IST 707 HOMEWORK 8

Joy (Qiaoyi) Liu

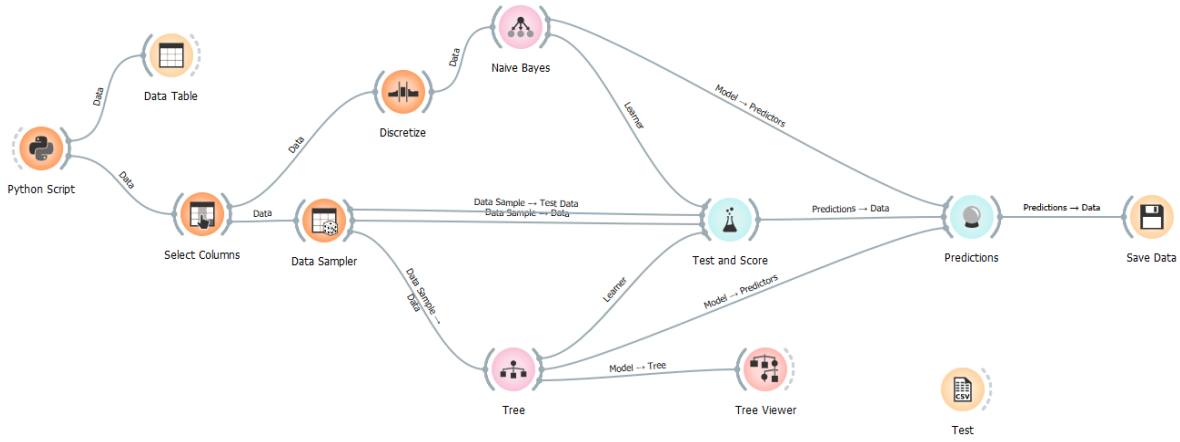
IST 707 Applied Machine Learning

Professor Joshua Introne

October 23rd, 2022

Introduction

The Modified National Institute of Standards and Technology (MNIST) dataset consists of 784 pixels/columns indicating each image with 28 pixels in height and 28 pixels in width. By showing an integer value between 0 to 255 for each pixel, it represents the lightness or darkness of that pixel to deduce the possible number of that image. With a train dataset of 42,000 rows and a label indicating the number, the Decision Tree and Naïve Bayes model could predict the possible number/label of the test dataset (28,000 rows).



Decision Tree

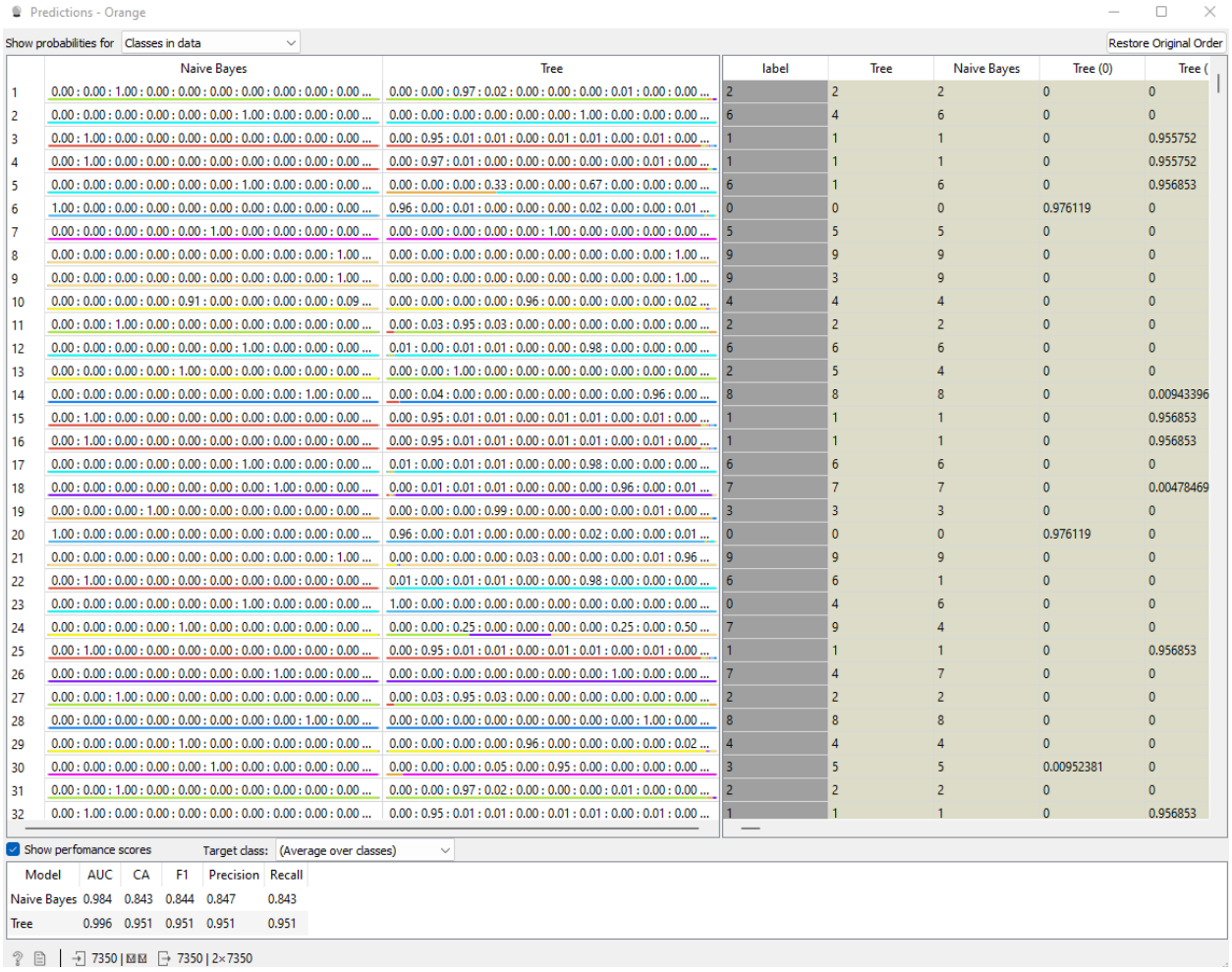
I first imported the train dataset into Orange via the Python Script widget. Then I selected the label column as the target for the predictions afterward. I compared the performance of the decision tree by tuning the sampling proportion from 70% to 90% in 3-fold validation. The predictions from the decision tree did not vary (AUC = .995, CA = .948, F1= .948, Precision = .948, Recall = .948).

Naïve Bayes

I added the discretize widget before Naïve Bayes and changed the setting from Equal-frequency discretization to Equal-width discretization. There was only a slight difference in its

prediction accuracy. The Equal-frequency discretization has F1 = .841 and the Equal-width discretization has F1 = .842.

Algorithm Performance Comparison



In general, the Decision Tree performs better than Naïve Bayes. The decision tree has a higher accuracy in predicting the labels and visualizing the results clearly. However, it has disadvantages in overfitting. The Decision Tree has the F-measure of .948, which is highly possible the model is overfitting to the dataset. However, to be certain whether the accuracy indicates overfitting would require a validation curve. I only found the solution to do so in Python, not in

Orange. Naïve Bayes is less likely to overfit due to the fact they “ignore” irrelevant features. Therefore, they have less accuracy in prediction. I personally did not recognize the difference in speed in generating the Decision Tree and Naïve Bayes. So I looked up and found that it is commonly accepted the Naïve Bayes is computationally faster in making decisions because it simplifies the computation by assuming there are no dependencies amongst attributes (Ashari et al., 2013).

Reference

Ashari, A., Paryudi, I., & Tjoa, A. M. (2013). Performance comparison between Naïve Bayes, decision tree and k-nearest neighbor in searching alternative design in an energy simulation tool. *International Journal of Advanced Computer Science and Applications (IJACSA)*, *4*(11).