CS342 Coursework 1: Plants vs Animals

Hersh Kataria u1620109

November 3, 2018

1 Data Pre-processing

I initially decided to analyse the data distribution of each chemical resolution by plotting histograms for each attribute in probes A and B. This indicated that the data for individual chemical compounds was corrupt. The resolutions for each compound were reordered from smallest to largest to fix this.

This proved to be the most significant pre-processing step which improved the accuracy of my models. Feature expansion was the next pre-processing step I carried out. I generated a new data frame with all the second degree polynomial features (multiplying all attributes with one another), giving me 91 attributes to work with rather than 12.

The final step was feature selection which allowed for higher accuracy and also increased performance by reducing the dimensionality. A tree-based estimator was used to calculate feature importances, and then unimportant features were dropped based on a threshold value using the SelectFromModel method.

AUC (Area Under the ROC Curve) was used to determine the accuracy of the classification models and R Squared was used for TNA prediction, whilst using 10-fold cross validation.

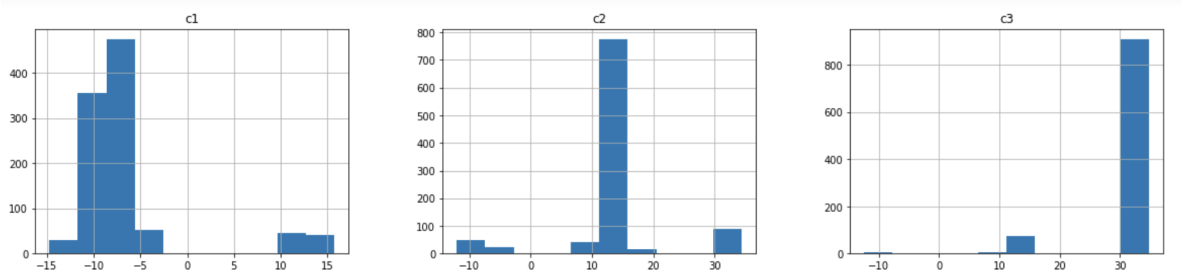

Figure 1: Histograms showing data corruption of a chemical compound.

2 Class Prediction

It was clear that 'tna' was the most predictive feature for class, by anaylsing feature importances from the decision tree classifier. Probe B did not have any data on 'tna' so this feature could not be used unless regression was used to predict 'tna' and then predicting class using that. However, this proved not to as successful as models without tna.

I tested using decision trees and kNN models, exhaustively searching for the optimum depth and neighbours respectively, and found kNN to give higher AUC scores.

Feature importances of the second degree polynomial features, without tna, were analysed using the feature_importances_ attribute of the decision tree model and highly predictive features were identified

according to their Gini impurity value. The most notable of these predictive features were 'c3*m2' and 'm2' as shown in Figure 2.

I then determined how many of these top important attributes to use, and how many neighbours to use in the kNN model. I carried out an exhaustive search to find these values. In the end, the strongest model proved to be a 24-NN classifier with features [m2, n2, p1, $c3^2$, c3 m2, c3 n2, m1 p3, $n2^2$, $n3^2$, n3 p1, $p2^2$], giving an AUC score of 0.7726.

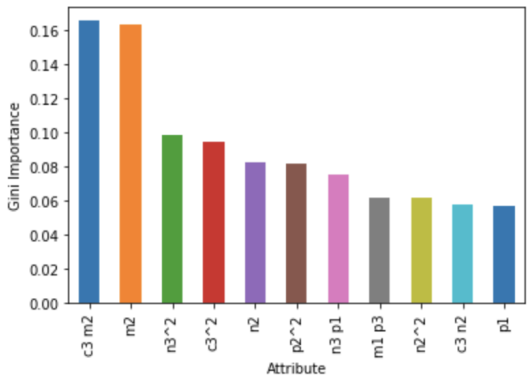

Figure 2: Highest Gini impurity values of 2nd order polynomial features for class prediction.

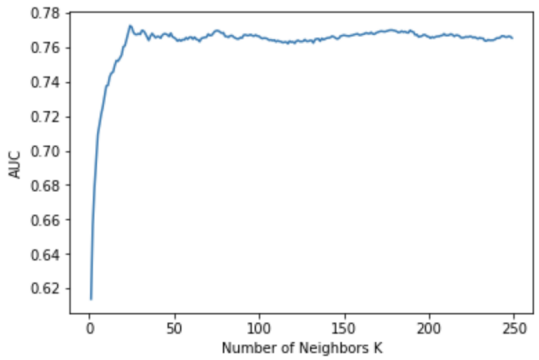

Figure 3: AUC against K (number of neighbours). Optimum K was 24.

3 TNA prediction

TNA predictions were tested using Lasso, Ridge and Decision Tree regression. Using the second order polynomial features on the decision tree regressor, with a max depth of 5, gave an R2 score of 0.7647. The same features were tested using LassoCV and RidgeCV, giving R2 scores of 0.9223 and 0.9145 respectively. Although the decision tree was not the most accurate model, it did provide us with the feature importances, indicating that c3*m2 and $n3^2$ were the most predictive features, as shown in Figure 4. After an exhaustive search on the value of alpha for Ridge and Lasso, the strongest model proved to be Lasso with an alpha value of 0.0011, giving an R2 score of 0.9230.

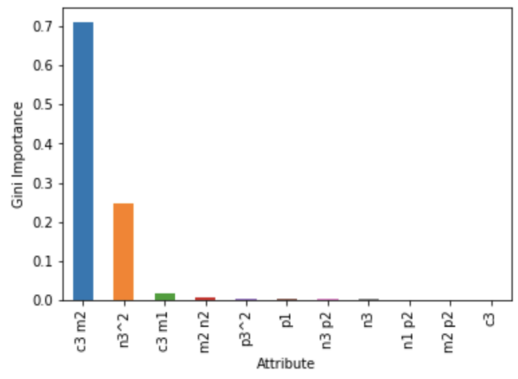

Figure 4: Highest Gini impurity values of 2nd order polynomial features for TNA prediction.

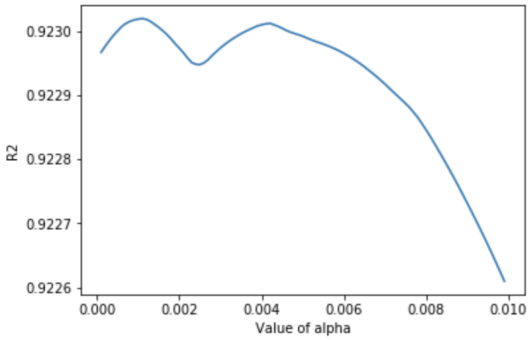

Figure 5: R2 against alpha for Ridge regression. Optimum alpha value was 0.0011.