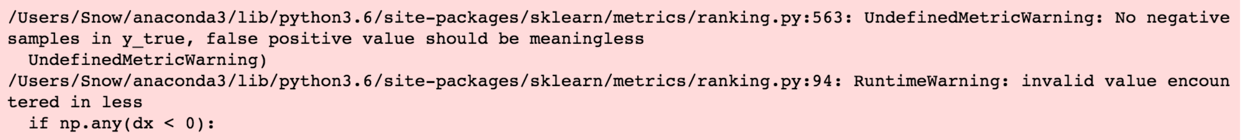
Snow Naing
BMI203
Final Project: Neural Net - Distinguishing binding sites of a transcription factor, RAP1

I adapted codes for neural net, learning from https://www.youtube.com/watch?v=pHMzNW8Agq4&t=5s. As for identity matrix, I tested in the test folder changing the input, hidden and output layer size as 8,3,8 respectively.

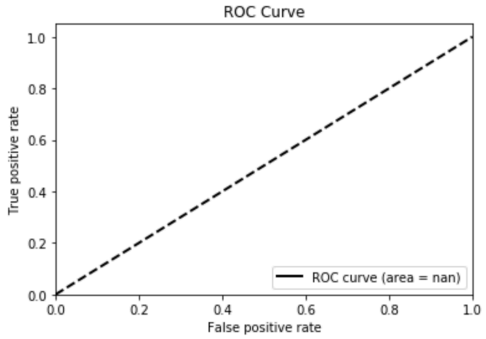
As for the design of my neural net, you can find the code in "neuralnet.py" file with the comments. And, the input of the file is in "io.py". The basic idea is that for input, I am changing the DNA string to a binary 4-bit vector, which makes the input into my neural net 68 nodes (17bp x 4-bit per bp). I also got rid of the positive binding site from the whole genome and the remaining sequences are used as negative test data. One thing I'd like to do if I have more time is to train my data with every possible 17bp in the genome (neg file). For now, just every 17bp stretch of the genome will have to do. As for the gradient descent, I follow the math as explained in the aforementioned youtube videos and added the bias matrices to it. I didn't play with lambda values but varying this value will affect how my descent gradient works and will impact on the scoring.

For the testing data, I tested with 100 iterations and used 200 hidden layers. The scoring for the testing data can be found in "yhat_test_snow.txt".

As for the performance of my neural net, I am having some problems getting ROC values. The error can be seen in the figure below. It seems the tpr values output are all nan. I am not quite sure what is causing this problem because when there are values in my true output vector.



```
/Users/Snow/anaconda3/lib/python3.6/site-packages/sklearn/metrics/ranking.py:563: UndefinedMetricWarning: No negative
samples in y_true, false positive value should be meaningless
  UndefinedMetricWarning)
/Users/Snow/anaconda3/lib/python3.6/site-packages/sklearn/metrics/ranking.py:94: RuntimeWarning: invalid value encoun
tered in less
  if np.any(dx < 0):
```



Below are the sample scores output using my neural net.

```
Actual binding site ACATCCGTGCACCTCCG [0.99999741]
Testing site AAAACCAAACACCTGAA       [1.34106696e-07]
```

Scanning through all the scores for testing file, my neural net performed decently with all the values very small. Will need to actually have the ROC working to grade the performance.