Question 5

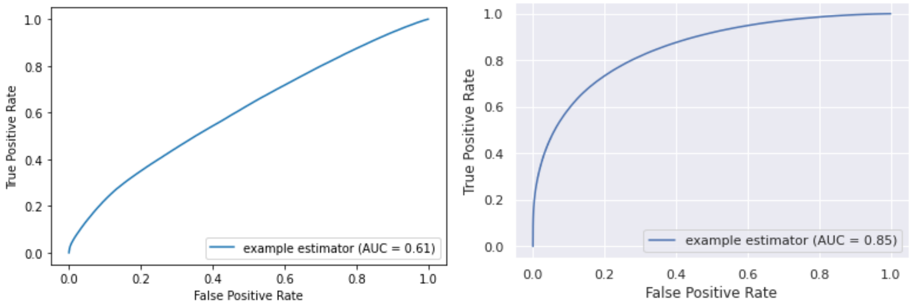

Figure 1: Untrained model(left) Trained model(right)

1. From figure 1, we can see after training, the models have improved its performance since:
    a. The AUC score has improved from 0.61 to 0.85 and the ROC curve is much closer to the top-left corner.
    b. A 0.85 AUC score indication there is 85% chance that the model will be able to distinguish better the positive and negative class.
2. For the normalization, the following steps is used.
    a. Convert the four vectors into a matrix using np.matrix() with matrix dimension 4*19.
    b. Normalize it using "normalize" function from sklearn.preprocessing;
    c. Give the limitation of the space, the table show the 1st column normalized value.

A	C	G	T
0.09529025	0.31872946	0.08324206	0.50273823

3. The following steps covers question 3-5.
    a. In the Basset model, create the activation object right after the 1st convolution layer (after batch norm and Relu activation).
    b. For each activation map, use 0.5*np.amax() to find the activation threshold;
    c. For each filter, use the F.unfold function to find the corresponding sequences and reshape them as [number batch*batch size(64),length(600),19,4]
        i. In each batch, among all the sequences, find the indices where the filter is activated and store the corresponding sequence in a list.
    d. For each filter, convert the list to stacked array and count the pair-bases over the sequence using np.sum() and reshape each PWM to [4*19].
    e. Calculate the Pearson Correlation coefficient between the flattened normalized 300 PWMs with the flattened normalized CTCF. (Note: given the computation limitation, the result is based on first 80 batches from test set.)

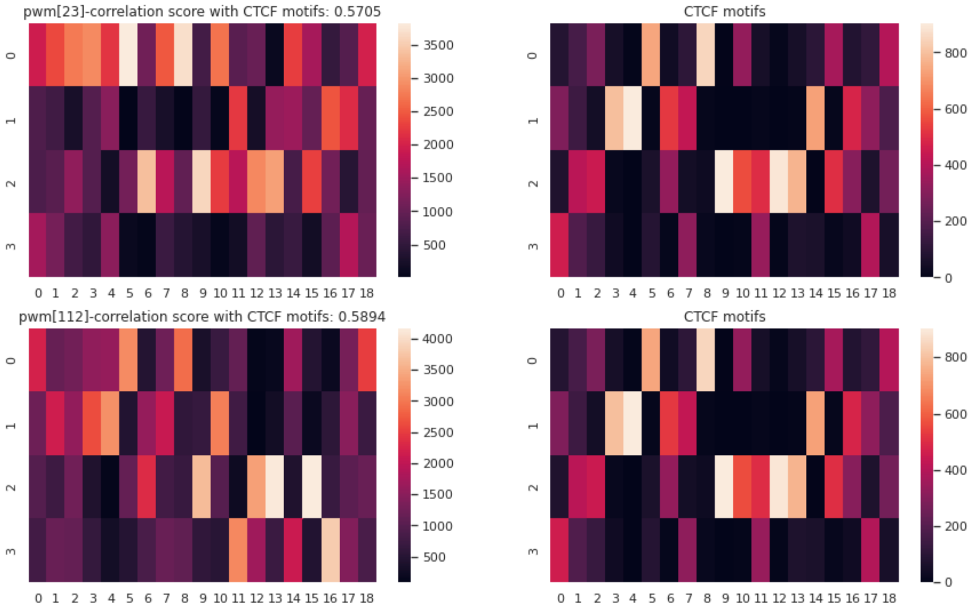

Figure 2 Heatmap for top two most similar PWMs to the CTCF motif