Coding AI HW 4

Stanley Oduor UNI: soo2117

November 2024

1 Part 1: Classifiers

a.) **Nearest neighbor**
Best training score: 0.9233333333333332
Testing score: 0.93
The decision boundary is very irregular and the regions are classified based on the closeness of the training points to each other. There is a high variance and the local patterns within the individual regions are captured well. The classifier has a high performance score of 0.93 which indicates the classification was effective as depicted.
b.) **Logistic regression**
Best training score: 0.6166666666666667
Testing score: 0.545
The classifier has a lower performance for the testing score due to the linear decision boundary and the classifier struggles with the decision boundary.
c.) **Decision tree** *Best training score: 0.97*
Testing score: 0.98
The decision tree created a box-like decision boundary. The decision tree classifier has a higher training score because it fits the data perfectly. The classifier also has high performance score meaning it generalized well for the dataset, with clear distinctions between the boundaries.
d.) **Random Tree**
Best training score: 0.9199999999999999
Testing score: 0.91
The random tree aggregates more decision trees, thus resulting to a more generalized boundary with less boxes compared to the decision tree classifier above. This reduces the chances of overfitting while maintaining the high performance of decision trees in both training and testing scores.
e.) **Adaboost**
Best training score: 0.6599999999999999
Testing score: 0.585
The adaboost classifier combines multiple weak learners to create a strong classifier. The decision boundary shows the specific concentrated points in the regions but doesn't show other patterns in the dataset. The relatively lower training score and testing scores depict that the model is struggling with generalization

for the dataset.
Images are below

2 Part 2 - High Dimensionality

High Dimensionality - Curse or Blessing?

While high dimensionality has mainly been considered a curse, especially due to the exponentially increased computation power required, I'm more attracted to the evolving view of it as a blessing. High dimensionality, despite the high computational power, exhibit properties that make it advantageous and makes the computation a bit more worth it. This becomes evident in several ways, two of which I aim to explore.

First, high dimension datasets tend to exhibit fairly simple geometric properties. While complementing the curse of dimensionality, its crucial to note that, if a dataset is highly dimensional, then suprisingly, some problems get easier and can be solved by simple and robust old methods(Gorban, 1). One such example of this cases is the handling of noise and random fluctuations in the data. High dimensional data is advantageous in this case because it increases the sample space thus reducing the likelihood of overfitting, which affects the performance of the model. The reduces overfitting and the increased sample space improve the performance and accuracy, and helps in generation of the simpler geometric shapes that enhance analysis.

Second, high dimensionality, enhances the concentration of measure where the high dimensional data tens to cluster around certain regions(as evident in the picture), thus being very helpful in classification tasks such as when using the KNN classifier. Moreover, this feature contributes to the blessing of high dimensionality as mistakes can be grouped in clusters allowing the creation of correctors for handling specific cases instead of specific mistakes which enhances the entire process(Gorban 3). This is further evidenced in the Lipschitz function for a high- dimensional sphere with uniform measure which is almost constant and seems to look like the tail of the normal distribution-'tails behave at worst like a scalar gaussian random variable with absolutely controlled mean and variance'(Donoho 18).

From the above, its evident the blessing of high dimensionality is advantageous and enhances interpretation and presentation as displayed in the picture from the code.

References.

1. High-Dimensional Brain in a High-Dimensional World: Blessing of Dimensionality
2. High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality

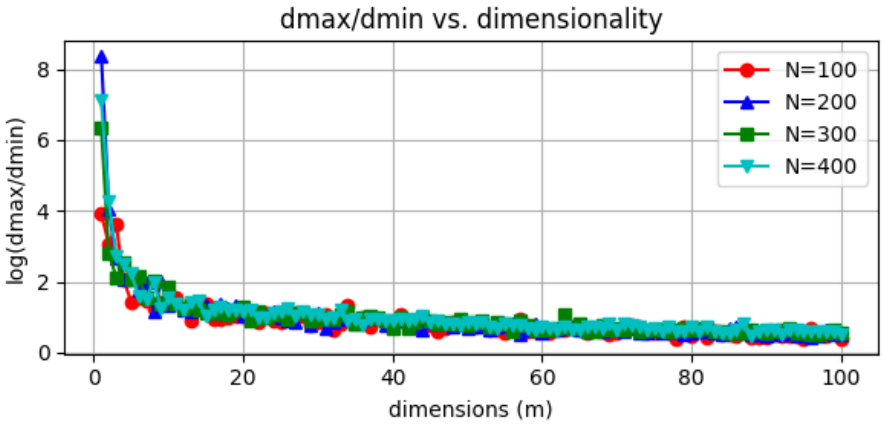

Figure 1: High dimensionality, curse or blessing plot

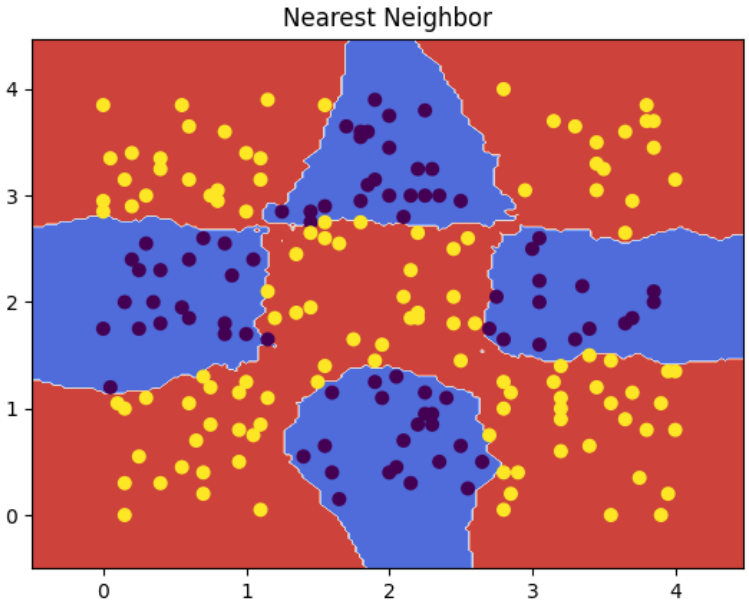

Figure 2: Nearest Neighbor plot

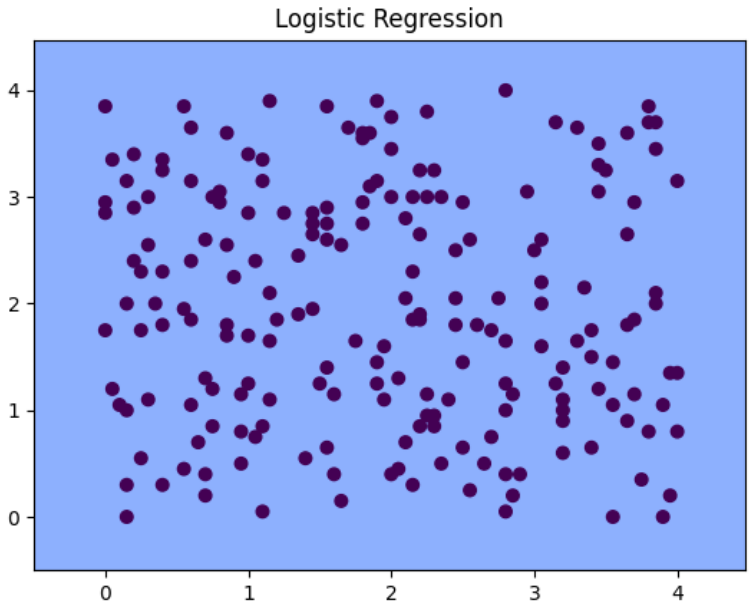

Figure 3: Logistic regression plot

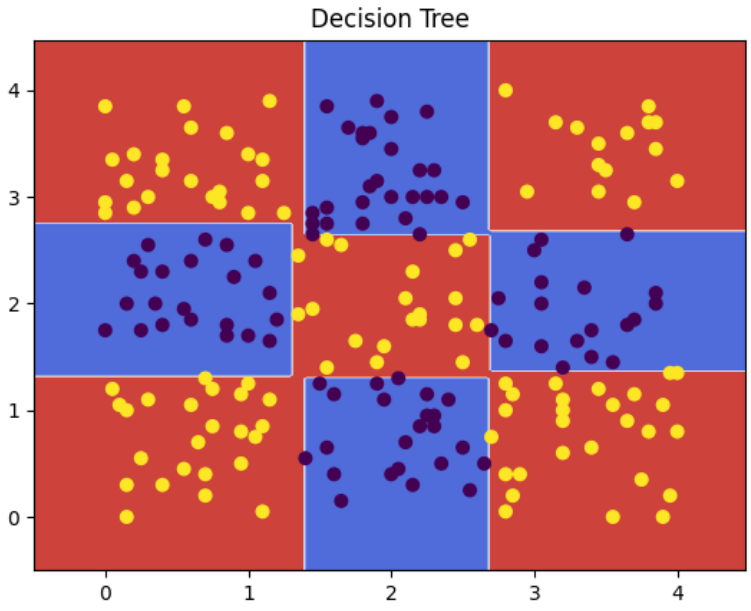

Figure 4: Decision tree plot

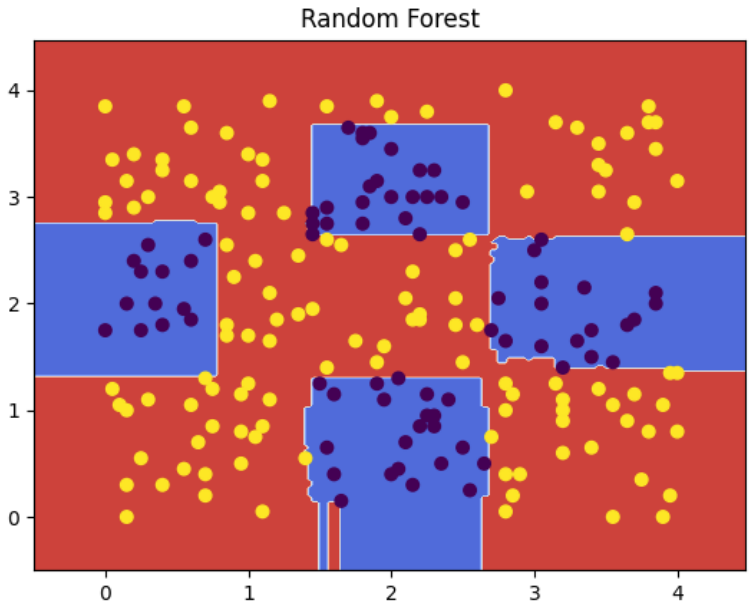

Figure 5: Random forest plot

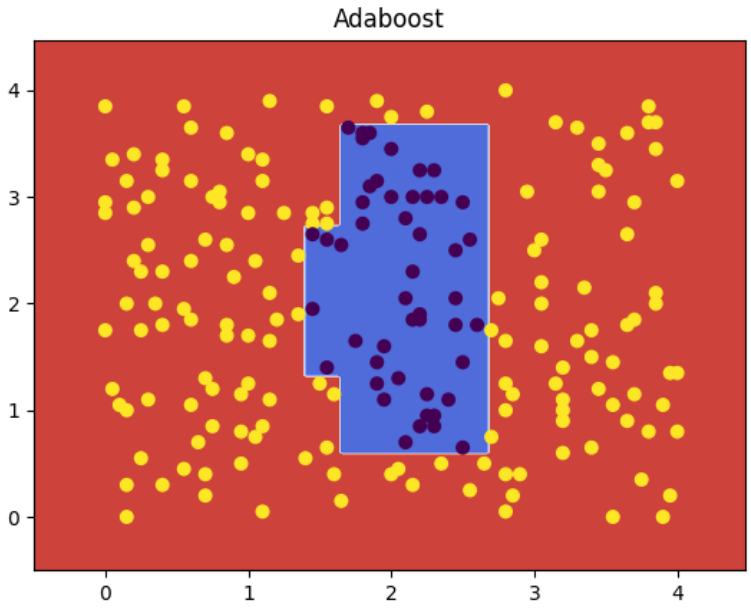

Figure 6: Adaboost plot