Homework 7

Due 11/27/2019

Background:

In this assignment, we will work with a dataset called MUV (maximum unbiased validation), which is commonly used as a benchmark for virtual screening. Each column represents a target (receptor protein) and each row represents a ligand (small drug-like molecule). Our task here is to predict the response (active/inactive) using only the features from ligands, which is the ligand-based virtual screening we talked about during the class.
In the CSV file, active response is marked as 1; inactives is 0, and empty cell means no data.

1. Feature extraction and data processing

We'll use two different features for this assignment: fingerprints (Circular Fingerprints) and descriptors (Mordred). Due to the long processing time, I have already done the process for you guys, but in case you want to try extracting features by yourself, here are the steps to follow:

First, install the packages (rdkit and mordred). I highly recommend doing this on local machine because of the installation of rdkit requires conda, which is an environment management software for python. However, if you are doing the homework on Google colab, try:

```
!wget -c
https://repo.continuum.io/miniconda/Miniconda3-latest-Linux-x86_64.s
h
!chmod +x Miniconda3-latest-Linux-x86_64.sh
!time bash ./Miniconda3-latest-Linux-x86_64.sh -b -f -p /usr/local
!time conda install -q -y -c conda-forge rdkit
%matplotlib inline
import matplotlib.pyplot as plt
import sys
import os
sys.path.append('/usr/local/lib/python3.7/site-packages/')
!pip install mordred
```

Then, follow the instructions (Circular Fingerprint, Mordred) to extract features from the SMILES strings in the dataset. There are several parameters for each featurization function, and feel free to read through the document and find out about the usages of these parameters. For this assignment, we will stick with the following 4 features:

```
mol = Chem.MolFromSmiles('c1ccccc1')
mordred_calc = Calculator(descriptors, ignore_3D=True)

# The 4 features required in this assignment
fp256 = AllChem.GetMorganFingerprintAsBitVect(mol, radius=2,
nBits=512).ToBitString()
fp1024 = AllChem.GetMorganFingerprintAsBitVect(mol, radius=2,
nBits=1024).ToBitString()
fp4096 = AllChem.GetMorganFingerprintAsBitVect(mol, radius=2,
nBits=4096).ToBitString()
mordred = mordred_calc(mol)
```

Question 1.1
Before you continue, could you guess which feature(s) will perform the best, based on the documents? Please write down the reasons behind your guess. There is no "correct" answer to this question.

Question 1.2
Plot the histogram(s)/bar plot(s) of the number of entries, the number of actives and the ratios of actives for each target. You can either put all the statistics in the same figure, or three seperate ones.

Question 1.3
Plot the histogram(s) of the number of entries, the number of actives and the ratios of actives for each ligand. You can either put all the statistics in the same figure, or three seperate ones.

2. Model comparison

Now pick the column (target) with the most number of entries and build 4 deep learning models for this single-task binary classification problem, using the 4 different features you have obtained in part 1. Remember to scale/normalize the input features before training.

You can use the same network configuration (except for the input layer of course) for all the models, and train them with enough epochs to make sure the training loss stabilizes. Report the ROC AUC of these 4 features.

Question 2.1
Based on your research and understanding, why do you think normalization is often required for neural networks? And which feature(s) is more likely to yield bad performance without normalization?

Question 2.2
Based on the plots from part 1, why do you think ROC AUC is the better metric for this classification problem, instead of accuracy?

Question 2.3
Which one of the 4 features has the best ROC AUC according to your results?

Question 2.4
If all the entries in the all columns/targets are used during the training, would you expect the prediction performance (of the specific target that you have chosen) to increase or decrease, and why? Again there is no "correct" answer here.

Now use all the columns/targets in the MUV dataset, and the best features you find from the previous comparison, to construct a neural network for this binary classification problem. An easy way to do this is to construct a network with two sets of inputs: the first one being the features of the ligands, and the second one is the one-hot encoding of the receptor.

Question 2.5
What is the ROC AUC of all the labels and the one specific receptor/target that you have picked in the previous run? Compare your results to the some of the latest results from MoleculeNet.