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. Background and motivation 1

1.2. Research Objectives 1

1.3. Thesis Structure 1

2. Literature Review 1

2.1. Machine Learning and Intelligence Data Analysis 1

2.2. Kernel Density Estimation (KDE) 1

2.2.1. Gaussian KDE and its Applications 1

2.2.2. Limitations of Traditional KDE 1

2.3. Robust Kernel Density Estimation (RKDE) 1

2.3.1. M-estimation and Hampel Funtion 1

2.3.2. Iterative Reweighted Least Squares (IRLS) 2

2.4. Naïve Bayes Classifier 2

2.4.1. Baisc Naïve Bayes Classifier 2

2.4.2. Challenges with Noisy and Tough Datasets 2

3. Methodology 2

3.1. Robust RKDE with IRLS 2

3.2. Harris Hawk Optimization (HHO) for bandwidth selection 2

3.3. Complete Cross-Validation Curve (CCV) 2

3.3.1. Unbiased Cross-Validation (UCV) 2

3.3.2. Biased Cross-Validation (BCV) 2

3.3.3. Bootstrap 2

3.4. Integration of Robust KDE and Bandwith selection into Naïve Bayes classifier 2

4. Experimental Setup 2

4.1. Dataset Description 2

4.2. Implementation Details 2

4.3. Performance metrics 2

5. Results and Discussion 3

5.1. Evaluation of Optimized Robust KDE 3

5.2. Impact of Bandwidth selection on classification performance 3

5.3. Comparison with traditional Naïve Bayes Classifier 3

6. Conclusion and Future Work 3

6.1. Summary of findings 3

6.2. Contributions to the field 3

6.3. Limitation and Future direction 3

7. References 3

Appendices 3

A. List of papers published by the author during his degree study 3

B. Python code display 3

# Introduction

# Background and motivation

In recent years, there have been significant developments in the field of machine learning, particularly in the area of intelligence data analysis. Machine learning algorithms have been extensively utilised across many sectors, such as healthcare and finance, due to their ability to process substantial datasets and leverage robust computational resources. These algorithms have proven effective in addressing intricate challenges that were previously deemed intractable. The widespread adoption of machine learning has fundamentally transformed our capacities in the domains of decision-making, automation, and predictive analytics. Nevertheless, the increasing dependence on machine learning methods has led to a heightened awareness of their susceptibilities, particularly in the presence of noisy and complex datasets. Real-world data frequently contains various defects, such as outliers, measurement mistakes, and instances of missing numbers. When left unattended, these flaws can have a substantial impact on the dependability and efficiency of machine learning models, resulting in less than optimal results.

Accurately understanding and modelling data distributions is a fundamental requirement for numerous machine learning techniques. Kernel Density Estimation (KDE) is a crucial technique in this field, providing the capability to represent the underlying data distributions. The KDE method plays a significant role in a wide range of applications, encompassing anomaly detection, classification, and clustering. The effectiveness of machine learning models is heavily reliant on precise density estimate, which forms the basis for essential activities such as categorization, decision-making, and risk evaluation. Nevertheless, conventional KDE techniques frequently encounter difficulties when faced with datasets that include outliers and noisy observations. Aberrant data points, sometimes referred to as outliers, have the potential to exert a disproportionate influence on the process of density estimation. Consequently, this might result in the production of skewed models and inferior outcomes in subsequent tasks. The task of addressing this difficulty and optimising the Kernel Density Estimation (KDE) technique holds significant significance in improving the resilience and dependability of machine learning models.

Outliers can be defined as data points that exhibit substantial deviation from the remaining data. Measurement errors, data corruption, or the occurrence of extraordinary occurrences might give rise to these issues. In numerous practical situations, outliers possess significant informational value and should not be disregarded, but rather be appropriately incorporated into the modelling process. Nevertheless, it is important to acknowledge that these techniques have the potential to alter the fundamental distribution of data, resulting in biassed estimations of density and, consequently, unreliable predictions from the model.

# Research Objectives

The central aim of this thesis is to contribute to the advancement of robust Kernel Density Estimation and its utilisation in the study of intelligence data. Our study aims to accomplish the following specific objectives:

To enhance or maximise the efficiency or effectiveness of a particular process or system. This study aims to develop and enhance robust kernel density estimation (KDE) approaches that can effectively estimate data densities in scenarios where outliers and noisy data points are present. The term "robust" refers to approaches that exhibit reduced vulnerability to the impact of outliers and offer more dependable density estimates.

Maximising the Efficiency of Bandwidth Selection: The utilisation of the Harris Hawk Optimisation (HHO) algorithm is proposed in order to optimise the selection of bandwidth parameters, which play a crucial role in Kernel Density Estimation (KDE). This optimisation aims to boost the accuracy and adaptability of KDE when applied to various datasets. The selection of bandwidth is a crucial and important task in the context of KDE (Kernel Density Estimation), and its optimisation has the potential to enhance the accuracy and precision of density estimates.

This study aims to explore and apply techniques to minimise the influence of outliers in classification prediction. Specifically, the focus will be on investigating strategies to mitigate the negative consequences of outliers in KDE-based models, with a particular emphasis on the Naïve Bayes classifier. Classification tasks are often observed in the field of machine learning, and the capacity to generate reliable predictions even when outliers are present holds substantial practical significance.

# Thesis Structure

This thesis is structured as follows:

**Chapter 2**: The research Review offers a comprehensive examination of the current body of research pertaining to machine learning, Kernel Density Estimation, robust statistics, and the difficulties presented by datasets including noise. We utilize influential literature and contemporary scholarly investigations to ascertain the present understanding in these fields.

**Chapter 3:** The methodology section provides a detailed explanation of the methodologies utilized in this study, encompassing the enhancement of robust KDE techniques, the utilization of the Harris Hawk Optimization algorithm, and the incorporation of these approaches into the Naïve Bayes classifier. In this thesis, we outline the fundamental technical principles that underpin our research.

**Chapter 4**: The Experimental Setup section provides a comprehensive overview of the datasets employed, the technical implementation details, and the performance criteria selected to assess the efficacy of the proposed approaches. The implementation of rigorous experimentation is of utmost importance in order to evaluate the efficacy of our methodologies.

**Chapter 5:** The section titled "Results and Discussion" summarizes the outcomes of the conducted tests, analyses their consequences, and evaluates the influence of bandwidth selection on the performance of categorization. By employing both quantitative analysis and qualitative talks, we thoroughly examine the practical consequences of our research.

**Chapter 6:** In conclusion, this study has provided valuable insights into the topic at hand. Moving forward, there are several avenues for further research that could further enhance our understanding of the subject The present study provides a comprehensive overview of the research's contributions, delineates its inherent limits, and proposes potential directions for further investigation within the domain of robust kernel density estimation (KDE) for machine learning. The present discourse entails a contemplation of the ramifications of our research and its capacity to propel progress in the field of intelligence data analysis.

**Chapter 7:** The references section encompasses an extensive compilation of sources, including influential works that have provided guidance for the present study. We express our gratitude for the valuable contributions made by the wider scientific community in influencing and refining our comprehension of the subject matter.

This thesis aims to make a contribution to the improvement of robust KDE approaches, with the ultimate goal of enhancing the capabilities of machine learning models in effectively managing tough and noisy datasets. The subsequent chapters will explore the complexities of these approaches and their empirical verification, facilitating a more profound comprehension of intelligent data analysis within the realm of machine learning.

# Literature Review

The literature review within this thesis offers a comprehensive examination of the fundamental principles and recent advancements in the domains of machine learning and intelligent data analysis. This chapter establishes the foundation by analysing the fundamental ideas and difficulties linked to these domains, thereby preparing for an in-depth investigation of KDE and its resilient variations.

# Machine Learning and Intelligence Data Analysis

The discipline of machine learning, which falls under the umbrella of artificial intelligence (AI), has experienced significant expansion and broadening in recent decades. Machine learning involves a diverse range of methodologies that facilitate the ability of computers to acquire knowledge from data, identify patterns, and make informed judgements without the need for explicit programming. The progression of machine learning can be succinctly encapsulated as follows:

**Early Developments**: The origins of machine learning may be historically attributed to the mid-20th century, during which influential figures such as Alan Turing and Arthur Samuel established the fundamental principles of rule-based learning and game-playing algorithms. These first advancements laid the foundation for subsequent investigations.

**Statistical Learning**: The field of machine learning has undergone a progressive transformation, transitioning into a statistical discipline that extensively incorporates principles from probability theory and statistical methodologies. Methods such as linear regression and decision trees have gained significant popularity in the field of pattern detection and prediction.

**The Era of Neural Networks**: During the 1980s and 1990s, artificial neural networks emerged as a prominent paradigm within the field of machine learning. Nevertheless, the potential of their capabilities was hindered by the constraints imposed by low computer resources and an inadequate amount of data.

**The Big Data Revolution**: The 21st century has saw a significant transformation in the field of machine learning, chiefly influenced by the abundant availability of extensive datasets and notable progress in computational capabilities. Consequently, the comeback of neural networks in the form of deep learning has emerged as a prevailing force in several domains, encompassing image identification, natural language processing, and autonomous cars.

**Interdisciplinary Impact:** Machine learning has expanded beyond its conventional limitations and has been implemented in other fields, such as healthcare (for purposes of diagnosis and medication development), finance (including algorithmic trading and fraud detection), and natural language processing (including chatbots and language translation).

The study of intelligence data, as it pertains to the field of machine learning, is of significant significance in the extraction of actionable insights, the facilitation of informed decision-making, and the automation of intricate procedures in many industries. The following are few fundamental elements that contribute to its significance:

**Decision Support:** The study of intelligence data offers decision-makers with significant insights derived from extensive databases. Data-driven insights play a pivotal role in enabling organisations across several sectors, including healthcare, finance, and marketing, to make informed decisions that optimise operational efficiency and enhance overall performance.

**Predictive Modeling:** Machine learning methodologies facilitate the creation of prognostic models that anticipate forthcoming occurrences or patterns. This holds significant value in various circumstances, including weather forecasting, stock market prediction, and disease outbreak monitoring.

**Anomaly Detection:** The identification of abnormalities or outliers in data is a critical undertaking that finds utility in various domains such as fraud detection, network security, and quality control. Machine learning algorithms have exceptional proficiency in identifying patterns that differ from the established norm.

**Pattern Recognition:** Machine learning models possess a high level of proficiency in identifying intricate patterns inside datasets. This capability is utilised in the fields of image identification, speech recognition, and natural language comprehension, leading to advancements such as autonomous vehicles and virtual personal assistants.

**Personalization:** The utilisation of intelligence data analysis plays a crucial role in driving recommendation systems, which are designed to offer personalised content and product choices. These systems have the ability to improve user experiences in various domains such as e-commerce, streaming services, and online advertising.

The rising reliance on data-driven decision-making across several industries underscores the growing importance of robust and accurate data analysis tools. The approach of Kernel Density Estimation holds significant importance in the modelling of data distributions, therefore making it a central focus of inquiry in this thesis.

The subsequent sections of this chapter will provide a more comprehensive consideration of the ideas and issues related to KDE, specifically within the framework of robust data analysis.

# Kernel Density Estimation (KDE)

The KDE technique is a widely employed statistical method utilized to estimate the probability density function (PDF) of a continuous random variable. This technique assumes a crucial position in diverse applications, encompassing data smoothing, density estimation, and outlier detection. This section will delve into the Gaussian KDE, which is considered one of the most extensively utilized variations, along with its various applications. Furthermore, we will examine the constraints associated with conventional KDE methodologies.

# Gaussian KDE and its Applications

The Gaussian KDE is a non-parametric method that aims to estimate the probability density function (PDF) of a given dataset. This methodology involves applying a Gaussian (normal) distribution, referred to as a kernel, on each individual data point and aggregating them to form the overall density estimate. The Gaussian kernel function, denoted as , is expressed as:

Where:

* is the kernel function.
* represents the distance from the data point of interest.

The Gaussian KDE estimate for a given data point with a dataset is calculated as follows:

Where:

* stands for the KDE estimate of the PDF at point .
* is the total number of data points in the dataset.
* signifies individual data points within dataset .
* is the bandwidth parameter, responsible for controlling the level of smoothing and is a pivotal factor in KDE as it determines the width of the kernels.

Gaussian KDE has gained widespread application in various domains:

**Probability Density Estimation:** The Gaussian KDE is a highly significant tool in the field of statistical analysis and hypothesis testing due to its ability to give a non-parametric estimation of the PDF .

**Data Smoothing:** The aforementioned methodology is employed to mitigate the impact of noise on data, resulting in the reduction of high-frequency noise and the accentuation of fundamental patterns in temporal data.

**Data Visualization:** The Gaussian KDE method plays a crucial role in producing visually appealing density plots, also known as kernel density plots. These plots are widely employed for the purpose of visualising the distribution of data.

**Outlier Detection:** The Gaussian Kernel Density Estimation (KDE) technique assists in the detection of outliers in datasets by constructing a model of the underlying distribution of the data and identifying data points that exhibit substantial deviations from this model.

# Limitations of Traditional KDE

While Gaussian KDE is a powerful tool, it is not without its limitations:

**Bandwidth Selection:** The selection of bandwidth (ℎ) has a substantial influence on the outcome of the KDE. An excessively limited bandwidth yields an estimate that is prone to overfitting, whereas an excessively broad bandwidth results in an estimate that is excessively smoothed. The process of choosing a suitable bandwidth is a complex undertaking.

**Curse of Dimensionality:** The effectiveness of KDE diminishes as the dimensionality of the data rises. The drop in volume of the kernel exhibits a rapid decline as the dimensionality increases, leading to the emergence of sparsity in spaces with high dimensions.

**Computational Demands**: The computing demands of KDE, particularly in high-dimensional domains, might be substantial, hence restricting its practicality for big datasets.

**Edge Effects:** The estimation of KDE towards the limits of the data domain may exhibit bias due to the absence of data points beyond the boundary, which is necessary for the formation of valid kernels.

**Sensitivity to Data Density:** The performance of KDE can exhibit substantial variations based on the density and distribution of the data. Sparse or multimodal datasets may result in poor performance.

**Boundary Issues:** The presence of boundaries or discontinuities in the data can result in inaccurate estimates due to the potential for undersmoothing or over smoothing.

Given the aforementioned constraints, scholars have undertaken investigations into resilient adaptations of the KDE method and have developed strategies to optimize the selection of bandwidth. In the following portions of this thesis, we will explore robust Kernel Density Estimation techniques and novel strategies to address these limitations.

# Robust Kernel Density Estimation (RKDE)

The KDE technique offers a non-parametric approach to estimate probability density functions based on observed data. The conventional KDE method exhibits a strong susceptibility to outliers and may yield density estimates that are biased when extreme values are present. In order to tackle this issue, the use of RKDE has emerged as a helpful tool, providing resilience against outliers and noise present in the data. It can be expressed using the following formula:

Let be a dataset with n data points, and represents the kernel function. The RKDE estimate of the probability density function at point is given by:

where:

* is the bandwidth parameter, determining the width of the kernel window.
* signifies individual data points within dataset
* is the weight assigned to each data point based on robust M-estimation, such as the Hampel function.

# M-estimation and Hampel Function

M-estimation is a statistical technique that is frequently utilized in the field of robust statistics for the purpose of estimating parameters of a given model. The primary purpose of this study is to identify the optimal parameter values that maximize a designated objective function, commonly known as the "M-estimator." One of the primary obstacles encountered in the field of robust density estimation pertains to the existence of outliers, which possess the potential to exert a substantial impact on the precision of conventional KDE techniques.

In order to tackle this particular difficulty, we propose the utilization of the Hampel function, which is a robust weight function employed in M-estimation techniques. But we have a lot of objective function in M-estimation technique such as *Huber*, *tuTukey’s biweight* and *bisquare* functions. In downweighing outliers while still keeping the influence of inliers, the Huber function is a reguarly used option in robust M-estimation. The formula 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 and its capacity to diminish or "trim" extreme values, so diminishing the impact of outliers in the estimate procedure. As 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 different trade-off between robustness and sensitivity.

The integration of the Hampel function within the KDE framework, the RKDE is defined as:

where:

* is the bandwidth parameter, determining the width of the kernel window.
* signifies individual data points within dataset
* is the weight assigned to each data point based on robust M-estimation, such as the Hampel function.
* is the Hampel function, which assigns weights to data points based on their distance from .

The bandwidth parameter ℎ determines the width of the kernel and plays a crucial role in regulating the smoothness of the density estimate. It requires to be carefully selected to balance bias and variance in the estimate.

The RKDE combines the weighted contributions of all data points, taking into account their proximity to the evaluation point x. This is achieved by the use of a kernel function, which helps to smooth out the contributions. The robustness of RKDE lies in its ability to offer more precise density estimations when confronted with outliers or data distributions that deviate from the standard, in contrast to conventional KDE techniques.

It is important to note that different implementations of RKDE may employ diverse kernel functions and robustness methods, resulting in changes in the specific formulation of the approach.

# Iterative Reweighted Least Squares (IRLS)

The IRLS method is a commonly employed numerical optimization approach in the field of robust statistics, particularly in the context of robust density estimation. It iterative process involves refining parameter estimations by updating them according to the weighted least squares criterion. M-estimation issues, which involve the maximization of a probability function, are particularly well-suited for this method.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

In the setting of RKDE, IRLS is employed to optimize the parameters of the Hampel function. The process involves iteratively updating the tuning parameters of the Hampel function to minimize the influence of outliers on the density estimation. The algorithm proceeds as follows:

1. Initialize the turning parameters.
2. Compute the weights for each data point using the current parameters.
3. Update the parameters based on the weighted least squares criterion.
4. Repeat steps 2 and 3 until convergence is achieved.

The iterative nature of IRLS allows the method to adapt to the data distribution, effectively reducing the impact of outliers while preserving the integrity of the density estimation for non-outlying data points.

In summary, robust Kernel Density Estimation (RKDE) leverages M-estimation with the Hampel function and Iterative Reweighted Least Squares (IRLS) to mitigate the influence of outliers and heavy-tailed data distributions. The Hampel function, with its adaptability to extreme values, offers a compelling solution for achieving robust density estimation in the presence of challenging data characteristics. In the subsequent sections, we explore how RKDE, aided by these robust techniques, is integrated into the Naïve Bayes classifier to enhance its performance in intelligence data analysis.

# Naïve Bayes Classifier

The Naïve Bayes Classifier is a foundational and widely used probabilistic classification algorithm in the realm of machine learning. Its roots trace back to Bayesian probability theory, a powerful framework for dealing with uncertainty and making predictions based on available data.

# Baisc Naïve Bayes Classifier

At the core of Naïve Bayes Classifier lies Bayes' theorem, a fundamental concept in probability theory. It allows us to calculate the probability of an event, based on prior knowledge of conditions that might be related to the event. In the context of NBC, we're interested in using Bayes' theorem for classification, specifically for predicting the class label of an observation () given its features ().

Mathematically, Bayes' theorem can be expressed as:

* is the posterior probability of class given the features
* is the likelihood of observing features given class
* is the prior probability of class
* is the marginal likelihood of the features

The beauty of Naïve Bayes Classifier lies in its simplicity. It makes the "naïve" assumption that the features are conditionally independent given the class label .

In other words, it assumes that knowing the value of one feature provides no information about the value of another feature when the class label is known. This simplifies the likelihood term:

Here, represents individual features, and can be estimated from the training data. Common techniques for estimation include Maximum Likelihood Estimation (MLE) or kernel density estimation, which provides a non-parametric way to estimate probability densities.

# Challenges with Noisy and Tough Datasets

While Naïve Bayes Classifier has proven effective in a wide range of applications, it does have limitations, especially when applied to noisy or challenging datasets. One significant limitation is its sensitivity to outliers and deviations from the independence assumption. In practice, real-world datasets often contain noisy or correlated features, which can violate the naïve independence assumption. When this assumption doesn't hold, NBC's performance can suffer.

Our research seeks to address these challenges by enhancing the Naïve Bayes Classifier with Robust KDE techniques. By employing robust statistical methods like M-estimation with the Hampel function, we aim to improve the estimation of and ultimately enhance the robustness and accuracy of the Naïve Bayes Classifier.

In subsequent sections, we will delve into the methodology used to optimize RKDE, incorporating Harris Hawk Optimization (HHO) for bandwidth selection, and evaluating the impact of these enhancements on classification performance.

By the end of this thesis, we aim to provide a comprehensive understanding of how RKDE can be applied to Naïve Bayes Classifier, offering a more robust and reliable classification tool, particularly in scenarios where data quality and the independence assumption are challenged.

# Methodology

# Robust RKDE with IRLS

# Harris Hawk Optimization (HHO) for bandwidth selection

# Complete Cross-Validation Curve (CCV)

# Unbiased Cross-Validation (UCV)

# Biased Cross-Validation (BCV)

# Bootstrap

# Integration of Robust KDE and Bandwith selection into Naïve Bayes classifier

# Experimental Setup

# Dataset Description

# Implementation Details

# Performance metrics

# Results and Discussion

# Evaluation of Optimized Robust KDE

# Impact of Bandwidth selection on classification performance

# Comparison with traditional Naïve Bayes Classifier

# Conclusion and Future Work

# Summary of findings

# Contributions to the field

# Limitation and Future direction

# References

# Appendices

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

# Python code display