Project-2:

Points: 200

Due Date: 12/03, 23:59

one-on-one interview: Week of 12/02

Description:

1. Assuming the output of project-1 is 20 files (10 DNA accessible files and 10 DNA inaccessible files or negative files). OR two big files.
2. There is no particular way to write a program, the logic differs from person to person, so take my advice and you are free to implement in any way.
3. You have to assign labels to the data, 1 for accessible and 0 for not-accessible. For example:
   ATTTAGG…CCCC 1
   TTTGGCC…AAAA 1
   ATTGGCC…CCAA 0
4. Combine the 20 files into one file. With labels in it. OR have two files with labels in it.
5. To practice, create a smaller file from the file generated in step-3/4. Lets say 1000 data points instead of the whole data set.
6. Convert the file into one-hot-encoding, preferably an numpy array. You can separate the labels and the data. For Example data[0] is data i.e. ATGGCC and a[1] is a label 0/1
7. Convert the one-hot-encoding into tensor format.
8. Pass the one-hot-encoding and the labels to a data_class (We talked about it in the class)
9. Create a custom data loader by calling the step-8 data_class. Basically you have to create a custom data loader for PyTorch
10. Keep 80% data for training and 20% data for testing. i.e. for example 800 data points for training and 200 data points for testing.
11. At first try it with Alexnet with 1D CNN.(refer to the program given in Moodle, remember it is in 2D)
12. The number of input channels will be 4.
13. If the program is working perfectly for 1D-CNN, test it with 1000 data points, Use two models 1) Alexnet 2) NiN
14. The Alexnet and NiN programs are in 2D.
15. Run a model for 5 epochs.
16. Use two different hyperparameters on each model. (Hint: Use 256 as hidden layers, faster execution)
17. **Total number of runs will be 4**:- Alexnet with two hyperparameters + NiN with two hyperparameters

18. Graphs: Plot an accuracy Vs epoch graph. All the four experiment results should be in one plot.
19. Confusion Matrix: Create a confusion matrix for the four experiments.
20. **IMPORTANT**: use a minimum of 10000 data points for each of the 4 experiments. You may not have to use all the data points, due to computing constraints. If you want good for you.
21. Do not forget to make *shuffle* = *True* for the training dataset.

Rubric:

Ability to create a custom data loader with one hot encoded genomic data from project-1. Ability to train the model (Alexnet and NiN) with a custom data loader. [150 points]. Plots [25] and confusion matrix [25 points]. No points will be awarded if you do not attend the one-to-one interview at the designated time.

Unable to upload a subset of the data and all the program to Moodle will result in 50% taken off. This is not a team project, if I find similarity I will not hesitate to give a zero.

We will talk more about this in the next class. If you have any doubts please ask me during the class.