Machine Learning in Life Sciences: Exercise 1

Tutors: Anup Kumar, Simon Bray, Sven Hauns
Date: 12.05.22

Exercise 1

a) Download the dataset and preprocess it.

https://github.com/BackofenLab/MLLS-exercise-SS22/blob/main/01-introduction-ml/ELAVL1_PARCLIP

Context: ELAVL1 (https://en.wikipedia.org/wiki/ELAV-like_protein_1) is a RNA-binding protein. Here, we are interested in predicting, based on the RNA sequence, whether it will bind to the ELAVL1 protein or not. The dataset contains 10000 samples like the following:

```
>ID0|1
AAAUCUUUAUUUUUCUAGGACAUGUUAUGCCUCCAUUUUCAAUUAAAAUAAAGUUAUCGGA
UUACACCACCACCAGGGGUC
```

On the first line, ID0 refers to the sample ID, and 1 to the target variable (1 = binding, 0 = non-binding). The second line is the RNA sequence.

Hint: To featurize RNA/DNA sequences, a good strategy is to use k-mers. k=3 is a good choice and will result in 4^3 = 64 predictor variables.

b) Split both the datasets created in the previous step into training and test datasets. **Hint**: Use ‘train_test_split‘ from scikit-learn.

Exercise 2

a) Choose two of the following classification algorithms and apply it to the dataset created above. **Hint**: Use scikit-learn.

a.1) Linear model - Logistic Regression (remember - it’s a classifier)

a.2) Nearest neighbor - K nearest neighbor

a.4) Classification tree - Decision tree

a.5) Ensemble model - Random forest, gradient boosting tree

b) Calculate the following metrics: accuracy, precision, recall, F1, AUROC. What do each of these tell you? How do you overall assess the performance of your two models?

Exercise 3

What is overfitting? What is underfitting? Why are they a challenge when developing a model? What are good strategies to avoid them? Write 1-2 sentences in answer to each of these questions.

Have fun!