**Training and Overfitting**

Student: Kabirat Nasiru

University of the Cumberlands

2024 Spring - Statistics for Data Science (MSDS-531-A01) - First Bi-term

Course Advisor: Dr. Danny Barnes

2/25/2024

**Training Model**

**Figure P4p4F1**

The scatter plot titled "Training set" shows the relationship between two variables, x1​ and x2​. The plot includes two types of data points, represented by '+' and 'o' markers, suggesting that there are two distinct categories or classifications within the dataset. The data points '+' points are more dispersed across the range of x1 and x2​ looks more inconsistent, the 'o' points appear to be more tightly clustered in different areas suggesting a more straightforward or consistent relationship between x1 and x2.

**Figure P4p4F2**

These are two decision trees used to model decision-making processes. Tree (a) is simpler with 11 leaf nodes, while tree (b) is more complex with 24 leaf nodes. The simpler tree (a) might have higher bias but better generalization potential due to its lower complexity, making it more interpretable and possibly more robust to unseen data. On the other hand, the more complex tree (b) has a seems to have a higher risk of overfitting, as it may adapt too closely to the training data, capturing noise instead of underlying patterns. This tree could have lower bias but higher variance, potentially leading to less accurate predictions on new data.

**Recreating Pdf1**

We would need to simulate two sets of points that resemble the '+' and 'o' classes shown in the scatter plot. Multiple times we generated data points with specific statistical properties (mean, variance) that capture the apparent clustering and spread in the plot until we get something that looks alike as a plot in R, please find code attached.

**Figure 1**

*Replica Training setA black and white image of a training

Description automatically generated*

**Differences between new plot and Plot Pdf 3**

The Replica I created has slightly clustered points in the mid region and towards the right-hand side, while Pdf 3 has more precise markers in three different spots. The replica also lacks color, and this can be adjusted further into the analysis using the Color function in R setting it to red for distinction. The replica I made has more randomly distributed compared to pdf 3 with precise data points.

**Recreating Decision tree for Pdf 3**

We Start by creating the initial node, known as the root node, which encompasses the entire dataset. Next, we Partition the dataset into two categories marked by "+" and "o." Choose the optimal attributes that best separate the data into these categories. We then split the dataset into two subsets based on the chosen attribute: one subset where the x-coordinate is less than or equal to 0.5, and another where the x-coordinate is greater than 0.5. We continue for each subset, making further splits until the points in each final subset belong exclusively to one category. At the end of this process, each terminal node, or leaf node, is assigned labels.

**Code**



**Simulating Data for Classification**

Simulating data for classification is like pilot training. It offers a safe and controlled environment where you can create various scenarios to rigorously test and refine classification algorithms, like how a pilot practices in a simulator before flying a real plane. This approach is highly cost-effective as it avoids the complexities and expenses associated with collecting real-world data, when there is a need. Simulated data also allows for exact reproducibility in experiments, which is crucial for verifying and comparing results across different studies. It's a practical tool for checking if a model is overfitting, which means it's too tailored to the training data and may not perform well in real-world situations.

**References**

Borshchev, A. (2013). The big book of simulation modeling - AnyLogic simulation software. Retrieved July 20, 2021, from Anylogic North America website: http://www.anylogic.com/big-book-of-simulation-modeling

Favero, L. P., & Belfiore, P. (2019). *Data Science for Business and Decision Making*. Elsevier S & T. <https://reader2.yuzu.com/books/9780128112175>