The Kernel Density Integral Transformation (TMLR, 2023)

Reviewer : Seokhun Park

Department of Statistics
Seoul National University

Contents

Outline

Introduction

- Traditional methods
  - Min-max scaling
    - Preserve the shape of each feature's distribution.
    - Not robust to the effect of outliers.
  - Quantile transformation
    - Not preserve the shape of each feature's distribution.
    - Robust to the effect of outliers.
- Proposed method (Kernel Density Integral Transformation)
  - Lies between min–max scaling and quantile transformation.

Proposed Method

- Consider univariate data $\{X_1, ..., X_N\}$.
- Gaussian Kernel density estimation:

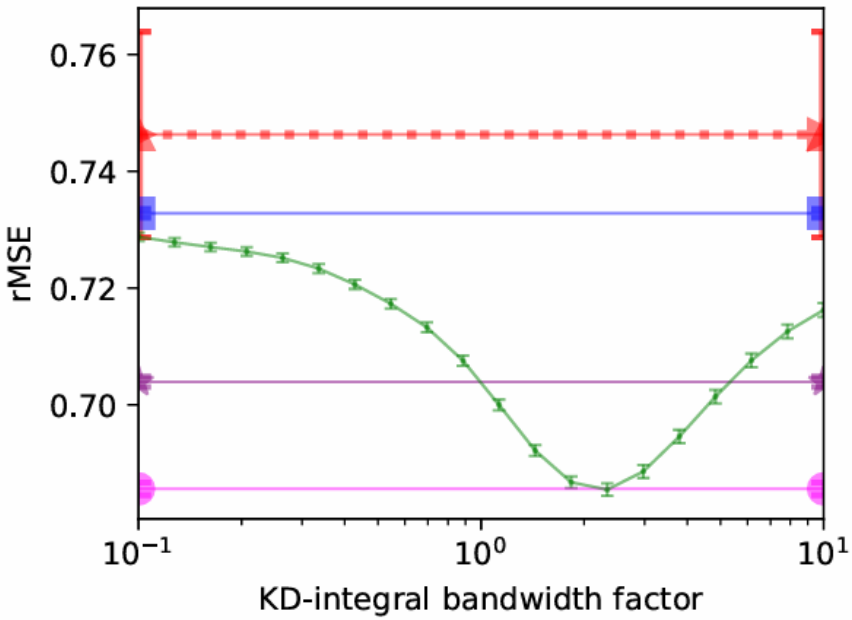
$$\hat{f}_h(x) = \frac{1}{Nh} \sum_{n=1}^{N} K\left(\frac{x - X_n}{h}\right), \tag{1}$$

  where $K(x) = \exp(\frac{-x^2}{2})/\sqrt{2\pi}$ and $h > 0$ is the bandwidth.
- Proposed method:
  1. Estimate density using the Gaussian Kernel density estimation.
  2. Preprocessing data via quantile transformation using a estimated density.
- $h \to \infty$: Proposed method $\approx$ min-max scaling
  $h \to 0$: Proposed method $\approx$ Quantile transformation.

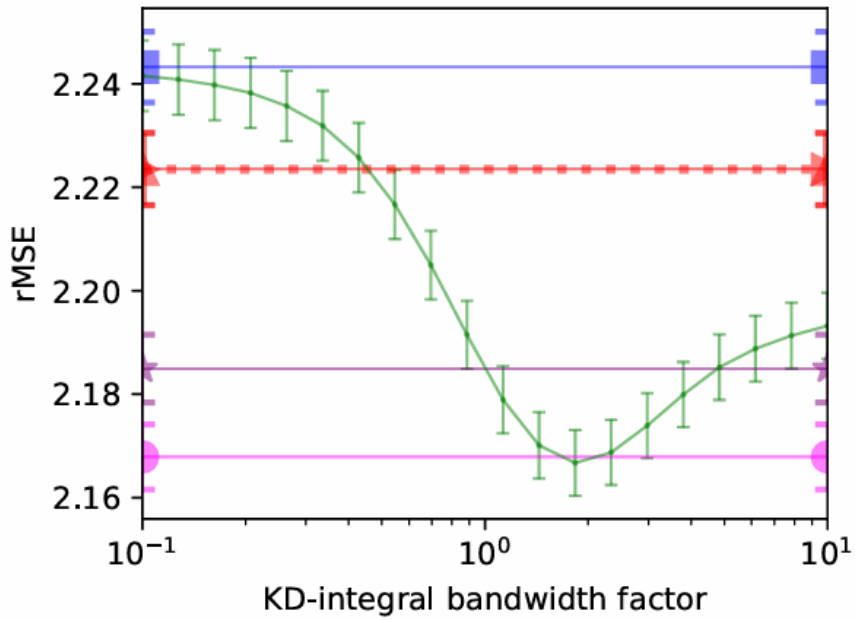
Outline

Experiments

(A) CA Housing ($N = 20640, D = 8$)



(B) Abalone ($N = 4177, D = 10$)



- Above figure presents the prediction performance of linear regression model based on data preprocessing methods.
- Red line : Min-max scaling, Blue line : quantile transformation
  Green line : proposed method depends on h

Experiments

Data preprocessing method	AUROC (std)
Min-max	0.864 (0.132)
Quantile	0.866 (0.131)
KDIT (h=1)	0.868 (0.129)
KIT (h=turned)	0.869 (0.129)

- Above Table presents the average of the AUROCs for Support Vector Classifier with various data preprocessing methods on various datasets.

Outline

Drawbacks of KDIT

1. It requires the entire data as a empirical quantile transformer.
2. Determining bandwidth parameter h is problematic and computational burden.
3. Since the experiments use only classical models, it is unclear whether the method would perform well when applied to models such as deep neural networks.