# Master Thesis Proposal: Robust Kernel Density Estimation for Naive Bayes Classifier

Boli Bi Iritié A-D

Sep 27, 2022

A classifier is a machine learning model used to distinguish between non-linear data objects according to specific attributes. Naturally, posterior probabilities can be used to achieved the aim, and Naive Bayes Classifier calculated them according to the Bayes Theorem. For handling the discrete attribute, the probability value simply obtains by its frequency value. Additionally, for tackling the calculation of the continuous attribute, an assumption of the density distribution model usually was made and then the parameters of the model are generated by using Maximum likelihood method. However, it is very difficult to decide the correct model assumption due to complexity of real-world dataset.

In this research, we try to use a modified Kernel density estimation for improving the performance of basic Naïve Bayes classifier. Kernel Density Estimation is a Nonparametric approach for calculating probability density using a dataset. The kernel effectively smooths or interpolates the probabilities across the range of possible outcomes for a random variable, ensuring that the total of probabilities equals one, as is required by well-behaved probabilities. The kernel function weights the contribution of observations from a data sample based on their relationship to a given sample. The window of observations, from the data sample that contributes to estimating the probability for a given sample is controlled by parameters called the **smoothing parameter** or **the bandwidth**. As a result, kernel density estimation is also known as a Parzen-Rosenblatt window. However, a big window may produce a coarse density with few features, whereas a tiny window may include too much detail and be too smooth or general to cover new or unseen cases adequately. It is important to choose an optimized method to obtain a suitable window width, and it will lead to accurate density estimation. In addition, existing outliers in the dataset are inevitable, which could affect the performance of density estimator and result in low accuracy of label prediction.

Our work tries to use a novel genetic algorithm for searching the optimized window width and we implement the **Robust Kernel Density Estimation** (RKDE) method to tackle the problem of outliers. The principle of RKDE is that the outliers can significantly affect the sample mean, which may reduce performance of the whole learner, so we need to reduce the weight of outliers until its reach 0. To fulfill our goal, we' ll use the Iteratively **Re-Weighted Least Square** (IRWLS) algorithm for improving the performance of the algorithm with optimized parameters