Derek Graves

07/12/2024

Applied Machine Learning Lab 3: Exploring and Manipulating Data

* Progress 1 - Screen shot prep code showing your name and environment

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

* Progress 2 - screen shot of model performanceA screenshot of a computer

  Description automatically generated
* Analysis 3 - Did model overfit the data?
  + The performance on the training set is consistently higher than on the test set by about 10% on all metrics. While this isn’t an extreme difference, it does suggest that the model may be somewhat overfitting the data, and that we could look at potentially making some improvements.
* Analysis 4 - Screen shot with performance on training set
  + Screenshot with performance on training set

A screenshot of a computer program

Description automatically generated

* + Which model do you prefer?
    - Based on the results of training both models, I would prefer the model which uses height as the feature.
  + Propose an explanation why one is better than the other.
    - The model using height as the feature has a more balanced performance between the training and test sets. Where the height model has a difference of around 10% between set performance metrics, the weight model, while having impressive performance in the training set metrics, shows some extreme differences between training and test metric scores, with test scores significantly lower. This indicates the weight feature model is overfitting. One explanation for this could be that the height feature in the data set has a more consistent relationship with gender, making it a better predictor for our model than weight, which may have more variability in its relationship with gender, making it a worse predictor and leading to the overfitting in the weight model.
* Analysis 5 - Which was the best feature set to train on?
  + The best feature set to train on has been the height only model up to this point as it clearly had best balance between training and test set performance metrics. Weight alone produced low test accuracy and F1 scores, and when we combined Height and Weight features, the combined model had perfect training performance but still showed a large drop in test performance, suggesting overfitting. It would be interesting to see if adding the age feature would have an improvement on model predictive power.
* Progress 6 - Screen shot of scatter plot

A screenshot of a computer

Description automatically generated

* Progress 7 - Screen shot tree plot

A diagram of a company structure

Description automatically generated

* Progress 8 - Screen shot of svc performance

A screenshot of a computer

Description automatically generated

* Analysis 9 - Compare SVC with other models
  + SVC vs. Height Only (Decision Tree)
    - The height-only decision tree model overall has a slightly higher training and test set performance with higher accuracy and F1 scores on both than SVC. While the height only decision tree model seems to be the best performing overall, SVC with both features included is a close second in performance to this model.
  + SVC vs. Weight Only (Decision Tree)
    - In training set performance, the weight-only decision tree model has a much higher training accuracy and F1, but SVC significantly outperforms this model in test set performance. The variability between the scores of the training and test sets show a much higher possibility of overfitting in the weight model than SVC, which had a less significant change between set performance metrics overall.
  + SVC vs. Height and Weight (Decision Tree)
    - The height and weight decision tree model had perfect training performance when compared to SVC but had a significant drop in performance between the training and test set. SVC again was the more realistic and stable model when comparing performance between the training and test sets and had less variability between the two.
* Analysis 10 - Support Vectors

A graph of a person and person

Description automatically generated with medium confidence

1. What can we learn from the support vectors?
   * Support vectors help to determine where the hyperplane, or decision boundary, lies in the data. They help to show the margin between the different classes represented by the data and can highlight where areas of uncertainty or overlap may be. They can help to show the model’s behavior and highlight where there could be future improvement.
2. Where are the boundaries of the positive/negative categories? (Sketch two lines on your figure)
   * (See screenshot of figure above)
3. In which regions should the prediction be tentative?
   * As the model gets closer to the decision boundary (the middle line between the margins for the support vectors), points are less likely to be classified correctly because they are equally close to both classes. The area between the margin lines should also be considered a region of tentative predictions, as it defines where the decision boundary lies and includes points that are closest to being identified as either one class or the other.

* Progress 11 - Screen shot NN performanceA screenshot of a computer

  Description automatically generated A screenshot of a training program

  Description automatically generated
* Analysis 12 - Compare NN with other models (and summarize your results)
  + Compared to the other models, our NN overall had a slightly lower test performance compared to the Decision Tree (Height Only) model, but this depends on the given run since performance can change with NN since it is randomized, so they are similar in performance. In comparison with the Height and Weight SVC model, the NN was able to slightly outperform it in test scores generally and was more consistent between training and test outcomes, though SVC was solid as well. In conclusion, where the Decision Tree for Height Only has the best performance, the NN comes in at a close second with slightly lower test performance but seems to have overall more balance in performance across its training and test sets. If we are only worried about performance, the Height Only Decision Tree seems to be the choice model for the data, but if we want to consider more real-world features, the NN model might be a better choice.