RET Project: The Effect of Bias in Training Data using the APDS996 Color Sensor on the Arduino Nano 33 BLE Sense Board

Research Experience for Teachers (RET) Summer 2021

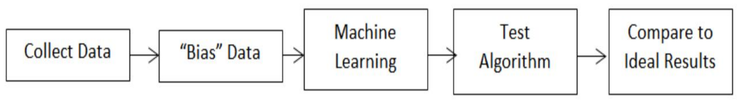
Brian Hawkins, Michael Stanley
SenSIP Center, School of ECEE, Arizona State University.

RESEARCH BACKGROUND/DESCRIPTION

- Machine learning is only as effective as the data used to create it.
- There are a variety of data sets used for training and testing machine learning algorithms so it is important to understand strengths and weaknesses of data sets.

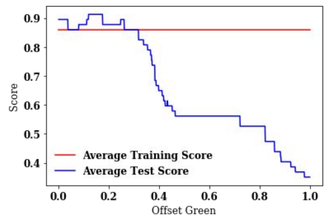
RESEARCH OBJECTIVES/PLAN

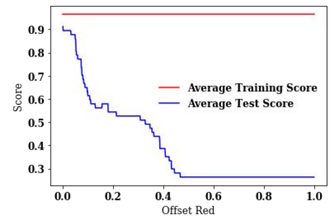
- Test the APDS996 color sensor using unequal and non normalized data sets.
- Modify a data set with an offset to study how it affects support vector machine algorithms and their corresponding confusion matrices.
- Study how the percentage of training data affects the performance of the support vector machine algorithm.

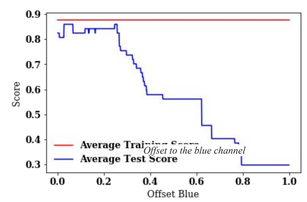


RESEARCH RESULTS/REMARKS

- Although the focus of the project was to examine how offsets added to the data affect the classification, there are several other parameters that can also affect misclassification.
- When using support vector machines, it is essential to understand how the selection of C, gamma, and the kernel function affect training and test scores.
- The radial basis function, C set at 100, and gamma set at .25 appeared to yield the optimum results for this project.
- Although the the training split is typically set to ⅔, slightly lower values may yield much lower training scores depending on the values chosen for the support vector machine.







Offset to the blue channel

REFERENCES

[1] Suresh, H., & Guttag, J. (2019). A Framework for Understanding Unintended Consequences of Machine Learning. *ArXiv, abs/1901.10002*.

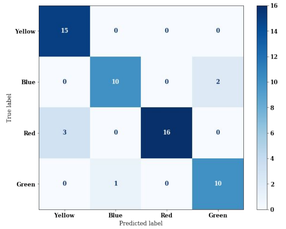
[2] Mehrabi, Ninareh & Morstatter, Fred & Saxena, Nripsuta & Lerman, Kristina & Galstyan, Aram. (2019). A Survey on Bias and Fairness in Machine Learning.

[3] Olteanu, A., Castillo, C., Diaz, F., & Kıcıman, E. (2019). Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. *Frontiers in big data*, *2*, 13. https://doi.org/10.3389/fdata.2019.00013

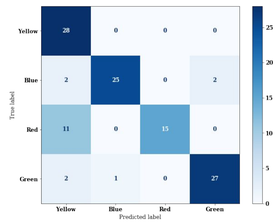
[4]Clark, C., Yatskar, M., & Zettlemoyer, L. (2019). Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. *arXiv preprint arXiv:1909.03683*.

LESSON PLAN OBJECTIVES

- Run the code in Google Colab to help students understand machine learning by examining confusion matrices and corresponding scatter plots.
- Test various data sets using the Support Vector Machine algorithm on Google Colab to create a confusion matrix and scatter plots.
- Extension - test their machine learning algorithms using items of student's choice



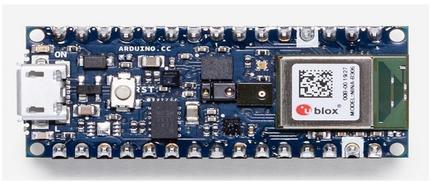
Confusion matrix for the testing set with .25 added to the red and green channels



Confusion matrix for the training set with .25 added to the red and green channels

LESSON IMPLEMENTATION/OUTCOMES

- This has not been accomplished yet but the expectation is that students would classification errors by examining the confusion matrices.
- I would also want students to examine the scatterplots created in Google Colab to understand if data points are close to the margin that they might be miscategorized.



Arduino Nano BLE 33 Sense Board

Ira A. Fulton Schools of Engineering
ARIZONA STATE UNIVERSITY

Sensor Signal and Information Processing Center
https://sensip.asu.edu

This work is funded in part by SenSIP center and NSF Award 1953745