Nucleome Aptitude Test ML

Instructions:

The following task is designed to be done in your own time on your own computer and should be answered in conjunction with the three corresponding files which will also be sent to you. Please provide the answer file (as described below), a copy or notebook of your code and a brief description of how you solved the problem.

You are charged with a sequence classification task. You will receive simulated DNA sequences of length 200 bp. Each sequence is associated with a unique class of enhancers (labelled: 0, 1, 2 or 3). The sequences contain different combinations of motifs that make them distinguishable. You will receive three files: A **training set file** with 10,000 and a **validation set file** with 1,000 sequences and associated labels (two-column, tab-separated format: class sequence) as well as a **test set file** with 1,000 sequences (one column, sequences only).
Your task is to train a classifier that accurately classifies the test sequences. Use the training and validation files for training, then classify the test sequences. Send back the test file with class labels associated with the sequences and we will compare them against our ground truth to determine your test accuracy.

You may use any computational framework and libraries you choose to. If you do not have anything set up locally, consider using google colab (https://colab.research.google.com/notebooks/welcome.ipynb) which gives you access to a free environment to run python code with tensorflow and keras pre-installed (pytorch can be installed).

Tipp: load local files inot colab:

```
# for local file upload
from google.colab import files
uploaded = files.upload()
```

Tipps:

- You may choose any classifier type you see fit.
- Pay attention to how you prepare your data for training

Files:

Pwm_seq_200bp_train_set_10k.txt - training sequences with labels (two-column tab-separated)

pwm_seq_200bp_valid_set.txt - validation sequences with labels (two-column tab-separated)

pwm_seq_200bp_test_set_TOSEND.txt - Test set (one column sequences only). Please classify the sequences using your classifier and send us the results as a two-column tab-separated file.