Introduction to Statistical Machine Learning for Bioinformaticians and Medical Informaticians, SS 2022
Dr. Ralf Eggeling
Dana Petracek
Methods in Medical Informatics
University of Tuebingen

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Assignment 9

Deadline: Monday, July 11, 2:00 p.m.

This problem set is worth 25 points. You can submit in groups of two people or alone. Submit your solutions by uploading them to moodle (none of the other students can see the files you upload). Name the files

`Assignment_9_[lastname].pdf` and `Assignment_9_[lastname].R`
for individual submissions and
`Assignment_9_[lastname1]_[lastname2].pdf` and `Assignment_9_[lastname1]_[lastname2].R`
for team submissions.

In the latter case, include names of both students at the top of both files. Both students must upload the identical files to moodle in time.

Moodle allows to upload **drafts**, which can then be further edited until the deadline. Use this for submitting finished tasks in case you might run out of time. Once you formally submit, the upload cannot be edited any longer and is ready for grading. If you never formally submit, the uploaded draft at the submission deadline will be graded.

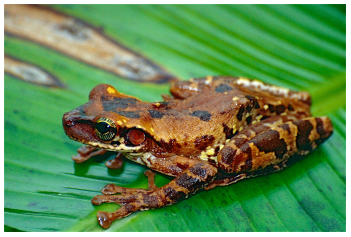


Figure 1: Exemplary members of the two Anura families of interest. Left: *Osteocephalus oophagus* as member of the *Hylidae* family. Right: *Adenomera andreae* as member of the Leptodactylidae family.

Task 1 (25 Points)

The data set Anuran Calls [1] contains acoustic features extracted from syllables of anuran (frogs) calls. The features are Mel-frequency cepstral coefficients (MFCCs) that were calculated from the .wav file of the original recording.

Your task is to build a machine learning model that disguishes the two Anura families Hylidae (tree frogs) and Leptodactylidae (southern frogs) based on their acoustic features as well as possible. For this purpose, you may freely choose among the methods covered in the course. In addition to base-R you may use all R-packages that are mentioned either in the ISLR book or on one of the assignment sheets (but nothing else).

(a) Download the file `AnuranCalls_train.csv`, a reduced version of the original UCI data set, from moodle. It contains 1000 data points, 22 continuous features, and a binary response *Family*. Build classifiers that predict the response based on all of the features and evalute their performance. Your submission will be judged according to validity of the evaluation, clarity of presentation, and overall depth of the analysis. For earning full points, you should try at least three different classification approaches. Use at most eight pages of text to describe your analysis and results; all pages beyond that will not be considered. (**25 Points**)

(b) Bonus task: Download the file `AnuranCalls_test.csv` from moodle. It contains the 22 features of 2000 additional data points, but the response *Family* is omitted. Predict the response using your best-performing

[1] https://archive.ics.uci.edu/ml/datasets/Anuran+Calls+%28MFCCs%29

Dr. Ralf Eggeling
Dana Petracek
Methods in Medical Informatics
University of Tuebingen

model from Task a) and store the predicted values in a .txt file, one line per prediction (same naming conventions as with .pdf and .R file). The order of predictions should be the same as the order of test data points. Your prediction will be evaluated with respect to the ground truth in terms of Matthews' correlation coefficient and ranked against the other submissions. The best-performing submissions will receive bonus points as follows:

Rank	1	2	3	4	5	6	7	8	9	10
Bonus points	25	18	15	12	10	8	6	4	2	1

Note:

- In case of a tie, the earlier submission in moodle receives precedence in the ranking.
- By submitting a solution to the bonus task, you agree that your position in the ranking, the numerical score, and a summary of the used analysis approaches is shared publicly with the class.