ECS171 HW2 Report

Jiahui Li : 915544392

Instruction to Run the Code

I split the code file for one file per question (problem1.py, problem2.py,etc) and put some shared code, eg.read the data, in one file (basicfunction.py).

Since we are using the same model and same sample to calculate the uncertainty in problem 7(bonus) as problem 6, and no more data and code we need, so I printed the probability of all the classes in problem 6 in code for the calculation of problem 7 and didn't make a problem7.py file.

The following report shows the results, plots and tables for each problem.

Problem 1

(a). We use Isolation Forest and LOF to detect the outliers. When we set the comtamination as "auto", there are 58 outliers by Isolation Forest method and there are 76 outliers by LOF method. So, there are some outliers on the dataset.

(b). In order to compare if LOF method agree with Isolation Forest, we set the contamination of LOF be the same as that of Isolation Forest, which is 0.039. Among 58 outliers detected by IF, there are 42 are detected by LOF too, about 72.0%. So these two methods are different since they based on different assumptions and having different detecting ways. By the way, some parameters such like the random state in IF and the n_neighbors in LOF may effect too.

(c). Assumptions:

Isolation Forest: (i) the number of splittings required to isolate a sample is equivalent to the path length from the root node to the terminating node. (ii) When the samples are abnormal, the number of splittings to isolate it will be small, and the path length will be short.

LOF: (i). local density is estimated by the distance of k-nearest neighbors. (ii). When the samples are abnormal, its local density will be lower than the local densities of its neighbor.

We choose Isolation Forest method to remove the outliers. The new, revised dataset has 1426 samples, and only has 9 different classes since all the 5 samples of "ERL" class have been dropped.

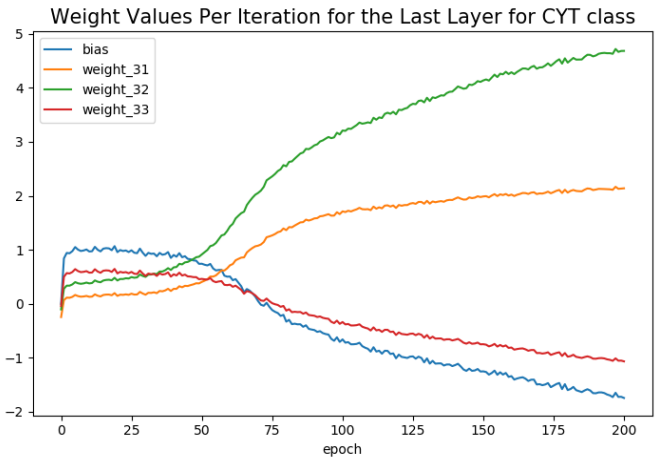
problem 2

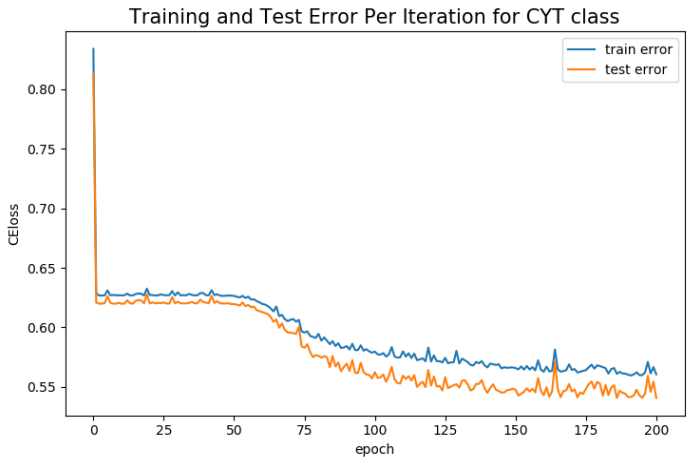
Since we need to use stochastic gradient descent, the batch_size = 1, and the epoch we choose is 200.

We plot 2 plots for the most popular class "CYT":

(i) Weight values per epoch for the last layer. Since "CYT" is the third node in my outpur layer, so the first index of weight is 3.

(ii) Training and test error epoch. We choose CEloss as error.





problem 3

Train with all 1484 data, 200 epochs, the training error is:

1-accuracy: 0.4609

CEloss: 1.1264

The "CYT" node in my output layer is the third one.

$$z^{(4)} = [z_1^{(4)} z_2^{(4)} \dots z_{10}^{(4)}] = [a_1^{(3)} a_2^{(3)} a_3^{(3)}] \begin{bmatrix} W_{11}^{(3)} & W_{21}^{(3)} & \dots & W_{10\_1}^{(3)} \\ W_{12}^{(3)} & W_{22}^{(3)} & \dots & W_{10\_2}^{(3)} \\ W_{12}^{(3)} & W_{21}^{(3)} & \dots & W_{10\_3}^{(3)} \end{bmatrix} + [b_1^{(3)} b_2^{(3)} \dots b_{10}^{(3)}]$$

$$a_3^{(4)} = \frac{exp(z_3^{(4)})}{exp(z_1^{(4)}) + exp(z_2^{(4)}) + exp(z_3^{(4)}) + \dots + exp(z_{10}^{(4)})} = \frac{exp(z_3^{(4)})}{\sum_{i=1}^{10} exp(z_i^{(4)})}$$

where, W is

```
[[ 4.220012    5.34821     5.074723   -6.1149693  -1.4222914  -1.9950147
  -2.166825    0.8284938  -0.72907996 -2.6233861 ]
 [ 2.5170481  -4.8777194  -2.4766107   3.9783995   2.5090015   1.8473928
  -3.4725847  -1.3675933   0.65391946  0.7910297 ]
 [-2.4195635  -0.15650998 -1.148507    1.0065137  -1.138264    0.46519175
   4.829911    0.60405207 -1.0910491  -0.2795592 ]]
```

b is

```
[-0.49482414  0.7680428   0.9999884  -1.1132219  -0.26453513  0.0874479
  1.123424    0.02331839  0.02008045 -1.1497556 ]
```

problem 4

Since the first sample is “MIT” node, so we choose the “MIT” node in output layer, which is the first node in my ANN. And from the second layers back to first layers, we also choose the first node in the second layers.

We initialize the eight weights that we need calculate to 1, and initialize the other weights and bias to 0.

From the code, we can get:

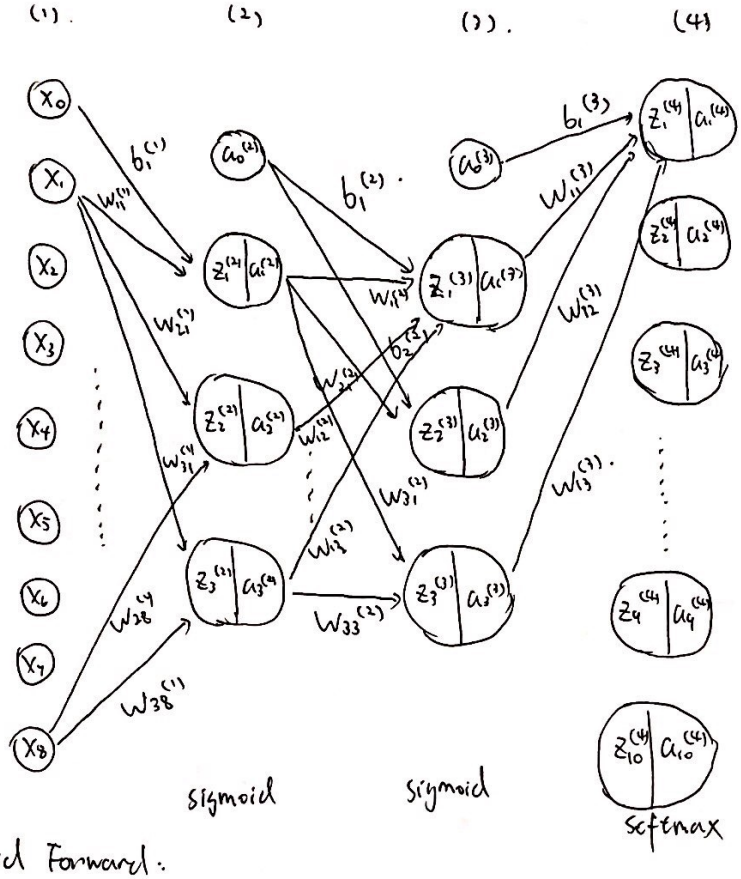
Bias from output to second hidden: 1.0032587051391602

Weights from output to second hidden: [1.0030115, 1.0016294, 1.0016294]

Bias from second hidden to first hidden: 1.0002284049987793

Weights from second hidden to first hidden: [1.0001142, 1.0001142, 1.0001142]

Then, we do it by hand:



The weights we want to update is $b_1^{(3)}$, $W_{11}^{(3)}$, $W_{12}^{(3)}$, $W_{13}^{(3)}$ and $b_1^{(2)}$, $W_{11}^{(2)}$, $W_{12}^{(2)}$, $W_{13}^{(2)}$.

Feed Forward:

$$z^{(2)} = [z_1^{(2)}\ z_2^{(2)}\ z_3^{(2)}] = [x_1\ x_2 \cdots x_8] \begin{bmatrix} W_{11}^{(1)} & W_{21}^{(1)} & W_{31}^{(1)} \\ W_{12}^{(1)} & W_{22}^{(1)} & W_{32}^{(1)} \\ \vdots & \vdots & \vdots \\ W_{18}^{(1)} & W_{28}^{(1)} & W_{38}^{(1)} \end{bmatrix} + [b_1^{(1)}\ b_2^{(1)}\ b_3^{(1)}] = [0\ 0\ 0]$$

$$a^{(2)} = [a_1^{(2)}\ a_2^{(2)}\ a_3^{(2)}] = \text{sigmoid}([z_1^{(2)}\ z_2^{(2)}\ z_3^{(2)}]) = \left[\frac{1}{1+e^0}\ \frac{1}{1+e^0}\ \frac{1}{1+e^0}\right] = \left[\frac{1}{2}\ \frac{1}{2}\ \frac{1}{2}\right]$$

$$z^{(3)} = [z_1^{(3)}\ z_2^{(3)}\ z_3^{(3)}] = [a_1^{(2)}\ a_2^{(2)}\ a_3^{(2)}] \begin{bmatrix} W_{11}^{(2)} & W_{21}^{(2)} & W_{31}^{(2)} \\ W_{12}^{(2)} & W_{22}^{(2)} & W_{32}^{(2)} \\ W_{13}^{(2)} & W_{23}^{(2)} & W_{33}^{(2)} \end{bmatrix} + [b_1^{(2)}\ b_2^{(2)}\ b_3^{(2)}]$$

$$= \left[\frac{1}{2}\ \frac{1}{2}\ \frac{1}{2}\right] \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} + [1\ 0\ 0] = \left[\frac{3}{2}\ 0\ 0\right] + [1\ 0\ 0] = \left[\frac{5}{2}\ 0\ 0\right]$$

$$a^{(3)} = [a_1^{(3)}\ a_2^{(3)}\ a_3^{(3)}] = \text{sigmoid}([z_1^{(3)}\ z_2^{(3)}\ z_3^{(3)}]) = \left[\frac{1}{1+e^{-2.5}}\ \frac{1}{1+e^0}\ \frac{1}{1+e^0}\right] = \left[\frac{1}{1+e^{-2.5}}\ \frac{1}{2}\ \frac{1}{2}\right]$$

$$z^{(4)} = [z_1^{(4)}\ z_2^{(4)} \cdots z_{10}^{(4)}] = [a_1^{(3)}\ a_2^{(3)}\ a_3^{(3)}] \begin{bmatrix} W_{11}^{(3)} & W_{21}^{(3)} \cdots W_{q1}^{(3)} & W_{10,1}^{(3)} \\ W_{12}^{(3)} & W_{22}^{(3)} \cdots W_{q2}^{(3)} & W_{10,2}^{(3)} \\ W_{13}^{(3)} & W_{23}^{(3)} \cdots W_{q3}^{(3)} & W_{10,3}^{(3)} \end{bmatrix} + [b_1^{(3)}\ b_2^{(3)} \cdots b_q^{(3)}\ b_{10}^{(3)}]$$

$$= \left[\frac{1}{1+e^{-2.5}}\ \frac{1}{2}\ \frac{1}{2}\right] \begin{bmatrix} 1 & 0 \cdots 0 \\ 1 & 0 \cdots 0 \\ 1 & 0 \cdots 0 \end{bmatrix} + [1\ 0 \cdots 0] = \left[2 + \frac{1}{1+e^{-2.5}}, 0 \cdots 0\right]$$

$a^{(4)} = [a_1^{(4)}\ a_2^{(4)} \cdots a_{10}^{(4)}] = \text{softmax}([z_1^{(4)}\ z_2^{(4)} \cdots z_{10}^{(4)}]) = \left[\frac{e^{z_1^{(4)}}}{\sum_{j=1}^{10} e^{z_j^{(4)}}} \quad \frac{e^{z_2^{(4)}}}{\sum_{j=1}^{10} e^{z_j^{(4)}}} \cdots \frac{e^{z_{10}^{(4)}}}{\sum_{j=1}^{10} e^{z_j^{(4)}}}\right]$

$= \left[\frac{\exp(2+\frac{1}{1+e^{-2.5}})}{\exp(2+\frac{1}{1+e^{-2.5}})+9\times e^0},\ \frac{e^0}{\exp(2+\frac{1}{1+e^{-2.5}})+9\times e^0},\ \cdots\ \frac{e^0}{\exp(2+\frac{1}{1+e^{-2.5}})+9\times e^0}\right]$

The first sample is MIT class, $y = [y_1\ y_2 \cdots y_{10}] = [1\ 0 \cdots 0]$

Back propagation:

$E_1 = -y_1 \log(a_1^{(4)}) - (1-y_1)\log(1-a_1^{(4)})$

$\frac{\partial E_1}{\partial b_1^{(3)}} = \frac{\partial E_1}{\partial a_1^{(4)}} \frac{\partial a_1^{(4)}}{\partial z_1^{(4)}} \frac{\partial z_1^{(4)}}{\partial b_1^{(3)}} = \left[-\frac{y_1}{a_1^{(4)}} + \frac{(1-y_1)}{(1-a_1^{(4)})}\right]\left[\frac{e^{z_1^{(4)}}(\sum_{j=2}^{10} e^{z_j^{(4)}})}{(\sum_{j=1}^{10} e^{z_j^{(4)}})^2}\right] \cdot 1$

$= (-\frac{1}{a_1^{(4)}})\, a_1^{(4)}(1-a_1^{(4)}) = a_1^{(4)} - 1 = \frac{\exp(2+\frac{1}{1+e^{-2.5}})}{\exp(2+\frac{1}{1+e^{-2.5}})+9\times e^0} - 1 = 0.67413 - 1 \approx -0.32587$

$\frac{\partial E_1}{\partial w_{11}^{(3)}} = \frac{\partial E_1}{\partial a_1^{(4)}} \frac{\partial a_1^{(4)}}{\partial z_1^{(4)}} \frac{\partial z_1^{(4)}}{\partial w_{11}^{(3)}} = (a_1^{(4)} - 1)\, a_1^{(3)} = -0.32587 \times \frac{1}{1+e^{-2.5}} \approx -0.30115$

$\frac{\partial E_1}{\partial w_{12}^{(3)}} = \frac{\partial E_1}{\partial a_1^{(4)}} \frac{\partial a_1^{(4)}}{\partial z_1^{(4)}} \frac{\partial z_1^{(4)}}{\partial w_{12}^{(3)}} = (a_1^{(4)} - 1)\, a_2^{(3)} = -0.32587 \times \frac{1}{2} \approx -0.16294$

$\frac{\partial E_1}{\partial w_{13}^{(3)}} = \frac{\partial E_1}{\partial a_1^{(4)}} \frac{\partial a_1^{(4)}}{\partial z_1^{(4)}} \frac{\partial z_1^{(4)}}{\partial w_{13}^{(3)}} = (a_1^{(4)} - 1)\, a_3^{(3)} = -0.32587 \times \frac{1}{2} \approx -0.16294$

$E_i = -y_i \log(a_i^{(4)}) - (1-y_i)\log(1-a_i^{(4)})$, for $i = 2$ to 10, $y_i = 0$ in this sample.

$\frac{\partial E}{\partial a_1^{(3)}} = \frac{\partial E_1}{\partial a_1^{(3)}} + \frac{\partial E_2}{\partial a_1^{(3)}} + \frac{\partial E_3}{\partial a_1^{(3)}} + \cdots + \frac{\partial E_{10}}{\partial a_1^{(3)}}$

$= \frac{\partial E_1}{\partial a_1^{(4)}} \frac{\partial a_1^{(4)}}{\partial z_1^{(4)}} \frac{\partial z_1^{(4)}}{\partial a_1^{(3)}} + \sum_{k=2}^{10}\left[\frac{\partial E_k}{\partial a_k^{(4)}} \frac{\partial a_k^{(4)}}{\partial z_k^{(4)}} \frac{\partial z_k^{(4)}}{\partial a_1^{(3)}}\right]$

$= (a_1^{(4)} - 1)\, w_{11}^{(3)} + \sum_{k=2}^{10}\left[\frac{1}{(1-a_k^{(4)})} \cdot a_k^{(4)}(1-a_k^{(4)})\, w_{k1}^{(3)}\right]$

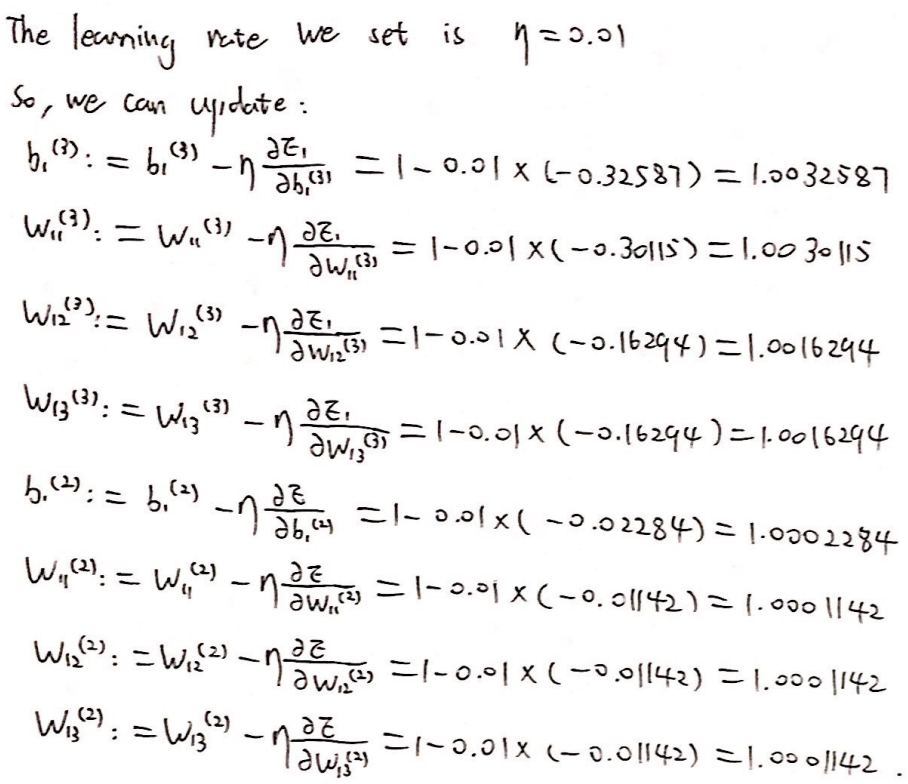
$= (a_1^{(4)} - 1)\, w_{11}^{(3)} + 0 = (a_1^{(4)} - 1)$ (since $w_{11}^{(3)} = 1$, $w_{k1}^{(3)} = 0$ for $k = 2$ to 10)

$\frac{\partial E}{\partial b_1^{(2)}} = \frac{\partial E}{\partial a_1^{(3)}} \frac{\partial a_1^{(3)}}{\partial z_1^{(3)}} \frac{\partial z_1^{(3)}}{\partial b_1^{(2)}} = (a_1^{(4)} - 1)\, a_1^{(3)}(1-a_1^{(3)}) \cdot 1 = -0.30115\left(1 - \frac{1}{1+e^{-2.5}}\right) \approx -0.02284$

$\frac{\partial E}{\partial w_{11}^{(2)}} = \frac{\partial E}{\partial a_1^{(3)}} \frac{\partial a_1^{(3)}}{\partial z_1^{(3)}} \frac{\partial z_1^{(3)}}{\partial w_{11}^{(2)}} = (a_1^{(4)} - 1)\, a_1^{(3)}(1-a_1^{(3)})\, a_1^{(2)} = -0.02284 \times \frac{1}{2} = -0.01142$

$\frac{\partial E}{\partial w_{12}^{(2)}} = \frac{\partial E}{\partial a_1^{(3)}} \frac{\partial a_1^{(3)}}{\partial z_1^{(3)}} \frac{\partial z_1^{(3)}}{\partial w_{12}^{(2)}} = (a_1^{(4)} - 1)\, a_1^{(3)}(1-a_1^{(3)})\, a_2^{(2)} = -0.02284 \times \frac{1}{2} = -0.01142$

$\frac{\partial E}{\partial w_{13}^{(2)}} = \frac{\partial E}{\partial a_1^{(3)}} \frac{\partial a_1^{(3)}}{\partial z_1^{(3)}} \frac{\partial z_1^{(3)}}{\partial w_{13}^{(2)}} = (a_1^{(4)} - 1)\, a_1^{(3)}(1-a_1^{(3)})\, a_3^{(2)} = -0.02284 \times \frac{1}{2} = -0.01142$



We can find that the the results made by hand and from the program are in agreement.

problem 5

We run each combination for 200 epochs and get the final testing error(CEloss).

	3	6	9	12
1	1.080121	1.085941	1.083406	1.088951
2	1.249665	1.078144	1.048840	1.049029
3	1.665847	1.528125	1.138714	1.320832

CEloss:

We can find that the combination of number of hidden layers = 2 and the number of nodes in each hidden layer = 9 has the least testing error, so this is the optimal configuration.

Except the combination of number of layers = 1 and the number of nodes = 3 (maybe because it has a very good initialization by random), we can find that for the same number of nodes, when the number of layers increase, the test error will decrease first and then increase. And we can also find that for the same number of layers, when the number of nodes increase, the test error will also decrease first and then increase. So we know that when the configuration is simple, maybe we are underfitting so the test error are large, and also , when the configuration is too complex, we will overfitting and the test error will be large too.

problem 6

We choose the optimal configuration, that number of hidden layers = 2 and number of nodes in each hidden layer = 9 and train the model with all the data (including training set and testing set) after removing outliers and predict the new sample.

It belong to “CYT” class, the probability is 0.3495.

problem 7

The quantitative measure of uncertainty for each classification is to calculate the entropy。 $H(X) = -\sum_{i=1}^{9}(p_i log p_i)$, if the entropy is larger, then the classification is high uncertainty. And if the entropy is low, it is less uncertain. We can easily see from the formula, if one of p_i is 1, and the others are 0, we are certain for this classification and the H(x) is 0. And if all the p_i are nearly equal, that is every p_i is near 1/9, then $H(x) = log(9)$, which is the largest, and under this situation, we are very uncertain about our results.

So the uncertainty of the new sample in problem 6 is 0.92.