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

Brian Hawkins
High School Engineering
Corona Del Sol High School, Tempe, AZ
Mentors: Michael Stanley, Dr. Andreas Spanias

NSF Award 1953745
https://sensip.engineering.asu.edu/ret/
RET Research and Training

https://sensip.engineering.asu.edu/ret/
Hands On Technical Training

- Arduino Integrated Development Environment (IDE)
  - Color sensor
  - Proximity sensor
- Jupyter was used for machine learning and graphing
  - Code management
  - Python
  - Support Vector Machine
    - Support vectors, C, gamma

https://sensip.engineering.asu.edu/ret/
RET Schedule and Training

Technical Exposition

• How sensors worked with derivation of formulas
• Kernal and sigmoid with probabilities
• Applications to speech and the complexity of speech
• Indoor air quality in cars
• How photodiodes scan and process data

https://sensip.engineering.asu.edu/ret/
RET Schedule and Training

Research Materials

• Mentor (Michael Stanley) - used extensively sometimes several times a day
• SciKit Learn - terminology and syntax of libraries
• Capabilities of Nano Sense board
• ASU online library - research potential topics of interest

https://sensip.engineering.asu.edu/ret/
RET Lab Experience Research Summary

Research Objectives

• What are the different ways that machine learning can be biased?
• What metrics are used to determine bias?
• How does the proximity sensor work?
• Types of data to collect with RGB sensor.

https://sensip.engineering.asu.edu/ret/
RET Lab Experience Research Summary

Research Background

• Normalization
• Under and Overfitting
• Equal data sets
• Sampling bias
• Data snooping
• Population bias

https://sensip.engineering.asu.edu/ret/
Research Proposal

• Difficult to pick a topic initially
• Focused too much on data collection instead of machine learning
• Abstract was straightforward
Research Conclusions

• Defining bias and fairness is difficult
• Color sensor data is uniform and has clearly defined classes so noise or an offset was necessary
• For students, it is meant to expose them to data that may not be clearly separated so they can see possible indicators in the graphs
• Metrics are necessary to tell the user if their machine learning is doing what they expect it to do
• Generalization error

https://sensip.engineering.asu.edu/ret/
Next STEPS in Research

• Examine more specific metrics that could be used to quantify the margin

• I mostly need to work on programming and creating additional graphs

• Possible work on margin distribution

• Original plan - Nano board, battery, micro SD card reader.
  • 3D printed enclosure

https://sensip.engineering.asu.edu/ret/
RET Instructional Lesson Implementation

Lesson 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

https://sensip.engineering.asu.edu/ret/
Lesson Description

Bias in Machine Learning

- Students will be working in groups to use the support vector algorithm to create confusion matrices and scatter plots on a set of given data. Groups will compare their results to the results of other groups.

- My goal is to have students understand a confusion matrix and corresponding scatter plots and how data bias can affect the outcome of a machine learning algorithm

- Most of the assessment will be formative while students are working, but students can have a quiz on confusion matrices and a short evaluation of the data that I provided their group.

Example of what I want students to see

https://sensip.engineering.asu.edu/ret/
Questions & Feedback

• What sorts of scaffolding opportunities do you think I should provide students, or any other suggestions?

https://sensip.engineering.asu.edu/ret/
Self Assessment

• The open ended nature of the program worked well and also didn’t.
  • Difficulty picking a problem, misunderstanding of machine learning, too much focus on the sensor and data collection, Python background
  • Early access to code samples
• Machine learning algorithms and terminology
• Code management in Jupyter
• ASU online libraries and Mendeley
• Flexible nature of the program
• One page summary and elevator pitches for students
References

Research Objectives


Thank You

• Andreas Spanias
• Michael Stanley
• Kristen Jaskie
• Jean Larson
• Ruby Sayed

If you are interested in looking at bias at all:
https://github.com/dssg/aequitas
https://dssg.github.io/fairness_tutorial/
https://textbook.coleridgeinitiative.org/chap-bias.html
https://fairmilbook.org/
https://www.amazon.com/Ethics-Data-Science-Mike-Loukides-ebook/dp/B07GTC8ZN7
https://www.solveforgood.org/

Best excerpt from my research

Disclaimer: No graduate students were harmed in the production of this paper. Authors are listed in order of increasing procrastination ability.