Assignment no. 4

Machine Learning:
Supervised Techniques
365.076 (1UE) WS 2017

Exercise 6 (20 points) Download the data table `Sequences_train.csv`. It contains 2,000 amino acid sequences with length 15 in the first column and class labels $-1/1$ in the second column. Apply an SVM approach for classifying these sequences. Use cross validation to determine the best possible kernel (see below) and the best model parameters (C and kernel parameters). Use a clear and transparent model selection procedure, i.e. grid search for parameters. Use the following kernels:

1. Implement the normalized spectrum kernel with adjustable subsequence length K.
2. Map sequences to vectorial data using a so-called *one-hot encoding* and apply linear and RBF kernel to these data.

Your final submission should include all your source code and a report documenting your model selection procedure and a summary of your results. Finally, train a model with your best parameters on the entire training set and predict the class for the sequences in the text file `Sequences_test_unlabeled.csv` and submit the predictions as a text file (one label $-1/1$ per line).

Data Analysis Contest: The three models performing best on an independent test set with 2,000 other sequences will be awarded 3 extra points.

Note: For the sake of fairness between R and Python users, it is not allowed to use sequence kernels available in existing software packages (e.g. R packages `kebabs` and `kernlab`). You are explicitly required to implement the spectrum kernel yourself!

Submission: electronically via Moodle:

`https://moodle.jku.at/jku2015/course/view.php?id=2634`

Please take the submission instructions into account! Deadline: Monday, January 8, 2018, 1:00pm.