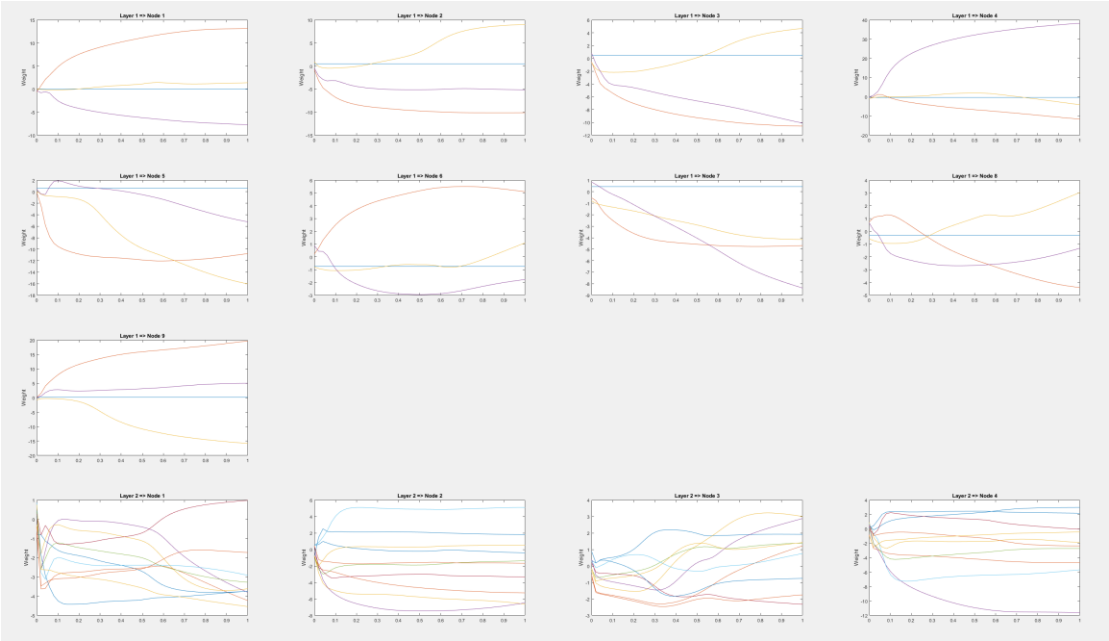
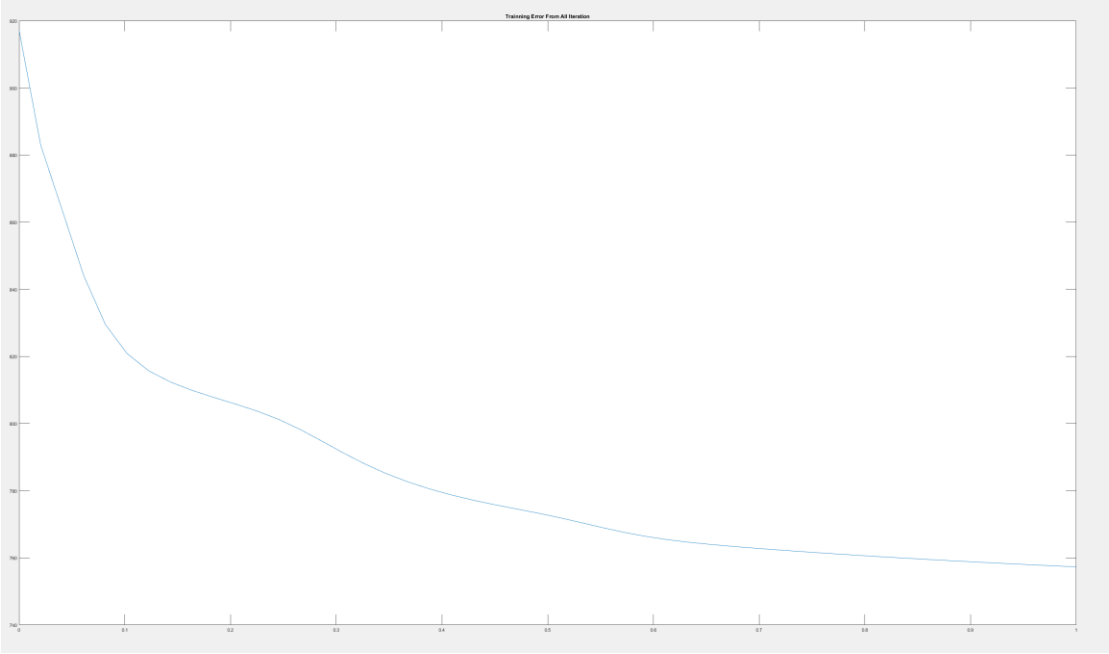
ECS 171 Homework Set 2 Haozhe Gu 999200555

1. Only Training have iterations, so Training Weight,Error are based on iteration. Training Output is based on the the single minimized error round.
   For the testing case, there's no iteration need. Also, weight does not change. So there are only plot for single minimized error round Error and Output.

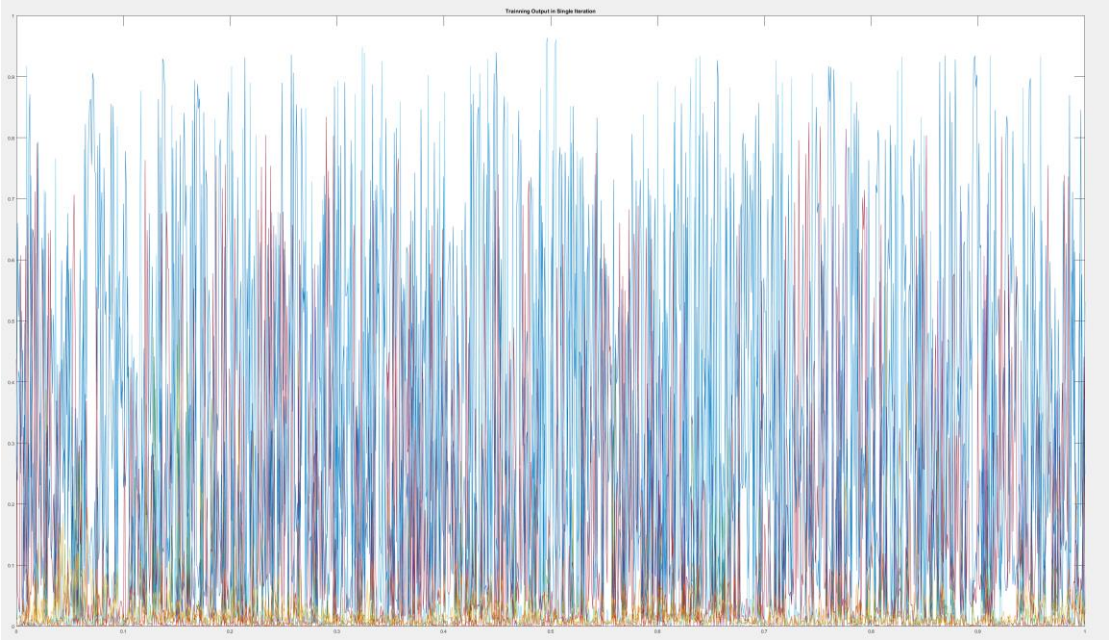
Training Weight Graph



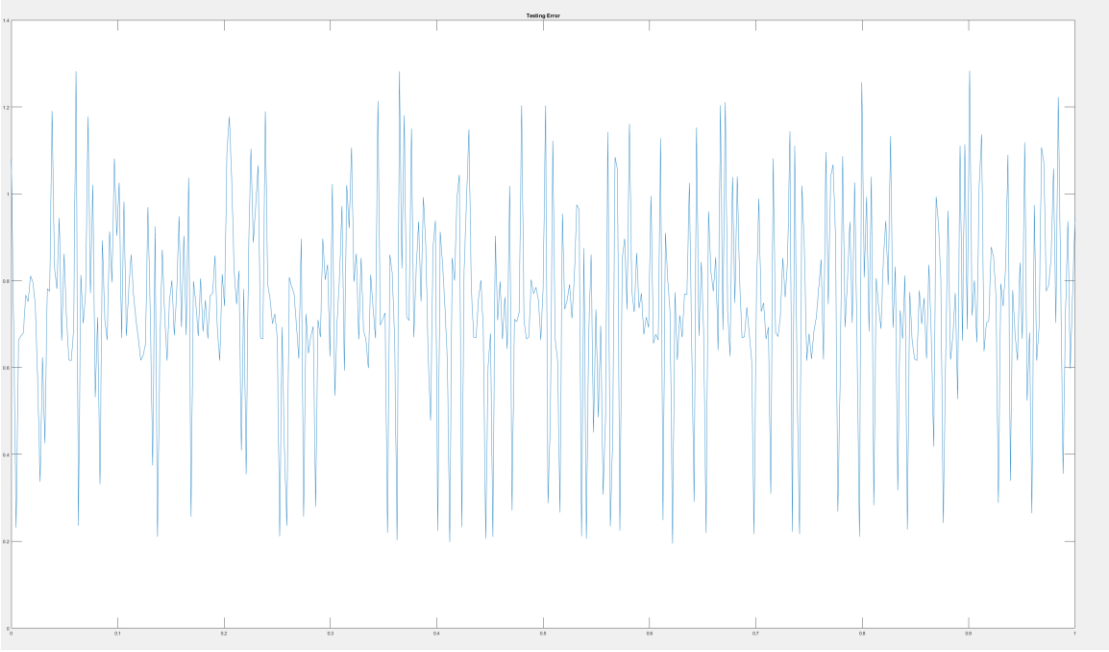
Training Error Graph



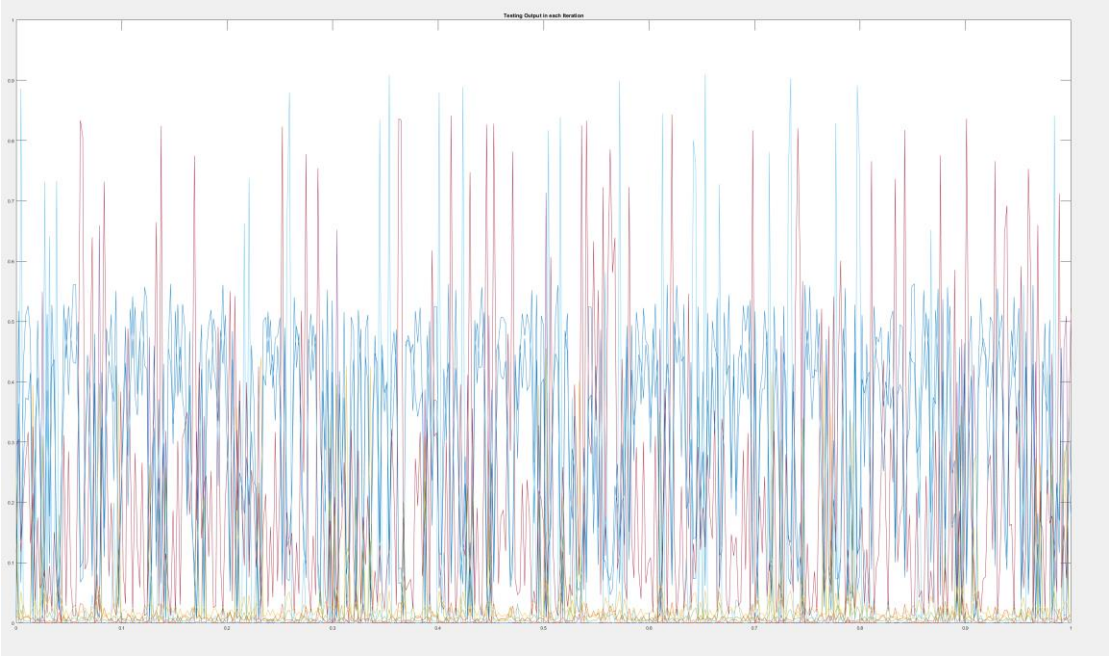
Trainning Output Graph



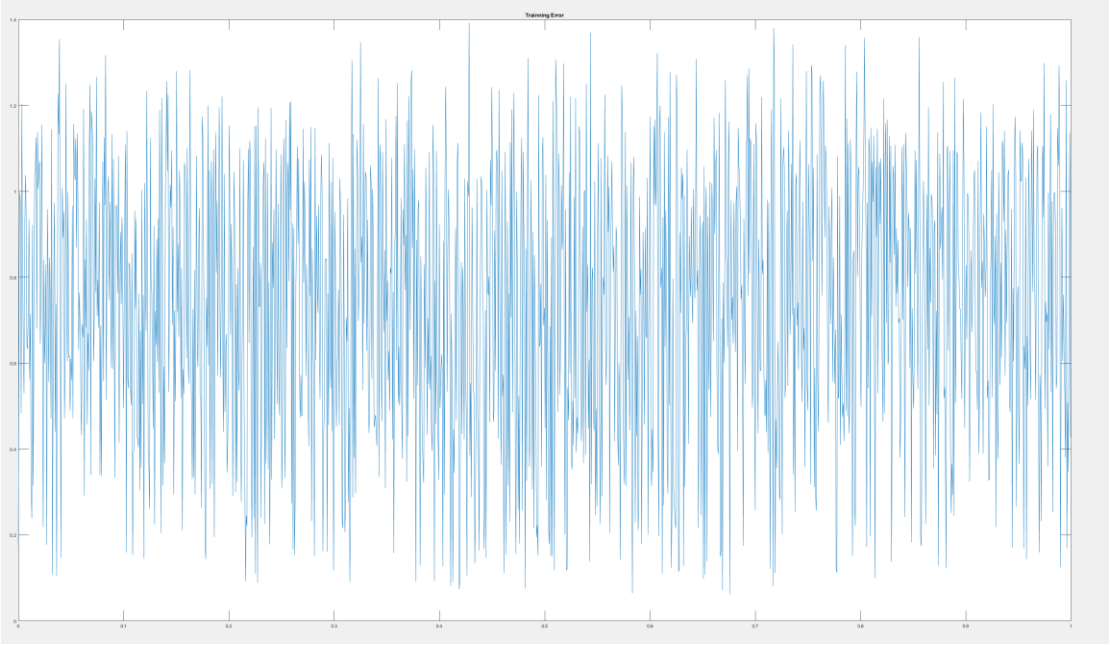
Testing Error Graph



Testing Output Graph



2. Training Error



Testing Training Error is : 1.094755e+03

Final Activation Function is : 0.1250 0.0000 0.0000 0.0001 0.0058 0.1363 0.0176 0.4829 0.0006 0.0052

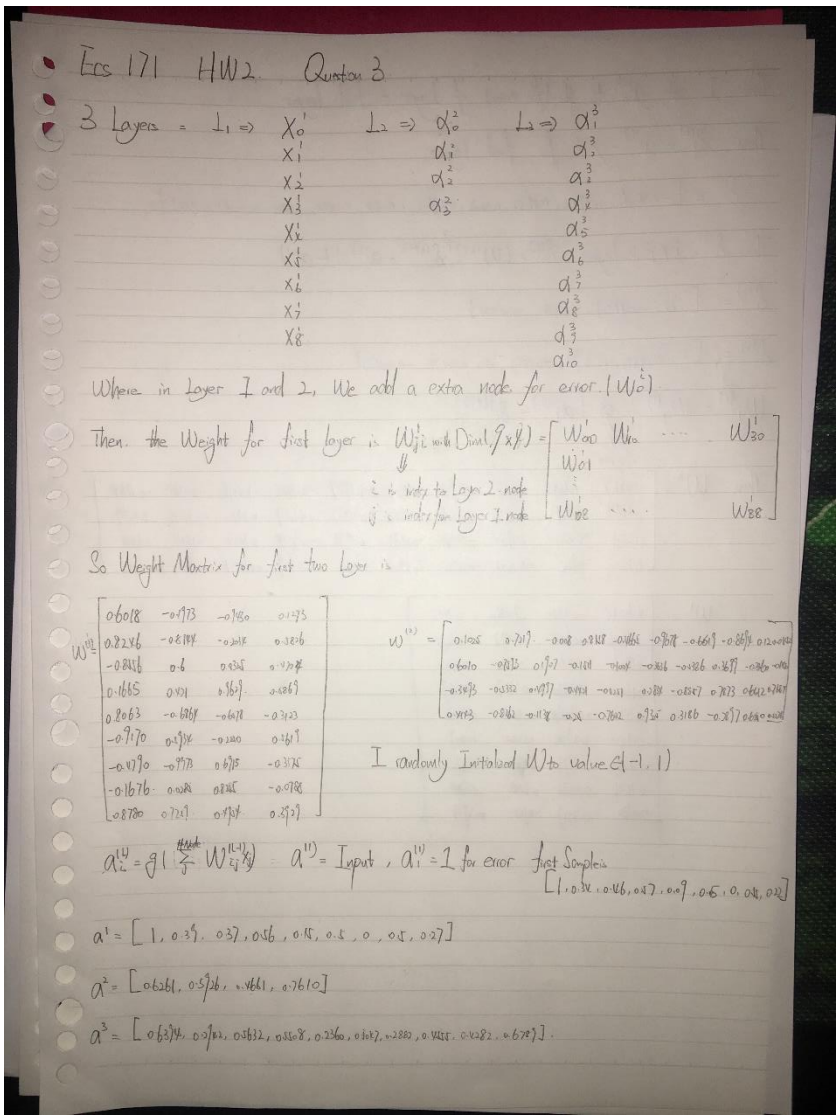
3. Output From Code (First is Weight for first Layer, Second is Weight for Second Layer)

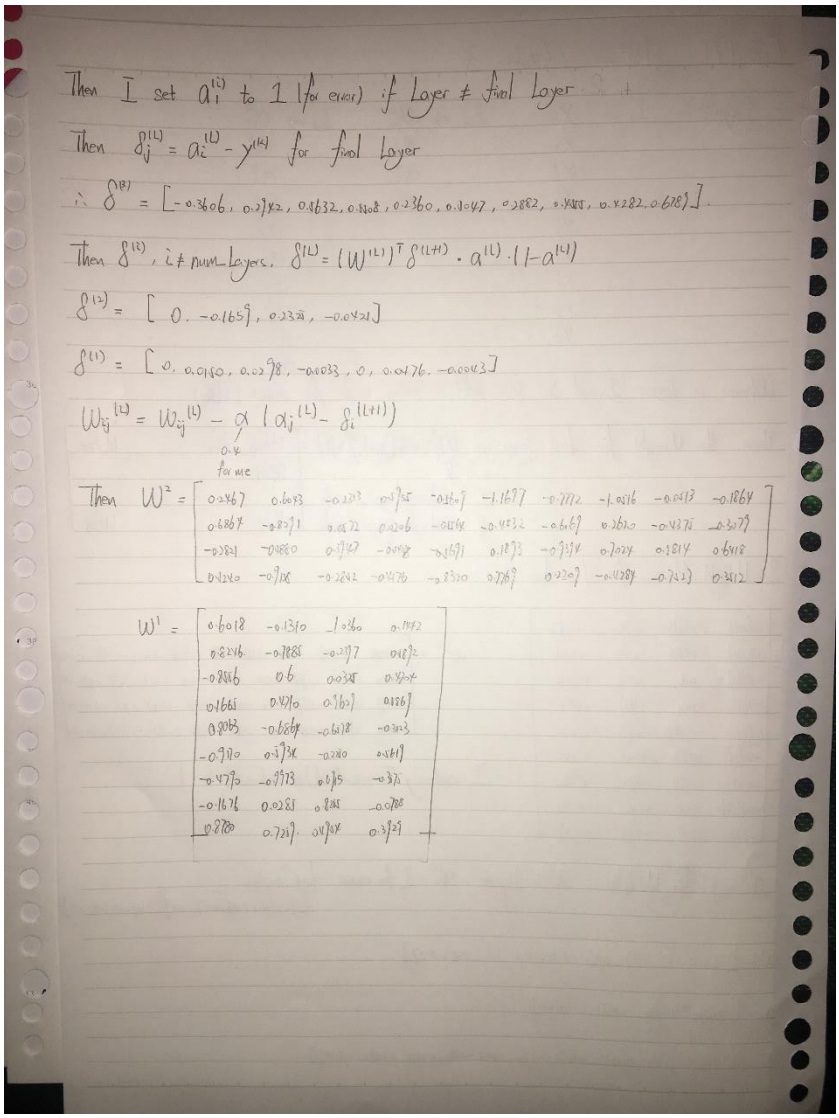
weight{1, 1}

	1	2	3	4
1	0.6018	-0.1310	-1.0360	0.1442
2	0.8246	-0.7885	-0.2397	0.5892
3	-0.8556	0.6000	0.0325	0.4704
4	0.1665	0.4710	0.9629	0.5869
5	0.8063	-0.6864	-0.6578	-0.3123
6	-0.9170	0.5934	-0.2250	0.5619
7	-0.4790	-0.9973	0.6915	-0.3575
8	-0.1676	0.0285	0.8255	-0.0788
9	0.8780	0.7259	0.4954	0.3929

weight{1, 2}

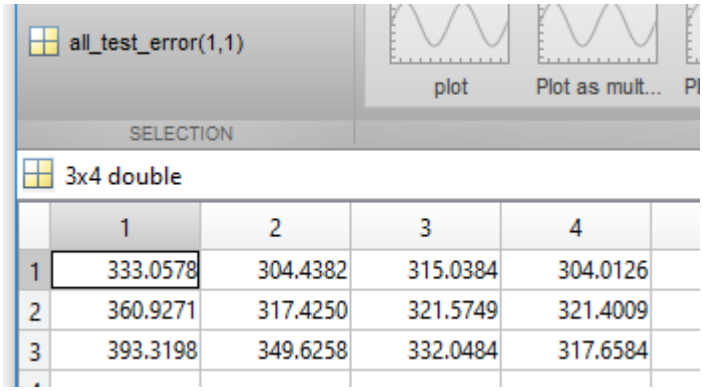
	1	2	3	4	5	6	7	8	9	10
1	0.2467	0.6043	-0.2333	0.5955	-0.5609	-1.1697	-0.7772	-1.0516	-0.0513	-0.1864
2	0.6864	-0.8291	0.0572	0.0206	-0.1564	-0.4832	-0.6069	0.2620	-0.4375	-0.3079
3	-0.2821	-0.5880	0.3947	-0.5448	-0.5691	0.1893	-0.9394	0.7024	0.5814	0.6418
4	0.5240	-0.9158	-0.2852	-0.4176	-0.8320	0.7769	0.2309	-0.4284	-0.7523	0.3512



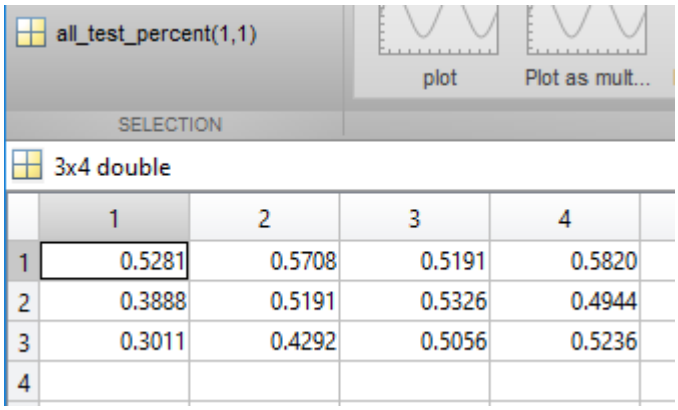


Both are in agreement.

4. Rows are Number of Layers {1,2,3}, Columns are Number of Nodes per Layer {3,6,9,12}
   The Second Graph is the corresponding Testing Classification Rate



	1	2	3	4
1	333.0578	304.4382	315.0384	304.0126
2	360.9271	317.4250	321.5749	321.4009
3	393.3198	349.6258	332.0484	317.6584



	1	2	3	4
1	0.5281	0.5708	0.5191	0.5820
2	0.3888	0.5191	0.5326	0.4944
3	0.3011	0.4292	0.5056	0.5236
4				

From the two matrix, the Optimal Configuration is 1 Hidden Layer with 12 Nodes per Hidden Layer.
The Relationship between these attributes can be summarized as: Generally, increase in Nodes per Hidden Layer will reduce the Error, while increase in Number of Layers might cause over-fitting which actually increase the Error in the test.

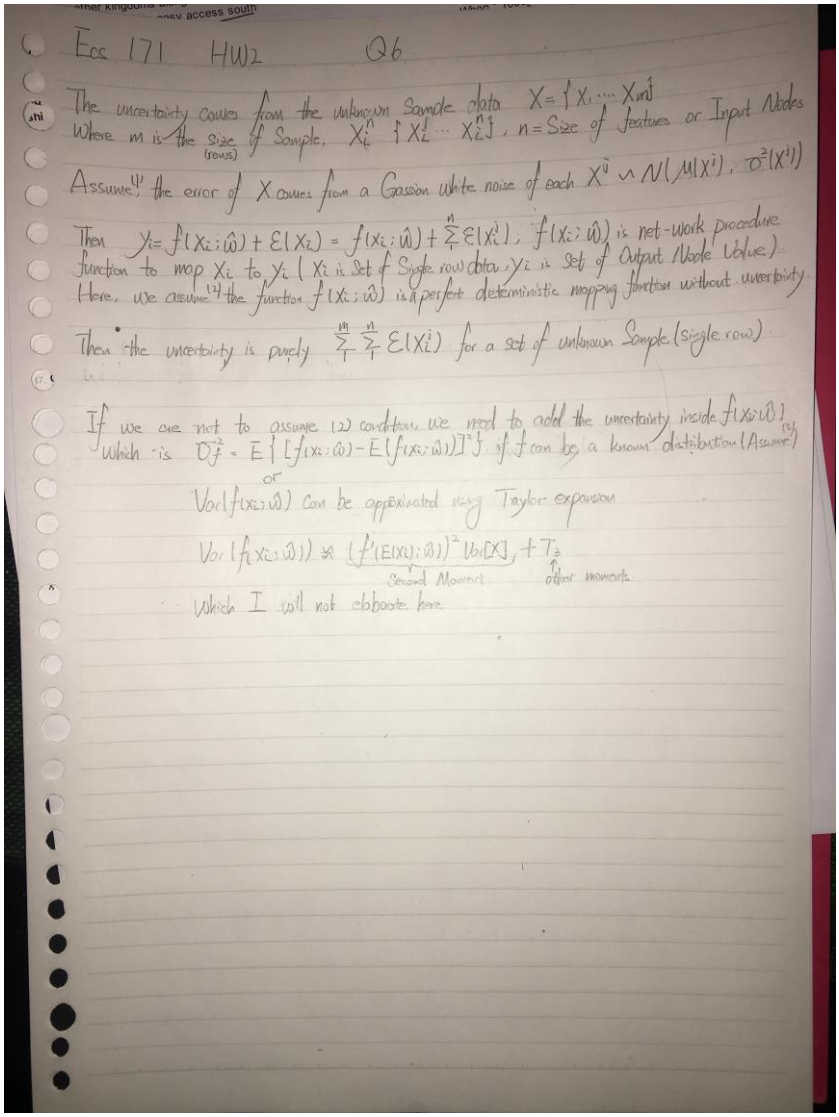
5.

```
Trainning percentage:
ans =

    0.5512

Predicted Category is: NUC
```

6.



Ecs 171 HW2 Q6

The uncertainty comes from the unknown Sample data $X = \{X_1, \dots X_m\}$
Where m is the size (rows) of Sample, $X_i = \{X_i^1 \dots X_i^n\}$, n = Size of features or Input Nodes

Assume (1) the error of X comes from a Gaussian white noise of each $X^i \sim N(\mu(X^i), \sigma^2(X^i))$

Then $y_i = f(x_i; \hat{w}) + \varepsilon(x_i) = f(x_i; \hat{w}) + \sum_1^n \varepsilon(x_i^j)$, $f(x_i; \hat{w})$ is net-work procedure function to map x_i to y_i (x_i is set of Single row data, y_i is set of Output Node Value)
Here, we assume (2) the function $f(x_i; \hat{w})$ is a perfect deterministic mapping function without uncertainty.

Then the uncertainty is purely $\sum_1^m \sum_1^n \varepsilon(x_i^j)$ for a set of unknown Sample (single row)

If we are not to assume (2) condition, we need to add the uncertainty inside $f(x_i; \hat{w})$, which is $\sigma_f^2 = E\{[f(x_i; \hat{w}) - E(f(x_i; \hat{w}))]^2\}$ if f can be a known distribution (Assume (3))

or

$Var(f(x_i; \hat{w}))$ can be approximated using Taylor expansion

$Var(f(x_i; \hat{w})) \approx \underbrace{(f'(E(x_i); \hat{w}))^2}_{\text{Second Moment}} Var[X] + \underbrace{T_3}_{\text{other moments}}$

Which I will not elaborate here