Danielle O'Sullivan
1402161

CS342: Assignment 1 Report

Abstract

This document describes the defining attribute to be TNA when classifying whether an organism is a Plant (Class 0) or an Animal (Class 1) from the 'Plants vs Animals' Assignment[1] along with details of how this conclusion came to be.

Results

The file *'classB.csv'* shows the probability of each entry being in either Class 0 or Class 1. The outcome was produced by splitting the data into training and test datasets. The training data (with corresponding target attribute) was fitted to a Decision Tree, *DecisionTreeClassifier*, using the *sklearn* module. The outcomes for the test dataset were compared to the corresponding labels and the proportion of correctly identified Class labels was noted. Outside of the '*predictClass.py*' script, a Cross Validation technique was applied to the classifier in order to combat overfitting and to ensure the Decision Tree was consistent across a number of tests. The average proportion of correctly identified Class labels was roughly 70%, which by the Receiver Operating Characteristic Area Under Curve (ROC AUC) is an acceptable result as the AUC should be above 50% for a more reliable binary classifier, as 50% is just as good as randomly choosing a class label for each entry.

This however was only achieved with the attributes 'TNA', 'Cryptonine-3' (C3) and 'Posidine-2' (P2). These attributes were decided to be the most informative after running all 10 available attributes through the Information Gain function (*infoGain*) where TNA, C3 and P2 gave the highest scores, 0.180, 0.102 and 0.031, respectively. For context, the other scores averaged around 0.00001. These calculations show that TNA is the most influential attribute by quite a large distance, and since Probe B has TNA values missing, the logical step is to generate some predictions for TNA values first and only after that produce the Class labels for Probe B.

Further investigation showed that when removing TNA from the classifier, the accuracy of predictions decreased quite substantially (the results have been omitted from the script for clarity). Thus, a lower predictive power for the training set surely means an even lower predictive power when predicting TNA values, as we not only remove the most influential attribute, but we also move from binary classification, to (continuous) regression; leaving much more room for error, and in mathematical theory, infinite room for error!

Conclusion

The attribute which is more predictive in whether a life-form is a plant or an animal is the TNA value due to its score for information gain and visually with the plots showing the clearest cut classification is when splitting the data by TNA value (please run commented code for a visualisation).

[1] CS342 Assignment 1 Briefing Sheet:
http://www2.warwick.ac.uk/fac/sci/dcs/teaching/material/cs342/cs342_assign1.pdf