Introduction to Statistical Machine Learning for Bioinformaticians and Medical Informaticians

SoSe 2022

Tutor: Dana Petracek

1	$\sum$

Marina Dittschar & Clarissa Auckenthaler

Assignment 9

(Submitted 11.07.2022)

Task 1

The train data set contains 22 features, one response variable and overall 1000 observations. The test set has also 22 feature variables but no response variable and overall 2000 observations. To get more details about our data set, we plotted the distribution 3 of the response variable for the training set. We can clearly see that with 315 occurrences in our training set the frog Family "Hylidae" occurs approximately half as often as the frog family "Leptodactylidae" (685 occurrences).

First, we decided to split our train data set further, into training (70% of 1000 = 700) and test set (30% of 1000 = 300) to train the different models and compare their accuracy and the Matthews correlation coefficient (MCC). Equation 1 shows the MCC formula. Before we started with prediction we normalized our data. Then, we selected different methods (logistic regression, tree, naive bayes classifier, random forest and support vector machine) to learn on our data. In some models, we also performed parameter tuning.

$$MCC = \frac{(TP * TN - FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{1}$$

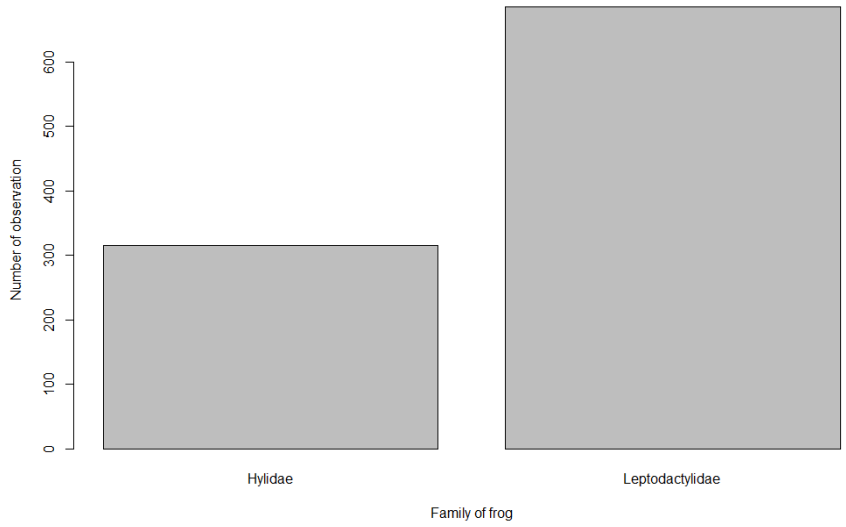

Figure 1: Distribution of the training set

Logistic Regression

As the first model of prediction we selected logistic regression.

Result:

$MCC = \mathbf{0.8922808}$
$ACC = \mathbf{0.9533333}$

	Hylidae	**Leptodactylidae**
Hylidae	88	6
Leptodactylidae	8	198

Leaving out the not significant features (MFCC_7, MFCC_22, MFCC_6, MFCC_9, MFCC_1, MFCC_2), lead to a slightly better result:
$MCC = \mathbf{0.9080882}$
$ACC = \mathbf{0.96}$

	Hylidae	**Leptodactylidae**
Hylidae	90	6
Leptodactylidae	6	198

Compared to the first logistic regression model it misclassifies two observations fewer.

Tree

For the next model, we selected a default tree model. Figure 2 displays our trained model and the decision variables.
Results:
$MCC = \mathbf{0.876505}$
$ACC = \mathbf{0.9466667}$

	Hylidae	**Leptodactylidae**
Hylidae	86	6
Leptodactylidae	10	198

Our tree-model performed even worse than the logistic regression model trained on all features. We decided not to continue with the tree based on this results.

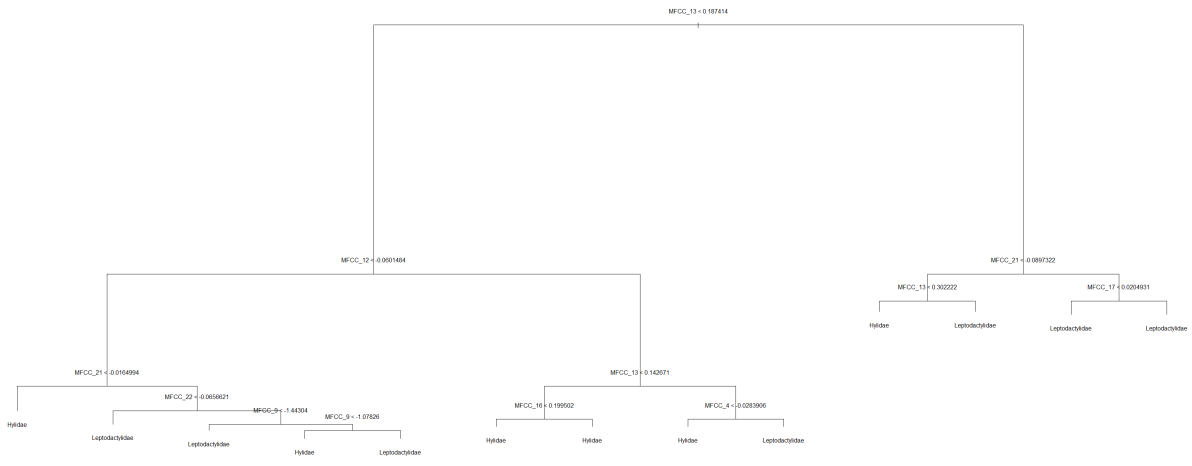

Figure 2: Our Tree model

Naive Bayes Classifier

Then we continued with the naive bayes classifier.
Results:
$MCC = \mathbf{0.8616427}$
$ACC = \mathbf{0.9366667}$

	Hylidae	**Leptodactylidae**
Hylidae	92	18
Leptodactylidae	4	189

This model performed better in prediction of the frog family "Leptodactylidae" compared to logistic regression model and tree model.
But clearly compared to the results before this is definitely the worst model, so we also decided to skip further improvement on this model. While 93% accuracy is still a good performance, our approaches was to find the best model.

Random Forest

Because we had good experience with trained random forest models we decided to train this model next. First, we trained a random forest model with default settings.
Results:
$MCC = \mathbf{0.9463491}$
$ACC = \mathbf{0.9766667}$

	Hylidae	**Leptodactylidae**
Hylidae	90	1
Leptodactylidae	6	203

Only one instance of the "Hylidae" family and 6 instances of the "Leptodactylidae" were misclassified. Since this was the best performance we achieved so far, we decided to tune the parameters *mtry* and *ntree*.

First we performed Bagging:
Results:
$MCC = \mathbf{0.9462645}$
$ACC = \mathbf{0.9766667}$

	Hylidae	**Leptodactylidae**
Hylidae	92	3
Leptodactylidae	4	201

This did not really improve misclassification rates.
As a next step we selected different mtry values (1 to 22), for each *mtry* value testing different *ntree* values (200, 500, 1000). In the beginning we performed it for even more *ntree* values (100, 200, 500, 1000, 2000) but to reduce the run time, we only selected 3 values when it became apparent that performance did not improve on either side of the range.
For each model, we calculated the MCC and stored all values in arrays. After the loop, we selected the maximum value of the MCC-array and identified the matching *mtry* (7) and *ntree* (200) values. We then again fitted the model on the training data.
Results:

$MCC = \mathbf{0.9693079}$
$ACC = \mathbf{0.9866667}$

	Hylidae	**Leptodactylidae**
Hylidae	93	1
Leptodactylidae	3	203

This model predicted nearly every observation correctly. The accuracy was 98.6% and only 4 instances were misclassified. Compared to all other methods before this, this was now the best performing model with an MCC of 96.9%.

Support vector machine

We had never trained a support vector machine before, so we decided to add this method and see if it performed better compared to the other methods and especially to our best performing model so far. We also tried different kernel types:

- Linear Kernel Results:
  $MCC = \mathbf{0.9152601}$
  $ACC = \mathbf{0.9633333}$

	Hylidae	**Leptodactylidae**
Hylidae	88	8
Leptodactylidae	3	201

  The simple linear kernel performed quite well but compared to our best model with 98.6% accuracy this model is over 2% less accurate.

- Radial Kernel For this model we tuned the parameter gamma and cost on different values with the radial kerne and selected then the best performing model. Results:
  $MCC = \mathbf{0.8715592}$
  $ACC = \mathbf{0.9433333}$

	Hylidae	**Leptodactylidae**
Hylidae	79	17
Leptodactylidae	0	204

  Because this model was not really performing well, we commented the code out, to reduce the run time.

- Tuned polynomial Kernel

  For this model we tuned the parameter gamma and cost on different values with the polynomial kernel and selected then the best performing model. Results:
  $MCC = \mathbf{0.9387255}$
  $ACC = \mathbf{0.9733333}$

	Hylidae	**Leptodactylidae**
Hylidae	92	4
Leptodactylidae	4	200

This model perform better compared to the support vector machine model with a linear kernel, but it was still 1.3% less accurate as the tuned random forest model. However, as the difference was so small, trying the support vector machine approach was still worth it.

Model Selection

Based on the accuracy and the Matthews correlation coefficient our trained random forest model with tuned *mtry* and *ntree* values performed best.
Because the model training is based on the number of observations we decided to select our best performing model and train it again but this time on the whole training set and predicted the values and we hope to reduce the test error and improve accuracy and mcc for the whole test set of 2000 observations.
Results prediction:

	Hylidae	**Leptodactylidae**
Hylidae	315	0
Leptodactylidae	0	685

Important: this results are not meaningful because we take the whole training set and calculate the performance based on the response of the response of the whole training set!

Results final prediction:

Hylidae	**Leptodactylidae**
682	1318

Distribution:

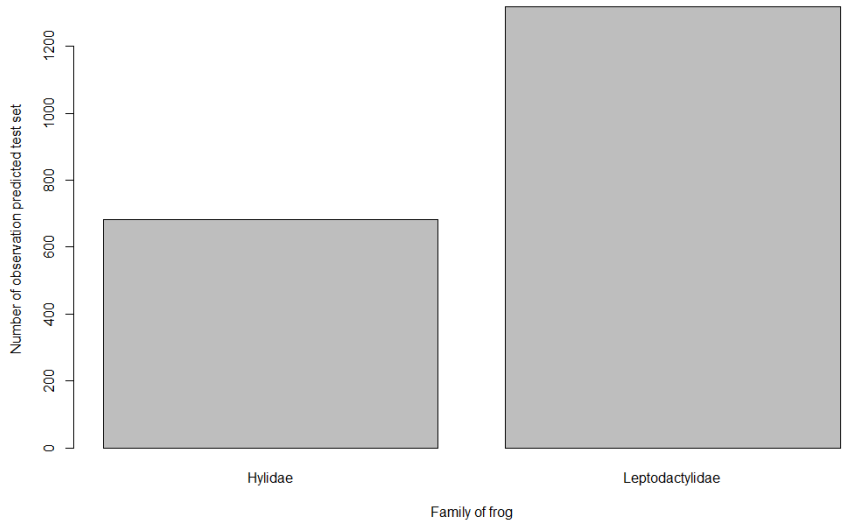

Figure 3: Distribution of the training set

The prediction of the test data response is stored in a text file named:
`dittschar_auckenthaler_test_response.txt`
The run time of our script is approximately 30 seconds.