CSCI3230 Fundamentals of Artificial Intelligence
Neural Network Project
Drug Molecular Toxicity Prediction
Due date: 23:59:59 (GMT +08:00), 15th November 2020

1. Project Specification

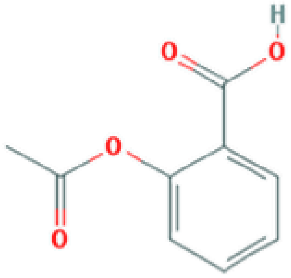
Almost all people are exposed to different chemicals during their lifetimes through different sources including food, household cleaning products and medicines. However, in some cases, these chemicals can be toxic and affect human health. As a matter of fact, over 30% of drugs have failed in human clinical trials because they are determined to be toxic despite promising pre-clinical studies in animal models. Consider real-world clinical trials for assessing drugs are extremely time-consuming, it is ideal if a computational drug molecular toxicity assessment method can be developed to quickly test whether certain chemicals have the potential to disrupt processes in the human body of great concern to human health.

Deep neural network has become a hot research topic in machine learning in recent years. Compared to other methods, deep learning has shown its advantages in handling large amount of data and achieving better performance. In this individual project, you will be given a dataset of drug molecules with their SMILES expressions (which will be explained later) and the binary labels indicating whether one drug molecule is toxic or not. You are going to develop a Deep Neural Network which can learn useful patterns from the data provided and predict the toxicity of a new list of molecules based on learned knowledge using TensorFlow package.

TensorFlow is an open-source software library designed for numerical computation using data flow graph and is widely used in deep learning community. It has many convenient APIs for implementing deep neural networks, more details would be introduced in our tutorial.

2. SMILES Expression

Simplified Molecular-Input Line-Entry System (SMILES) is a linear representation for molecular structure using 1D ASCII strings. For example, aspirin, a commonly used drug in daily life, its 2D structure is



and its SMILES is

CC(=O)OC1=CC=CC=C1C(=O)O

The one hot format of SMILES is a 2D {0,1} matrix, where each column represents a symbol in the SMILES notation of the current molecule, and each row is one ASCII character appeared in the dataset's SMILES dictionary. The size of the 2D matrices is the size of the dataset's SMILES dictionary * the length of the longest molecule SMILES, which means we have zeros padded after short molecule SMILES. For a SMILES notation, one at row i, col j means the jth symbol of that SMILES is the ith character in the dictionary. The one-hot example for aspirin is:

	C	C	(	=	O	)	O	C	1	=	C	C	=	C	C	=	C	1	C	(	=	O	)	O
C	■	■						■			■	■		■	■		■		■					
(			■																	■				
=				■						■			■			■					■			
O					■		■															■		■
)						■																	■	
1									■									■						

3. Dataset

The dataset provided is about the toxicity of some small molecules. We provide two folders for you, one is the training data SR-ARE-train (7167 samples), and the other one is testing data SR-ARE-test (234 samples). There are three files in each folder:

File	**Type**	**Description**
names_smiles.txt	String	A text file, each line contains a drug molecule's name and its SMILES expression, separated by comma (,)
names_labels.txt	String	A text file, each line contains a drug molecule's name and its toxicity label, where 0 means non-toxic and 1 means toxic, separated by comma (,)
names_onehots.pickle	Numeric	A pickle file which can be loaded by pickle package, storing a python dictionary containing (1) a list of names of the molecules, and (2) an numpy array of the one-hot representations for the SMILES expressions of drug molecules

The source data are *names_smiles.txt* and *names_labels.txt* files. Your neural network is supposed to learn from (SMILES, label) records, and be able to predict label from SMILES in the end. The *names_onehots.pickle* file is derived from *names_smiles.txt*, storing one-hot representations of SMILES expressions of drug molecules. Providing *names_onehots.pickle* is to ease the burden of you on data preprocessing and focus on neural network construction. You are not constrained on how to use these data files as long as they are the only training data. You can build Convolutional Neural Networks or other kinds of neural networks as you like.

The molecules in SR-ARE-train and SR-ARE-test do not overlap with each other. The data we use to mark your model is in a folder named SR-ARE-score (hundreds of samples), and you have no access to it. There are two files in the SR-ARE-score folder: *names_smiles.txt* and *names_onehots.pickle*. The format of these two files are exactly the same with those for training and testing, but the molecules are new.

4. Assignment Requirement

1) This is an individual project.
2) Data
   a) SR-ARE-train
   b) SR-ARE-test
   c) SR-ARE-score (not accessible)

   You can train your model on the SR-ARE-train and test its performance on SR-ARE-test, or you can train your model with both.
3) Features

   You can either use the one hot of SMILES as features for molecules, or directly use the SMILES notation as you wish.

4) Model
You can use any deep neural network architecture in this assignment to achieve good performance, e.g. convolutional neural network (CNN), recurrent neural network (RNN), graph neural network (GNN), etc.

5) Prediction task
You are going to predict toxicity of the molecules based on their structures. The output of your model is the {0,1} label indicating that the current molecule is toxic or not, where 0 means non-toxic and 1 means toxic.

6) Output and Marking
Your model will be tested on the SR-ARE-score dataset. Your submitted folder will be extracted and put alongside the SR-ARE-score folder, and it means that you must use the relative path "../SR-ARE-score/" in your submitted *test.py* file to access the marking data.
In your *test.py* file, you need to first restore your model parameters and then test your model on the marking data ("../SR-ARE-score/"), and you should output your predictions into a file named *labels.txt*, which should be in the same directory as *test.py*. Each line in *labels.txt* file is the {0,1} toxicity label for the corresponding sample.
Your final score of this assignment will depend on the Balanced Accuracy among your predictions and the true labels:

$$balance\_accuracy = \frac{1}{2} \times \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right)$$

7) Deep learning library
TensorFlow 1.5.0 and NumPy only. Python's native packages are allowed. Please do not use other libraries. Otherwise, you will get zero marks for this assignment.

8) Programming language
The only supported language for this assignment is python3. We do not accept python2 program.

5. Submission Requirement

1) Submission list
   a) Source file for training. Name it as *train.py*
   b) Source file for recovering your network model. Name it as *test.py*
   c) TensorFlow generated files, which store your trained model.
   d) Any other files that help your programs to work, such as preprocessing files, format-converting files.

2) Submission packaging
Put everything in the submission list above into a folder, name the folder as your student id, zip the folder WITHOUT encryption, also name the zip file as *your_student_id.zip*. Submit the zip file to our submission system.

3) DO NOT add any of the data in training and test folder in your zip file, because we will grade your model solely on the score folder, which is already put in the server.

6. Important Points

To make this project fair and meaningful, there are some other points you MUST follows:

1) The time limits to run your *test.py* is 60s
2) The size of the whole zip file should be less than 200 MB
3) Plagiarism will be SERIOUSLY punished (ZERO mark plus reporting to department)

7. Late Submission

No. of Days Late	Marks Deduction
1	10%
2	30%
3	60%
4 or above	100%