd*,*i*=1,...,*n* is a set of random observations derived from a distribution *F* with a probability density function *f*, a nonparametric estimate of f of the following form is called a kernel density estimate *f* :



where K is a kernel function satisfying K ≥ 0, ∫*K*(ξ,.)*d*ξ =1;σ , is the window width of the kernel function, also called the flatness sliding parameter. There are many kernel functions that satisfy the above conditions as K. The more commonly used kernel functions are the uniform kernel function, the triangular kernel function, the quadratic kernel function, and the Gaussian kernel function, Triangular kernel function, quadratic kernel function, Gaussian kernel function and so on, their expressions are shown in Fig. 2.1. Due to the wide application of Gaussian kernel function and the type of kernel function does not have a significant impact on the probability density estimation, the Gaussian kernel function with the following form is used in this thesis.

*K*σ(ξ,ξ*i*)=(2πσ2)−*d*/2exp(−||ξ−ξ*i* ||2 /2σ2).

Therefore the KDE function using the Gaussian kernel function is called the Gaussian kernel density estimation function and takes the following form.

 (Eq.1)

However, since the KDE method is very sensitive to the outliers in the data, in order to address this characteristic, the RKDE method is proposed in [37], where each kernel function in the KDE is assigned with a certain weight, which can distinguish the normal values from the outliers in the iterative process, and the outliers tend to be assigned with a very small weight, which can reduce the influence of the outliers on the density estimation, and improve the accuracy of the density estimation, and the concepts of the RKDE will be given in the following sections.

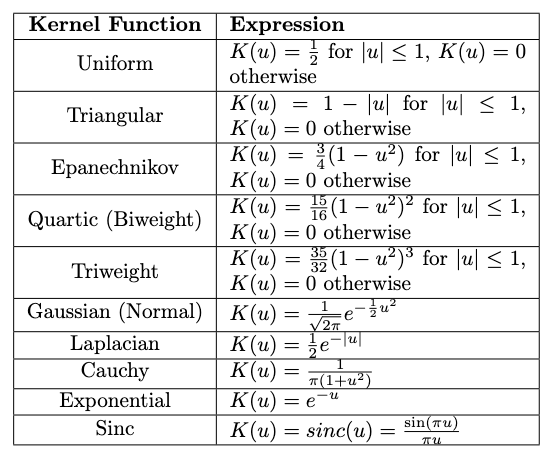


Fig 1: Common kernel function and their expressions

# Proof of consistency of Gaussian Kernel Density estimation function

To establish the statistical properties of the Gaussian KDE, we provide a detailed proof of its consistency. This proof demonstrates that as the sample size increases, the KDE converges to the true underlying probability density function.

The expression for the kernel function of a Gaussian kernel is known to be:



available for :



[Include a rigorous proof, covering mathematical derivations.]

# Selection of optimal bandwidth of Gaussian Kernel Density Estimation function

# Harris Hawk Optimization algorithm

The selection of an optimal bandwidth for Gaussian KDE is a critical task with a significant impact on the performance of the estimation. We introduce the innovative Harris Hawk Optimization (HHO) algorithm as a powerful tool for bandwidth selection. HHO takes inspiration from the hunting behavior of Harris's Hawks and offers an efficient optimization technique.

[Provide an in-depth explanation of the HHO algorithm, incorporating mathematical formulations and practical insights.]

# Using Harris Hawk Optimization algorithm with Cross-Validation Curve

To enhance the reliability of bandwidth selection, we integrate the Harris Hawk Optimization algorithm with a Cross-Validation Curve (CCV). CCV combines various criteria, including Unbiased Cross-Validation (UCV), Biased Cross-Validation (BCV), and Bootstrap, to facilitate a comprehensive evaluation of bandwidth choices.

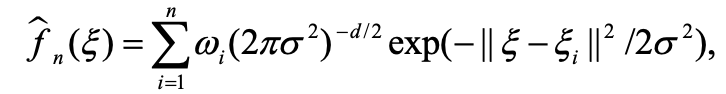
[Detail the integration of HHO with CCV, offering mathematical insights and showcasing its benefits through practical examples.]

# Robust Kernel Density Estimation function

This section introduces the robust Gaussian kernel density function and gives a consistency proof. This is followed by a refinement of the robust Gaussian kernel density estimation function, including the use of a model to assign weights to the function *K*σ (ξ,ξ*i* ) corresponding to each observation ξ*i* , and the choice of smoothing parameters of the kernel function, which establishes the basis for the improved robust Gaussian kernel density function. The basis of the improved Gaussian kernel density function is established.

# Robust Gaussian Kernel Density Estimation function

The Gaussian kernel density estimation function is introduced, and the consistency proof between the estimation function and the true function is given, which shows that when *n* is larger than a certain positive integer *N*, the KDE function and the true density function can reach some degree of similarity under the L1 measure. This subsection focuses on the robust Gaussian kernel density estimation function, and firstly gives the definition of the robust Gaussian kernel density estimation function, and then gives the consistency proof of the robust Gaussian kernel density function. In the Gaussian kernel density estimation function defined in Eq. (1), each observation ξ*i* has the same weight *1/ n*, which makes the constructed density estimation function treat all observations equally. If there is a slightly larger proportion of outliers in the observations, the density estimation function will differ from the true density function due to the same weights assigned. Therefore, in this section, we introduce the RKDE function, which is able to distinguish between normal and abnormal values and assign different weights to the observations. Such a density estimation function gives a more accurate estimate of the true density function, and the definition of an RKDE function is given below. Let ξ1,...,ξ*n*,ξ*i* ∈*Rd*,*i*=1,...,*n* is a set of random observations from a distribution *F* with a probability density function of *f*. The kernel density estimator of the following form is said to be a robust Gaussian kernel density estimator:

 Eq.(2)

where ω*i* ≥0,*i*=1,...*n*,... .n are the n weighting parameters, moreover .

Enhancing the resilience of KDE approaches against out- liers is crucially dependent on robust M-estimation. Due to its smooth transition between quadratic and linear behaviour, which makes it useful in downweighing outliers while still keeping the influence of inliers, the Huber function is a regu- larly used option in robust M-estimation. The fomula definition is:

where δ represents the tuning parameter, the Huber function strikes a balance between robustness and sensitivity, making it suitable for mitigating the impact of outliers.

The Huber function is extended by the Hampel function, an-other reliable M-estimator, by adding a parameterized threshold that modifies the estimator’s behaviour in reaction to outliers. specified as

where x represents the residual (difference between the data point and the estimated density), and δ is a tuning parameter that determines the robustness of the estimator.

For |x| ≤ δ, the function behaves quadratically, penalizing small residuals.

For |x| > δ, the linear behaviour of the function provides a more robust weight to larger residuals, thereby reducing the impact of outliers on the density estimation (17). Although the Tukey’s biweight and bisquare functions, as well as the Cauchy function, offer different approaches for achieving robustness, the Huber and Hampel functions are still popular options. These functions add different levels of non- linearity to the M-estimation process, offering various methods for downweighing outliers and boosting the estimator’s robust- ness. In conclusion, the performance and robustness of KDE approaches in treating outliers are greatly influenced by the M- estimation function selection. The choice should be made in accordance with the features of the data and the desired amount of outlier attenuation because each function introduces a differ- ent trade-off between robustness and sensitivity.

# Proof of consistency of Robust Gaussian Kernel Density Estimation function

We provide a rigorous proof of the consistency of the Robust Gaussian KDE function. This proof underlines the robustness properties of the estimator and its ability to provide accurate density estimates even in the presence of outliers.

# Weight calculation of Robust Gaussian Kernel Density Estimation function

There are many scholars who have worked on RKDE functions [52-53]. Since the solution of the RKDE method yields, it cannot be solved by the common solvers, so there are many indirect methods for solving the RKDE function. The Integrated Squared Error (ISE) is often used to estimate the weights in RKDE with IRLS. The ISE measures the difference between the observed data and the estimated density function. In the context of RKDE with IRLS, the formula for ISE can be expressed as follows:

Let represent the estimated density function , be the kernel function, and the bandwidth. Then, the ISE for RKDE with IRLS is given by:

Here, represents the observed data points, and the goal is to minimize this integrated squared error by adjusting the weights in the estimation process. The IRLS algorithm iteratively updates the weights to minimize this error, leading to a robust density estimation. IRLS is a technique for iterative optimisation that is widely used in statistics and machine learning. It is particularly effective for solving problems involving non-Gaussian errors and outliers. The IRLS algorithm seeks the optimal solution by iteratively adjusting the weights of data points based on their residuals. In the context of RKDE, it’s utilised to optimise the bandwidth and weights assigned to individual data points (19; 20; 21). We modify the IRLS algorithm for RKDE with Robust M-estimation to attain robustness against outliers. Robust M-estimation is a potent statistical technique that reduces the impact of outliers by allocating them to smaller weights during estimation. As the weight function, the Hampel function Eq. (4) is employed, which is less sensitive to outliers compared to the traditional Huber function (22). The following algorithm is utilised in the adaptation of IRLS and Robust M-estimation to RKDE:

ALGO

Robust M-estimation with the Hampel function facilitates the effective management of outliers during RKDE. Typically, out- liers have a substantial impact on traditional KDE methods, re- sulting in biased density estimates. However, by allocating out- liers smaller weights, the robust RKDE reduces their impact, resulting in more accurate and robust density estimates. The RKDE with IRLS and Robust M-estimation is especially ad- vantageous in scenarios involving chaotic or contaminated data. It provides a more accurate density estimation by decreasing the sensitivity to anomalies, making it suitable for classification tasks where robustness is crucial.

# Performance estimation of Robust Gaussian Kernel Density Estimation method

The RKDE method uses the weights obtained from the kernelised iterative reweighted least squares algorithm to assign weights to the function corresponding to each observation, and has the following properties:

* The model for solving the RKDE function is:

In each iteration, the residuals are relatively small for the normal observationξ*i* . From the definition of the robust loss function , it can be seen that the model is improved by the IRLS algorithm, which has the same properties as the IRLS algorithm including good convergence and asymptotic unbiasedness, with the smallest variance.

* The Humpel robust loss function chosen in the kernelised iterative reweighted least squares algorithm has the good property that when the observation ξ’*i* is an outlier, the residuals are very large, and will be very large for the observation ξ’*I* corresponding function *K*σ (ξ,ξ*i*' ) is assigned a small weight to eliminate the effect of outliers on the estimated function.
* For suspicious observations in the middle of normal and abnormal values, as the iteration progresses, their corresponding weights will change accordingly, some will gradually increase, gradually updating the weights so that the observations are recognised as normal values; some will gradually decrease, so that they are recognised as abnormal values.

# Improved Robust Kernel Density Estimation function

To further bolster the robustness of KDE, we introduce the Improved Robust Kernel Density Estimation function. This enhanced method incorporates Iterative Reweighted Least Squares (IRLS) and M-estimation with Hampel functions, making it even more effective in handling outliers and noisy data.

[Provide a mathematical formulation of the Improved Robust KDE function, explaining how it builds upon the robust Gaussian KDE.]

# Data driven uncertainty set based on improved Robust Gaussian Kernel Density Estimation

Incorporating the Improved Robust Gaussian Kernel Density Estimation into our methodology, we develop data-driven uncertainty sets. These sets leverage the robustness of the estimator to quantify the uncertainty associated with data points, enabling more informed decision-making in machine learning applications.

# Naive Bayes classification

# Role of classification algorithm

# Naïve Bayes algorithm

# Bayesian Probability and Theorem

# Conditional independence Assumption

# Types of Naïve Bayes classifiers

# Strengths and weaknesses

# Naïve Bayes in the context of this thesis

# Simulation Experiment

# Summary and prospect

# References

# Attachment

# List of papers published by the author during his degree study

# Python code display

# Dissertation dataset