1R Mammals Report

Name: Austin Sampson

eMail: aws9t5@mst.edu

Course: CS 5402

Date: 02-25-2020

Concept Description:

Train a system from existing data to classify animals as amphibian, anthropod, bird, fish, insect, mammal or reptile. Calssification will be done through Naive Classification and ID3 and model will be selected after comparing their respective accuracy and error rates to present to the empoloyer.

Data Collection:

Data has been provided from the client based off the observation of their feild agents. The training and test data sets provided by the client are tranimal-taxonomy-train.csv and animal-taxonomy-test.csv respectivly.

Example Description:

animal.name

nominal attribute name of the animal or species This attribute was removed from the training and test data sets to prevent overfitting

hair

Nominal boolean attribute that displays output as: True False

feathers

Nominal boolean attribute that displays output as: True False

eggs

Nominal boolean attribute that displays output as: True False

milk

Nominal boolean attribute that displays output as: True False

airborne

Nominal boolean attribute that displays output as: True False

aquatic

Nominal boolean attribute that displays output as: True False

preditor

Nominal boolean attribute that displays output as: True False

toothed

Nominal boolean attribute that displays output as: True False

backbone

Nominal boolean attribute that displays output as: True False

breathes

Nominal boolean attribute that displays output as: True False

venomous

Nominal boolean attribute that displays output as: True False

fins

Nominal boolean attribute that displays output as: True False

legs

Ratio Lable displaying the number of legs. null value or 0 indicates the absince of legs. This data was filtered out of the test and training data due to containing zero values.

tail

Nominal boolean attribute that displays output as: True False

domestic

Nominal boolean attribute that displays output as: True False

catsize

Nominal boolean attribute that displays output as: True False

gestation

Interval attribute displays a measure of time it took for the gestation of a species. this data was removed from the training and test data sets do to missing values.

type

Nominal, Main classificationn variable for this data set. Output displayed as: mammal fish arthropod bird insect amphibian reptile

Data Import and Wrangling:

Importing test and training data

```
#import main file
train <- read.csv("animal-taxonomy-train.csv")
test <- read.csv("animal-taxonomy-test.csv")
```

We remove the name attribute to prevent the moddel from overfitting by training for name attributes. We also set the legs and gestation attributes to null do to missing and zero values.

```
#remove the name attribute from the test and training data

train$animal.name <- NULL
test$animal.name <- NULL
train$gestation <- NULL
test$gestation <- NULL
train$legs <- NULL
train$legs <- NULL
```

Mining and Analytics:

Naive Bayes Classifiers

Traing the Confusion Matrix for bayes classification

```
train.nb <- naiveBayes(type ~ ., data=train)
train.nb

##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = X, y = Y, laplace = laplace)
##
## A-priori probabilities:
## Y
##  amphibian  arthropod       bird       fish     insect     mammal
reptile
## 0.04395604 0.10989011 0.18681319 0.13186813 0.08791209 0.38461538
0.05494505
##
## Conditional probabilities:
##            hair
## Y                FALSE       TRUE
##   amphibian 1.00000000 0.00000000
##   arthropod 1.00000000 0.00000000
##   bird      1.00000000 0.00000000
##   fish      1.00000000 0.00000000
##   insect    0.50000000 0.50000000
```

```
##   mammal    0.02857143 0.97142857
##   reptile   1.00000000 0.00000000
##
##            feathers
## Y           FALSE TRUE
##   amphibian     1    0
##   arthropod     1    0
##   bird          0    1
##   fish          1    0
##   insect        1    0
##   mammal        1    0
##   reptile       1    0
##
##            eggs
## Y           FALSE TRUE
##   amphibian   0.0  1.0
##   arthropod   0.1  0.9
##   bird        0.0  1.0
##   fish        0.0  1.0
##   insect      0.0  1.0
##   mammal      1.0  0.0
##   reptile     0.2  0.8
##
##            milk
## Y           FALSE TRUE
##   amphibian     1    0
##   arthropod     1    0
##   bird          1    0
##   fish          1    0
##   insect        1    0
##   mammal        0    1
##   reptile       1    0
##
##            airborne
## Y                FALSE       TRUE
##   amphibian 1.00000000 0.00000000
##   arthropod 1.00000000 0.00000000
##   bird      0.23529412 0.76470588
##   fish      1.00000000 0.00000000
##   insect    0.25000000 0.75000000
##   mammal    0.94285714 0.05714286
##   reptile   1.00000000 0.00000000
##
##            aquatic
## Y               FALSE      TRUE
##   amphibian 0.0000000 1.0000000
##   arthropod 0.4000000 0.6000000
##   bird      0.6470588 0.3529412
##   fish      0.0000000 1.0000000
##   insect    1.0000000 0.0000000
```

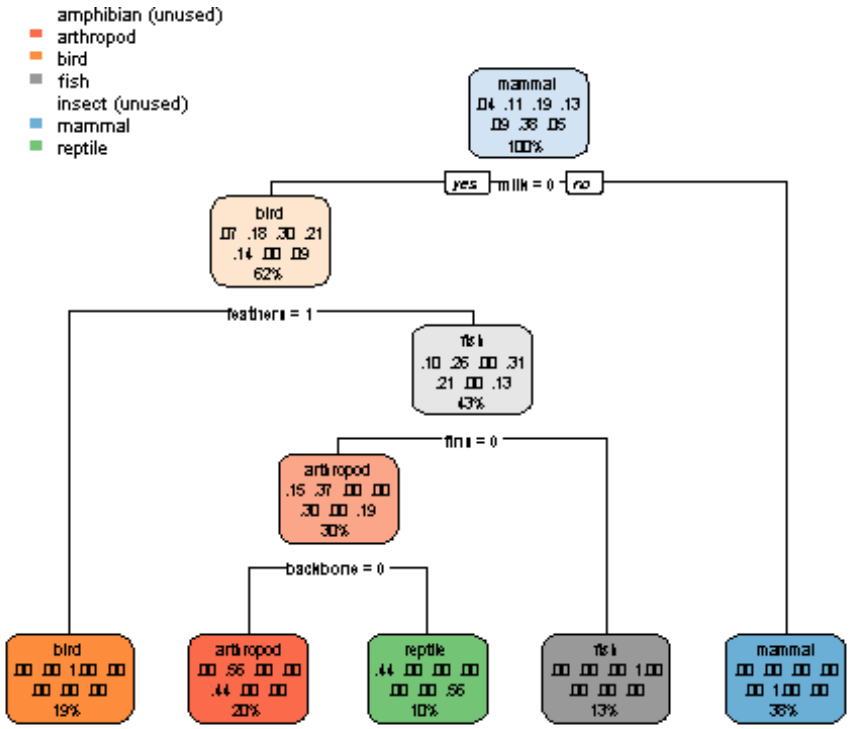
```
##   mammal    0.8857143 0.1142857
##   reptile   0.8000000 0.2000000
##
##            predator
## Y               FALSE      TRUE
##   amphibian 0.2500000 0.7500000
##   arthropod 0.2000000 0.8000000
##   bird      0.4705882 0.5294118
##   fish      0.3333333 0.6666667
##   insect    0.8750000 0.1250000
##   mammal    0.5142857 0.4857143
##   reptile   0.2000000 0.8000000
##
##            toothed
## Y           FALSE TRUE
##   amphibian   0.0  1.0
##   arthropod   1.0  0.0
##   bird        1.0  0.0
##   fish        0.0  1.0
##   insect      1.0  0.0
##   mammal      0.0  1.0
##   reptile     0.2  0.8
##
##            backbone
## Y           FALSE TRUE
##   amphibian     0    1
##   arthropod     1    0
##   bird          0    1
##   fish          0    1
##   insect        1    0
##   mammal        0    1
##   reptile       0    1
##
##            breathes
## Y           FALSE TRUE
##   amphibian   0.0  1.0
##   arthropod   0.7  0.3
##   bird        0.0  1.0
##   fish        1.0  0.0
##   insect      0.0  1.0
##   mammal      0.0  1.0
##   reptile     0.2  0.8
##
##            venomous
## Y                FALSE       TRUE
##   amphibian 0.75000000 0.25000000
##   arthropod 0.80000000 0.20000000
##   bird      1.00000000 0.00000000
##   fish      0.91666667 0.08333333
##   insect    0.75000000 0.25000000
```

```
##   mammal    1.00000000 0.00000000
##   reptile   0.60000000 0.40000000
##
##            fins
## Y                FALSE       TRUE
##   amphibian 1.00000000 0.00000000
##   arthropod 1.00000000 0.00000000
##   bird      1.00000000 0.00000000
##   fish      0.00000000 1.00000000
##   insect    1.00000000 0.00000000
##   mammal    0.91428571 0.08571429
##   reptile   1.00000000 0.00000000
##
##            tail
## Y               FALSE      TRUE
##   amphibian 0.7500000 0.2500000
##   arthropod 0.9000000 0.1000000
##   bird      0.0000000 1.0000000
##   fish      0.0000000 1.0000000
##   insect    1.0000000 0.0000000
##   mammal    0.1428571 0.8571429
##   reptile   0.0000000 1.0000000
##
##            domestic
## Y                FALSE       TRUE
##   amphibian 1.00000000 0.00000000
##   arthropod 1.00000000 0.00000000
##   bird      0.94117647 0.05882353
##   fish      0.91666667 0.08333333
##   insect    0.87500000 0.12500000
##   mammal    0.77142857 0.22857143
##   reptile   1.00000000 0.00000000
##
##            catsize
## Y               FALSE      TRUE
##   amphibian 1.0000000 0.0000000
##   arthropod 0.9000000 0.1000000
##   bird      0.7058824 0.2941176
##   fish      0.7500000 0.2500000
##   insect    1.0000000 0.0000000
##   mammal    0.2571429 0.7428571
##   reptile   0.8000000 0.2000000
```

ID3

To train the ID3 Classifier I untailzied the rpart package and their graphing tools

```
#train the id3 model using the rpart library
model.id3 <-rpart(type ~., data = train)
rpart.plot(model.id3)
```



Evaluation:

ID3 Calculating the Confusion matrix for ID3 decision tree classification Using the model create a prediction for the dataset and compare it to the actual

```
#predict classification using id3 model
pred.id3 <- predict(model.id3, test,type="class")
#print the generated classification decision tree

#Calculated the confusion matrix based off the prediction
pred.id3 <- droplevels(pred.id3)
confusionMatrix(pred.id3,test$type)

## Confusion Matrix and Statistics
##
##           Reference
## Prediction bird fish mammal
##     bird      3    0      0
##     fish      0    1      0
##     mammal    0    0      6
##
## Overall Statistics
##
##                Accuracy : 1
##                  95% CI : (0.6915, 1)
##     No Information Rate : 0.6
```

```
##     P-Value [Acc > NIR] : 0.006047
##
##                   Kappa : 1
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: bird Class: fish Class: mammal
## Sensitivity                  1.0         1.0           1.0
## Specificity                  1.0         1.0           1.0
## Pos Pred Value               1.0         1.0           1.0
## Neg Pred Value               1.0         1.0           1.0
## Prevalence                   0.3         0.1           0.6
## Detection Rate               0.3         0.1           0.6
## Detection Prevalence         0.3         0.1           0.6
## Balanced Accuracy            1.0         1.0           1.0
```

Bayes Calculating confusion matrix for Bayes I had re-added the levels that where missing from both tables in order to calculate the confusion matrix.

```
m_predictions <- predict(train.nb, test)

#table(m_predictions);
#table(test);



m_predictions <- droplevels(m_predictions)
cfm <- confusionMatrix(m_predictions, test$type)
cfm

## Confusion Matrix and Statistics
##
##           Reference
## Prediction bird fish mammal
##     bird      3    0      0
##     fish      0    1      0
##     mammal    0    0      6
##
## Overall Statistics
##
##                Accuracy : 1
##                  95% CI : (0.6915, 1)
##     No Information Rate : 0.6
##     P-Value [Acc > NIR] : 0.006047
##
##                   Kappa : 1
##
##  Mcnemar's Test P-Value : NA
##
```

```
## Statistics by Class:
##
##                      Class: bird Class: fish Class: mammal
## Sensitivity                  1.0         1.0           1.0
## Specificity                  1.0         1.0           1.0
## Pos Pred Value               1.0         1.0           1.0
## Neg Pred Value               1.0         1.0           1.0
## Prevalence                   0.3         0.1           0.6
## Detection Rate               0.3         0.1           0.6
## Detection Prevalence         0.3         0.1           0.6
## Balanced Accuracy            1.0         1.0           1.0
```

Accuracy and Error Rates

As can be shown above due to both the bayes and ID3 algorithm generating accuracies of 1, we know both models have an error rate of 0%.

ID3

Results

Since both the generated classification modesl produced a 100% accuracy rating from their confusion matrixes based off our test data and therefore had error rates of 0%. We can conclude that both the Naive Bayes and ID3 models are equivalent in their ability to produce accurate and relaible results for the classification of our data.

Out of the two models examened I would recomend presenting the ID3 model to the client due to its ease of use and graphical representation. Another reason to choose ID3 would be because it does not operate under the assumption that all the data attributes are independent Like Naive Bayes.

References:

https://www.rdocumentation.org/packages/caret/versions/3.45/topics/confusionMatrix
https://www.rdocumentation.org/packages/utils/versions/3.6.2/topics/data
http://www.learnbymarketing.com/tutorials/naive-bayes-in-r/
https://www.rdocumentation.org/packages/rpart/versions/4.1-15/topics/rpart
http://www.milbo.org/rpart-plot/prp.pdf