Q1. We have our activation function.

$$\text{LReLU}(x) = \begin{cases} 0.01x & \text{if } x < 0. \\ x & \text{otherwise.} \end{cases}$$

Our loss function is

$$E(w, v \mid x) = -\sum_t \left(r^t \log y^t + (1 - r^t) \log(1 - y^t)\right) + \sum_h \|w_h\|^2$$

where $y^t = \tanh\left(\sum_{h=1}^{H} v_h z_h^t + v_0\right)$

and $z_h^t = \text{LReLU}\left(w_h^T x^t\right)$.

We now will derive the update equations of the MLP with one hidden layer.

where

$$\Delta v_h = -\eta \frac{\partial E}{\partial v_h} = -\eta \left(\frac{\partial E}{\partial y^t}\right) \cdot \left(\frac{\partial y^t}{\partial v_h}\right) \quad \text{---(1)}$$

and $$\Delta w_h = -\eta \frac{\partial E}{\partial w_h} = -\eta \left(\frac{\partial E}{\partial y^t}\right)\left(\frac{\partial y^t}{\partial z_h}\right)\left(\frac{\partial z_h}{\partial w_h}\right) \quad \text{---(2)}$$

We first derive Δv_h.

$$\frac{\partial E}{\partial v_h} = -\eta \underbrace{\frac{\partial E}{\partial y^t}}_{(A)} \cdot \underbrace{\frac{\partial y^t}{\partial v_h}}_{(B)} \quad \text{---(1)}$$

A $$\frac{\partial E}{\partial y^t} = -\sum_t \frac{r^t}{y^t} - \frac{(1-r^t)}{1-y^t}$$

$$= -\sum_t \frac{r^t(1-y^t) - (1-r^t)y^t}{y^t(1-y^t)} \quad \text{——— (i)}$$

B $$\frac{\partial y^t}{\partial v_h} = 1 - \tanh\left(\sum_{h=1}^{H} v_h z_h^t + v_0\right) \cdot \sum_{h=1}^{H} z_h^t$$

$$= (1-y^t) z_h \quad \text{——— (ii)}$$

Replacing (i) and (ii) in (1)

$$\nabla v_h = \frac{\partial E}{\partial v_h} = \{\eta\}\left(-\sum_t \frac{(1-y^t)r^t - (1-r^t)y^t}{y^t(1-y^t)} \cdot (1+y^t)(1-y^t) \cdot z_h^t\right)$$

$$= (-\eta)\left(-\sum_t \frac{(r^t - y^t)}{y^t}(1+y^t) \cdot z^h\right)$$

$$\nabla v_h = \eta \sum_t \frac{(r^t - y^t)(1+y^t)}{y^t} z_h^t$$

we now derive w_h.

$$\frac{\partial y^t}{\partial z_h^t} = \left(1 - \tanh^2\left(\sum_{h=1}^{H} v_h z_h + v_0\right)\right) \cdot v_h$$

$$= (1 - (y^t)^2) \cdot v_h \quad \text{——— (iii)}$$

$$z_h^t = \text{LReLU}(w_h^T x^t)$$

$$\frac{\partial z_h^t}{\partial w_h} = \frac{\partial \text{LReLU}(w_h^T x^t)}{\partial w_h}$$

$$= \text{LReLU}'(w_h^T x^t) \cdot x^t$$

$$= 0.01(w_h^T x^t) \quad \text{when } (w_h^T x^t) < 0. \quad \text{——— (iv)}$$

$$= w_h^T x^t \quad \text{when } (w_h^T x^t) \geq 0.$$

we know that

$$E(w, v/x) = \underbrace{\sum_t \left(r_t \log y_t + (1-r_t)\log(1-y_t)\right)}_{A} + \underbrace{\sum_h \|w_h\|_2^2}_{B}$$

$$\frac{\partial E}{\partial w} = \frac{\partial A}{\partial w} + \frac{\partial B}{\partial w}$$

$$\frac{\partial A}{\partial w} = \left(\frac{\partial A}{\partial y_t}\right)\left(\frac{\partial y_t}{\partial z_{ht}}\right)\left(\frac{\partial z_{ht}}{\partial w_h}\right)$$

$$\frac{\partial A}{\partial y_t} = -\sum_t \frac{r_t(1-y_t) - (1-r_t)y_t}{y_t(1-y_t)} \quad \text{---(iii)}$$

using our previously calculated $\frac{\partial y_t}{\partial z_{ht}}$ and $\frac{\partial z_{ht}}{\partial w_h}$

in $\frac{\partial A}{\partial w}$

we get

$$\frac{\partial A}{\partial w} = (-\eta)\left(-\sum_t \frac{r_t(1-y_t)(1-r_t)y_t}{y_t(1-y_t)} \cdot (1+y_t)(1-y_t) \cdot v_h \cdot \text{LReLU}(w_h^T x)\right.$$

$$= \eta \sum_t \frac{(r_t - y_t)(1+y_t)}{y_t} v_h \cdot \text{LReLU}'(w_h^T x)$$

$$\frac{\partial B}{\partial w} = \frac{\partial}{\partial w}\left(\sum_h \|w_h\|_2^2\right)$$

$$= 2w_h.$$

combining these two we get

$$\Delta w_h = \eta \left(\sum_t \frac{(r^t - y^t)(1+y^t)}{y^t} \cdot v_h \cdot \text{LReLU}(w_h^T x) \cdot x^t + 2w_h \right)$$

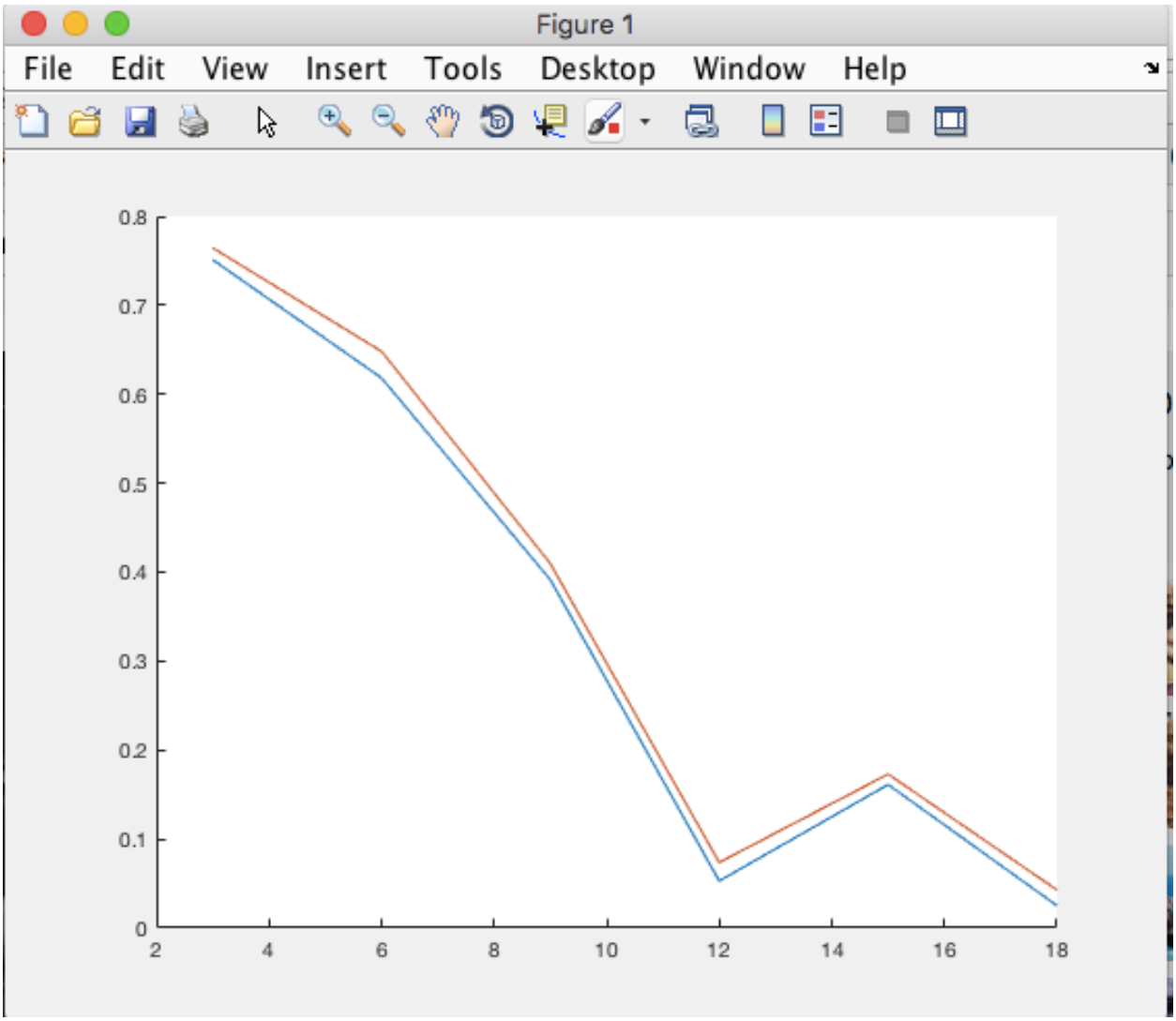
$$= \eta \left(\sum_t \frac{(r^t - y^t)(1+y^t)}{y^t} \cdot v_h \times 0.01 \cdot (w_h^T x^t) \cdot x^t + 2w_h \right)$$

when $(w_h^T x^t) < 0$

$$= \eta \left(\sum_t \frac{(r^t - y^t)(1+y^t)}{y^t} \; v_h w_h^T x^t + 2w_h \right)$$ for all other values of $w_h^T x^t$.

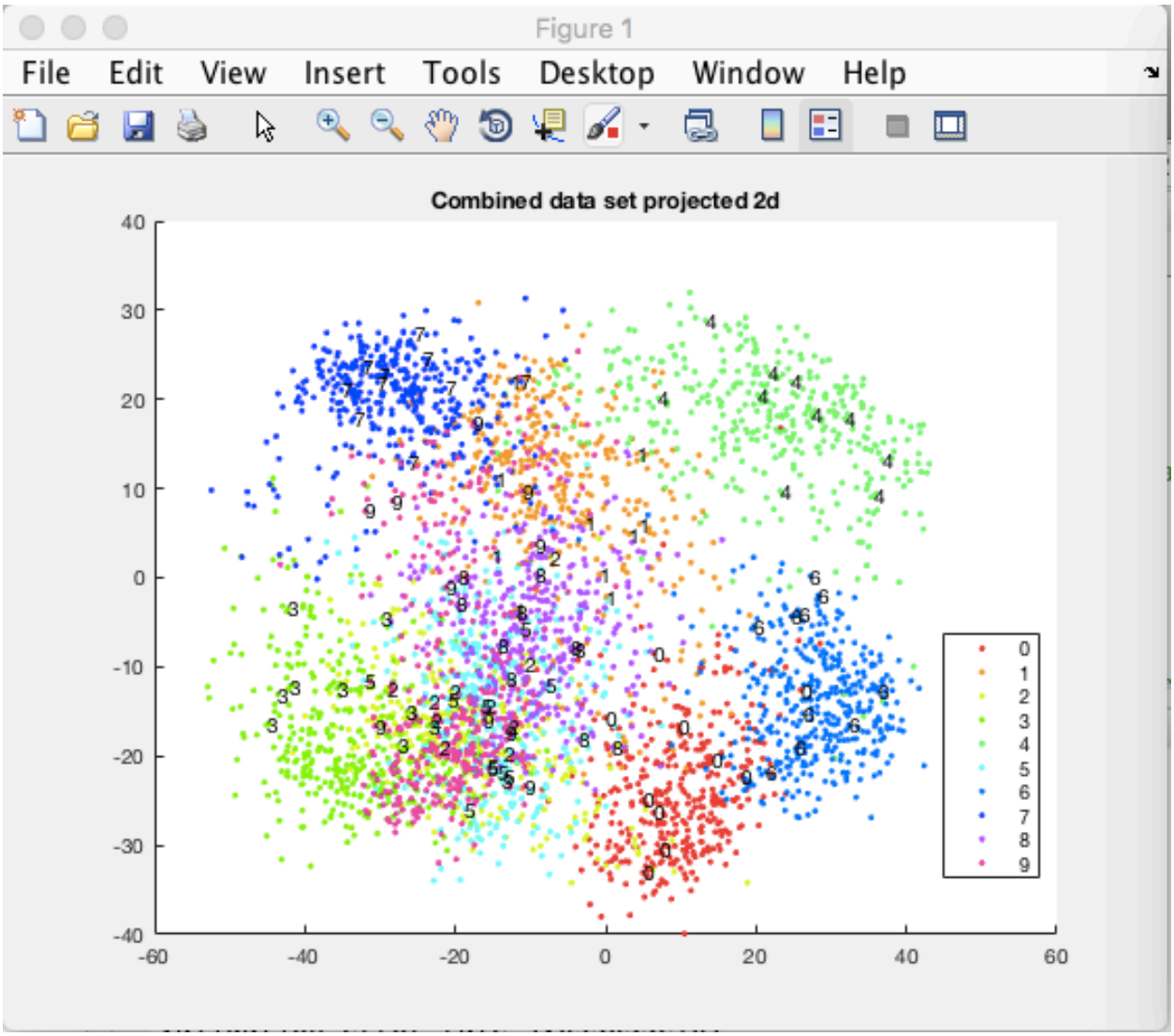
Q2c. On running the mlp training for the given number of hidden units we get the following results.

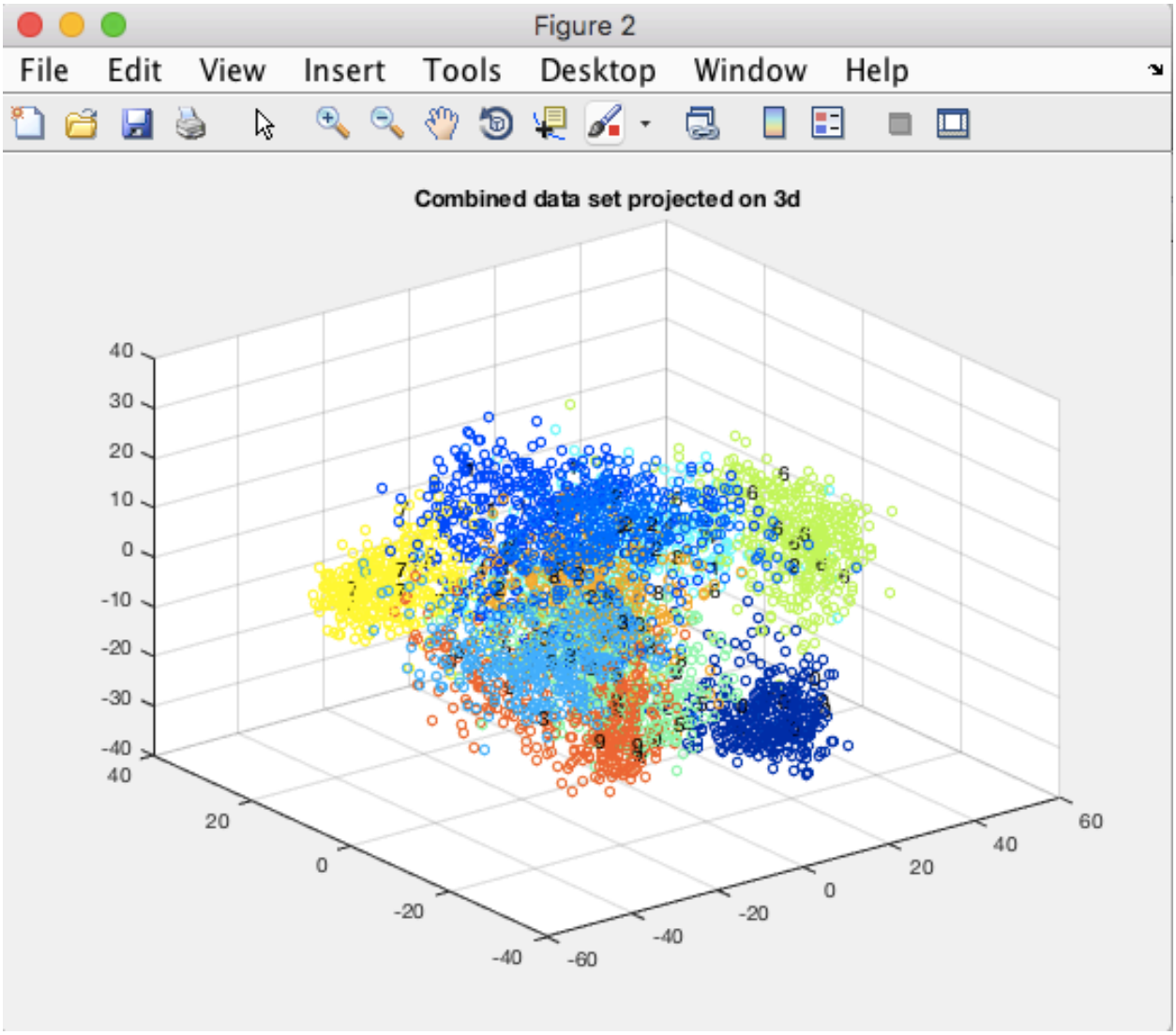
```
For 3 hidden units
converged error -3.657286e+03
Train error_rate 7.512013e+01
Validation error rate 7.645489e+01
For 6 hidden units
converged error -3.034975e+03
Train error_rate 6.182595e+01
Validation error rate 6.481580e+01
For 9 hidden units
converged error -2.003261e+03
Train error_rate 3.913508e+01
Validation error rate 4.095035e+01
For 12 hidden units
converged error -3.214572e+02
Train error_rate 5.285638e+00
Validation error rate 7.367859e+00
For 15 hidden units
converged error -8.608202e+02
Train error_rate 1.612387e+01
Validation error rate 1.729845e+01
For 18 hidden units
converged error -1.553472e+02
Train error_rate 2.509343e+00
Validation error rate 4.271223e+00
Error rate on test set for 18
Error rate is 4.162220e-02
>>
```



From the data we can see that we obtain the lowest validation error when the validation unit are equal to 18.
Using this set of hidden values we get the an error rate of 4.16% on the test set of data.

Q2b.
Using the w from our previous training on the combined training and validation set we arrive at newly calculated z-values. We use PCA to arrive at the first 2 and first 3 principal components and then plot these data points and label them by their classes which results in the following plots.



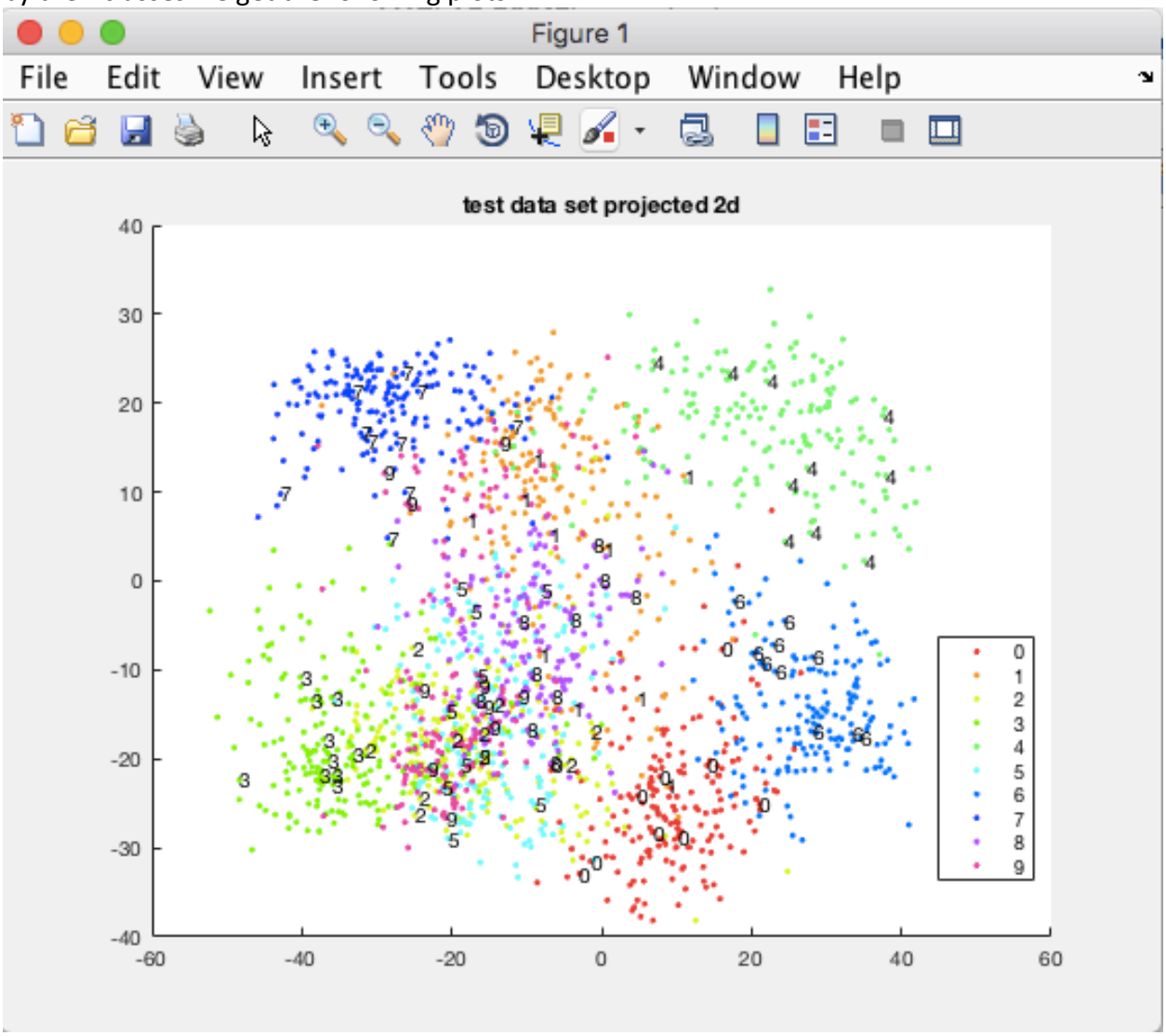


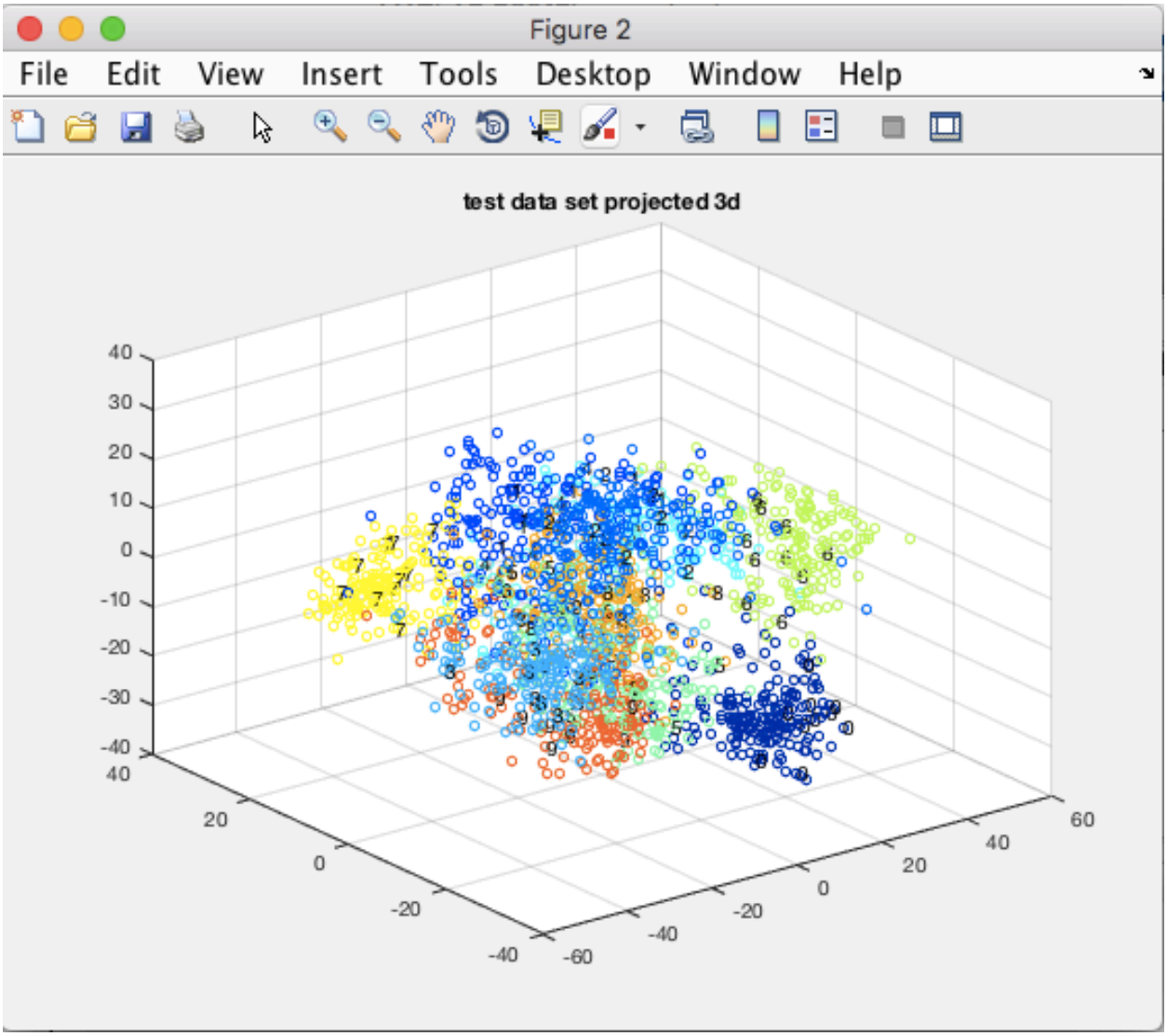
From our plots, we observe definite clusters for each class with a few outliers. Due to the reduction of our z-values into 2 and 3 dimensions, there is considerable overlap in these clusters, however, this would not be the case if we were able to plot the z-values in all of it's dimensions.

The noticeable difference between the 2d and the 3d plots are the fact that the 3d plots contains more separation of the classes.

Q2c

Using the principal components learned from the earlier question, we now generate the first 2 and first 3 principal components for the test dataset. On doing so and plotting the data-points by their classes we get the following plots.





On plotting these points using the principal components learned in the previous question, we notice that the clusters of each class are almost similar to the ones generated by the combined dataset. Except for the reduced number of data points in the test data set plots, the corresponding 2d and 3d plots are identical.