Binarization using Binary Trees in Machine Learning

03.27.2021

—

Yijie Guo, Yu Chen
yg2418@nyu.edu, yc4902@nyu.edu
CSCI-GA.2565 Machine Learning
New York University

Overview

Discretization and Binarization are common feature engineering techniques to transform continuous data to discrete or binary data, since real world data is noisy and may have highly skewed or non-standard distribution. Also, many machine learning algorithms only accept discrete values and many other machine learning algorithms work better with discrete values. However, due the disruptive transformation process, the transformed data could lose its characteristics and become meaningless. We propose a unique binarization process using binary trees, which could potentially capture the original data distribution as much as possible, while offering options allowing researchers to balance between the degree of smoothness and generalization and the representativeness of original distributions and characteristics.

Goals

1. Find the most common public machine learning datasets for our experiment
2. Apply our binary tree feature engineering to convert continuous data to binary data
3. Apply the K-Bins Discretizer as the baseline to compare with our method
4. Feed all three types of data into several machine learning and deep learning algorithms, such as random forest, gradient boosting, average perceptron, support vector machine and neural networks.
5. Develop and test hyper parameters for our method
6. Analyze the error, bias, variance and convergence rate of all types of data on all tested algorithms.
7. If possible, conclude with mathematical proofs for our methods as a regularizer

Specifications

Our project would use Python and Jupyter notebook as the development environment. Also, we would like to use the sklearn machine learning library and the pytorch deep learning library.