ECS 171 HW2

Wenhao Su 915297212

Prob1

(a) there're outliers on the dataset.

#outliers for Isolation forest: 149

#outliers for LOF: 149

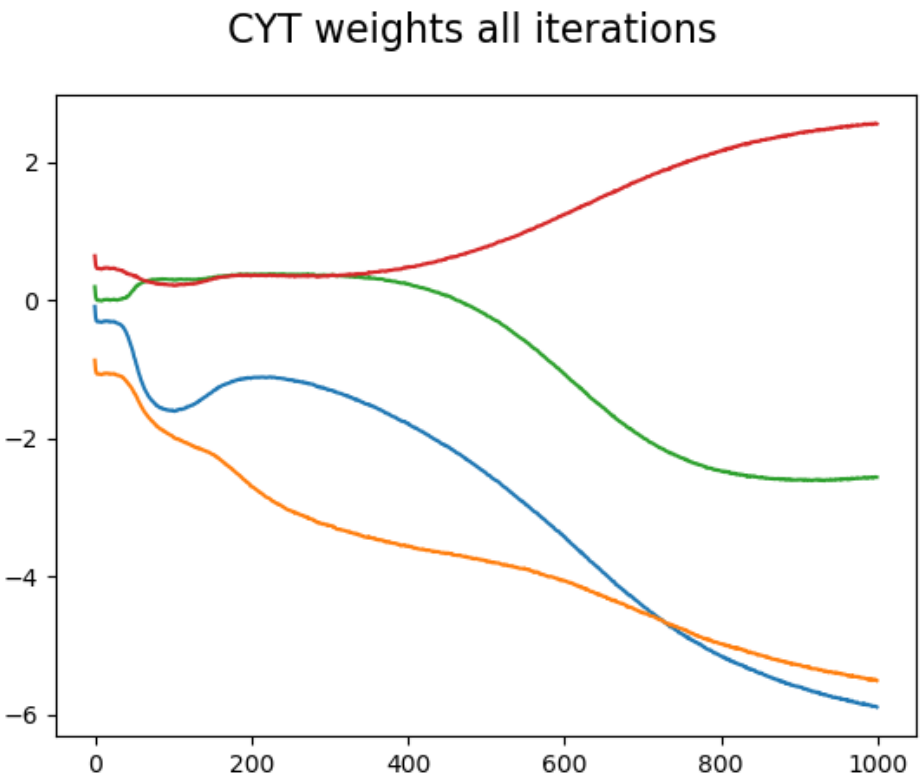
(b) #number of disagreement is: 174

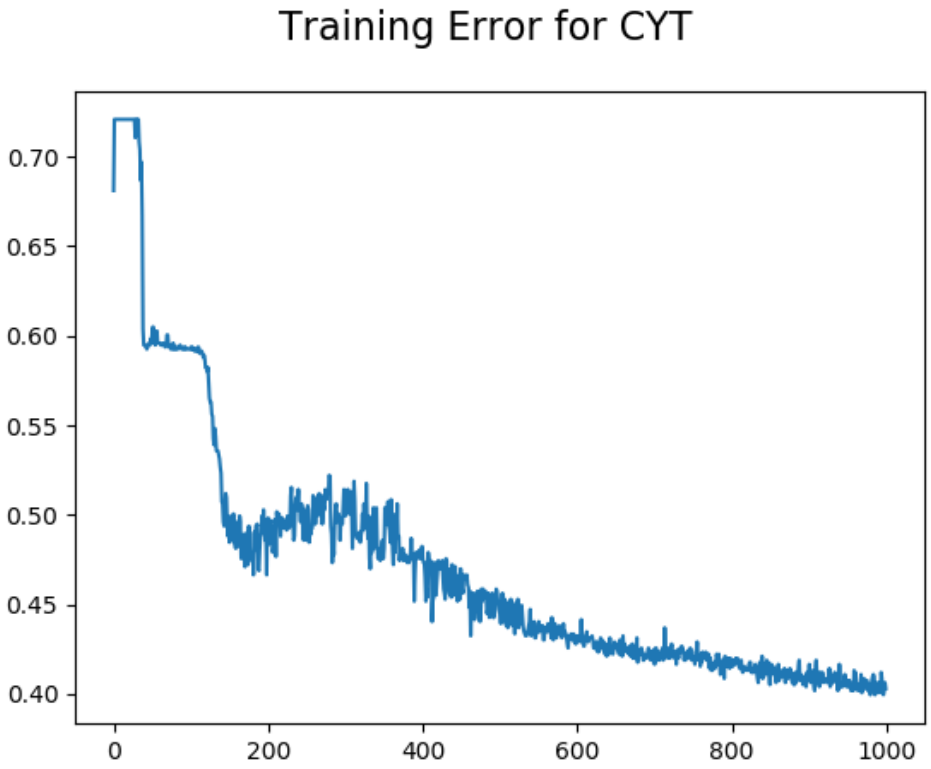
Two methods' difference include following reasons: Isolation forest doesn't need to calculate the distance between clusters or try to converge to the centroid. Instead it generates a binary tree and traverse the path. However, conversely, local outlier factor depends on calculating distance within k-th nearest neighbor based on the clustering model.

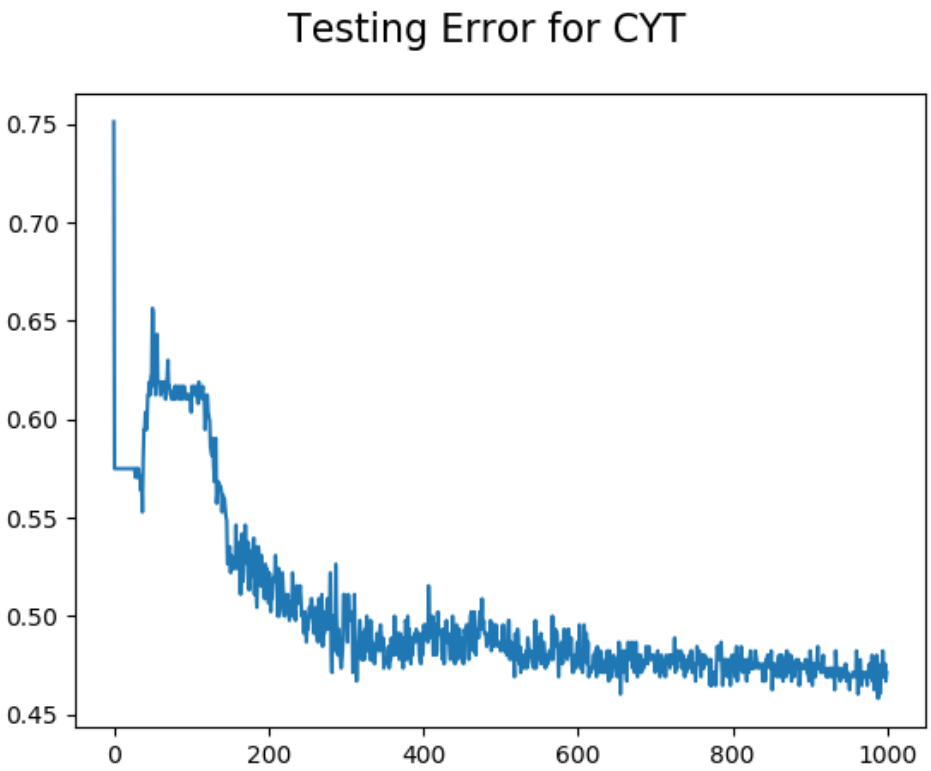
(c) Assumption of Isolation forest is based on the assumption that the anomaly data points are always rare and far from the center of normal clusters.

Assumption of local outlier factor is based on the pre-determined number of clusters used as a key parameter.

Prob2 **Run the script and plots will be saved to the current working directory (Red line is bias)**




Training Error for CYT
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0
200
400
600
800
1000


Testing Error for CYT
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0
200
400
600
800
1000

Prob3

Training Error for final model is: 0.3153558056899225

g(x) = $\frac{1}{1+e^{-(0.31113145 - 1.8810582*X1 - 0.04873516*X2 - 0.7547411*X3)}}$

w^t = [-1.8810582, -0.04873516, -0.7547411], bias = 0.31113145

Prob4

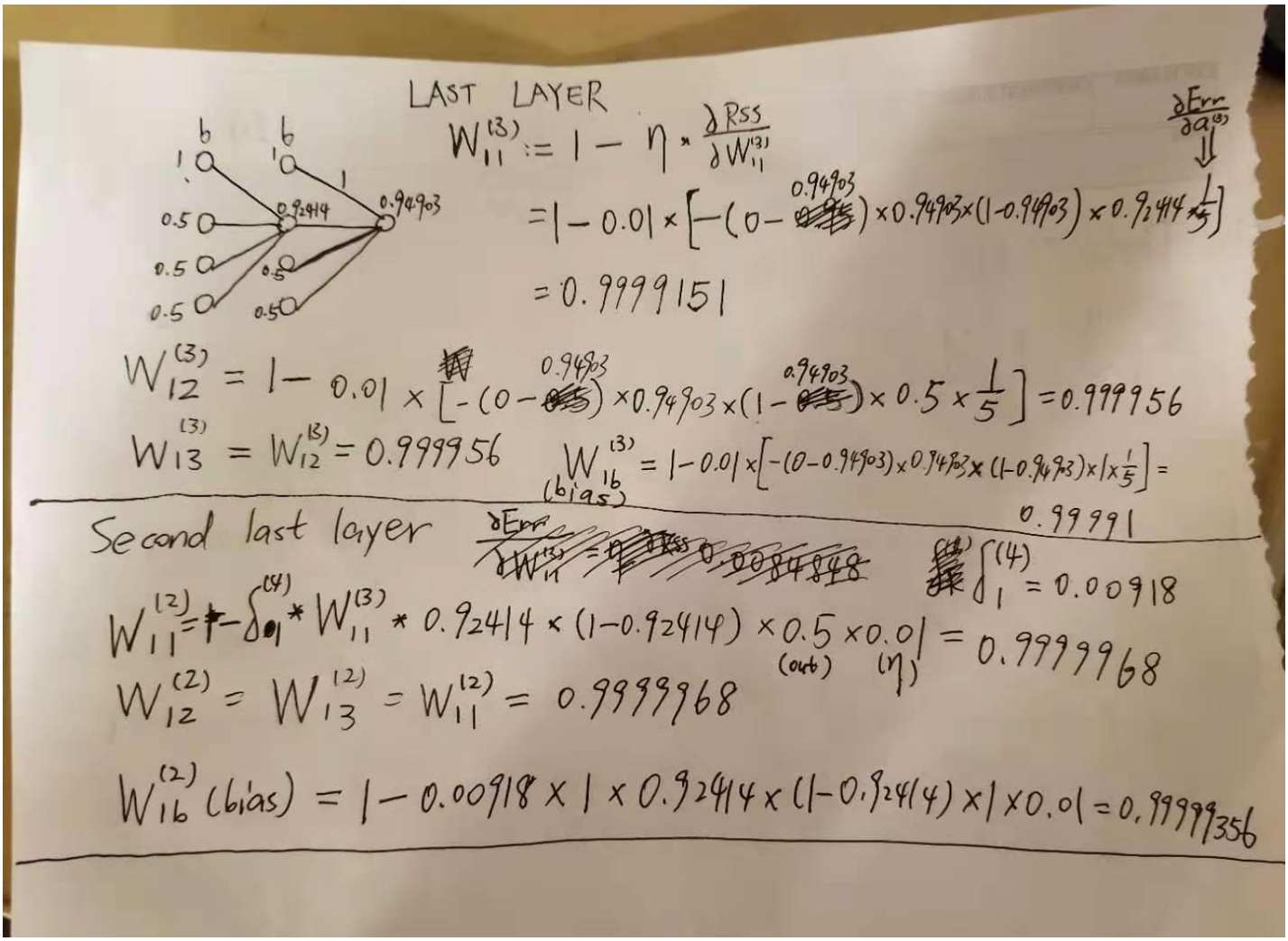
Last layer weight is [0.9999151, 0.9999541, 0.9999541]

Last layer bias is 0.9999082

Second last layer weight is [0.9999968, 0.9999968, 0.9999968]

Second last layer bias is 0.99999356

All error is within a range of 10^{-5}



Prob5

Misclassified Ratio Error:

	3 nodes	6 nodes	9 nodes	12 nodes
1 layer	0.188482	0.403141	0.413613	0.376963
2 layer	0.172775	0.167539	0.361257	0.413613
3 layer	0.172775	0.183246	0.382199	0.371728

The optimal setting is 2 layer 6 nodes with the lowest error. Apparently, such implementation can have the attributes of a medium amount of nodes and layers, which can help to avoid the situations of underfitting and overfitting. High amounts of layers and nodes will cause overfitting, which will increase error. Low amounts of layers and nodes will cause underfitting, which will also increase error.

Prob6

the unknown sample has the output of

[0.34324172 0.55863 0.02393821 0.10706931 0.02115896 0.00836715 0.01389998 0.01080599 0.0083985 0.00186314]

the unknown sample is likely to be NUC

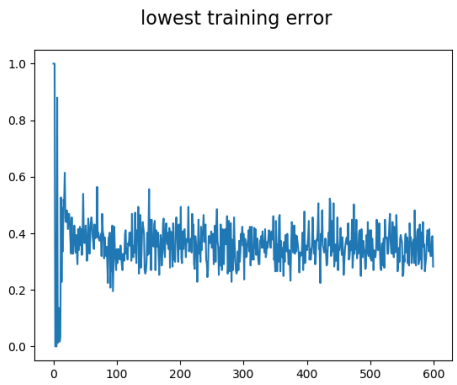
Prob7 Misclassified Ratio Error:

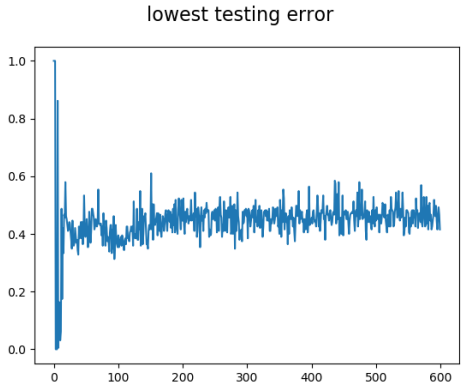
	3 nodes	6 nodes	9 nodes	12 nodes
1 layer	0.435897	0.415385	0.451282	0.420513
2 layer	0.307692	0.461538	0.4	0.425641
3 layer	0.276923	0.461538	0.410256	0.415385

Unfortunately, ReLU isn't better than sigmoid for the simple reason that ReLU isn't a suited model to simulate the correlation in this case since ReLU is more inclined to linear function compared to sigmoid.

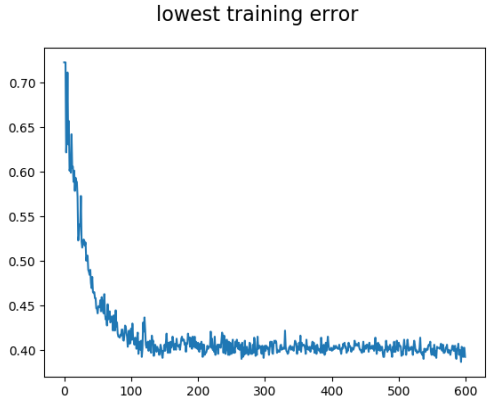
Run the script and plots will be saved to the current working directory:

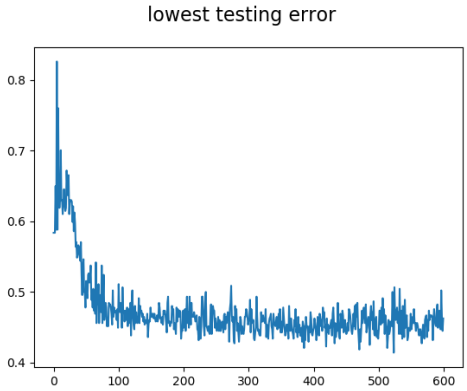
This is the plots of lowest training error in terms of misclassified ratio for **CYT**





This is the error in terms of misclassified ratio for **ALL** classes





Prob 8

To get uncertainty, our final activation function can be softmax in order to get us probability of each output predicted class, and use 1 minus each entry of prediction gives us uncertainty for each class.

the unknown sample has the uncertainty of [0.6840593 0.4230746 0.98809063 0.9192814 0.9979057 0.9995316 0.99829555 0.9939454 0.99613905 0.9996767]

As we can see here, for the unknown sample has the lowest uncertainty of 0.423 for the second output class which is NUC, so the result matches with the former prediction. Therefore, the uncertainty measurement is justified.