**Normal/Tumor and Cancer Typing: Programming Assignment**

**Classifying Normal Samples and Tumors Samples from RNAseq profiles**

**and**

**Classifying Types of Tumors from RNAseq profiles**

**Due November 14th, 2020 at Midnight.**

In this assignment we are going to work with both the **Normal/Tumor** datasets and the **Cancer Type** datasets discussed in class. The data and the starting models for this assignment can be

Found at the following URL <https://bit.ly/3efMC6O> also same directory at

<https://anl.box.com/s/wzjkajk3jcqkxspg0gra61ptmekrdgmb> .

There are two tarballs **NT.tar.gz** and **TC.tar.gz** that contain starter Keras models and datasets for the initial part of the assignment which is to get running a Normal/Tumor classifier (NT3) and a basic Type classifier (TC1). Use the command gunzip NT.tar.gz to get the “tar” file and the command “tar -xvf NT.tar” to unpack the file. (repeat for the TC.tar.gz file). You should end up with a NT3.py and two .csv files which will be read for training and testing. You will need to run these in an environment that has tensorflow and Keras installed and you might need some other dependencies such as pandas and numpy. The starter codes train the models but you will need to work out how to run the trained models on additional input data.

Once you have those working, we would like you to do the following:

1. Use the Type Classifier to classify the Tumor samples from the NT data set. Using the type IDs from the Type classifier output for the tumor samples create labelled tumor data that you can use to add to the training/testing data for the Type Classifier. Retrain the Type classifier on the expanded set of data and report on the results.
2. Using the “normal” samples from the NT dataset add a “normal” type to the Type Classifier training/testing data and retrain the model. Report on the results (accuracy, confusion matrix etc.).
3. Using the newly trained Type Classifier that now has the new class of “normal” as one of the categories re-run the classifier on all of the NT data and determine how many of the Normal samples are misclassified? What could be a possible explanation of this misclassification? Note that you will need to do some strict partitioning of training and testing data to not run the model on the same data it was trained on.
4. Build a new type classifier that attempts to assign the tissue type to the sample independently of whether the sample is Normal or Tumor. You might want to use the metadata files in the directory to provide information about which samples are of which type.
5. Using ideas discussed in class, use an unsupervised learning method to cluster the Cancer Type sample data and assign your own “types” to the clusters and use those labels to train a type classifier. Describe your approach and report on the results.
6. Using any method you like, attempt to produce a model that can generate new data samples labelled as Normal and Tumor and run those newly generated samples through the classifier developed in Step 3. Report on the results.

For each of these problems, you might want to use 5-fold cross validation and display the confusion matrix, or other evaluation metrics in addition to the accuracy achieved.

There are two input files for each problem, one containing all the genes (60,483 features) and one containing just the protein coding genes (19,560 features). You will likely want to use some kind of normalization for these features. Note the datasets in the NT.tar.gz and TC.tar.gz have been normalized as we mentioned in class.

The other input files are **nt.coding.csv**, **nt.all.csv** and **type.coding.csv**, **type.all.csv**

There are two other files in that directory which map Ensembl IDs to Gene Symbol and Gene Types. You might find those useful when developing gene signatures. There are 18 types of cancer represented, a subset of the GDC types.

For each part please turn in your code (a python notebook is a reasonable way, but a python script is also fine), turn in output from the program (text or graphics, graphics preferred (screen shot, log file or notebook). And a short (1 paragraph write up for each part and section, explaining what you did and your critique of the results, comments on problems or difficulties and possible future approaches that might do better).