Introduction to Machine Learning HW-4

Ujval Bangalore Umesh 5202540

Problem 1

$$E(w, v|x) = -\sum_t r^t \log y^t + (1-r^t)\log(1-y^t) + \sum_h \|w_h\|_2^2$$

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

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

The update equations are Δv_h and Δw_{hj}

$$\frac{\partial E}{\partial v_h} = \frac{\partial E}{\partial y^t} \times \frac{\partial y^t}{\partial v_h}$$

$$= -\sum_t \left[r^t \frac{1}{y^t} + \frac{(1-r^t)(-1)}{(1-y^t)} \right] \cdot \left[1-y^{t2}\right] z_h^t + 0$$

$$\Delta v_{th} = -\eta \frac{\partial E}{\partial v_h}$$

$$= \eta \sum_t \frac{\left(r^t\left[1-y^t\right] + (1-r^t)y^t\right)\left[1-y^{t2}\right] z_h^t}{y^t(1-y^t)}$$

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

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

$$\Delta w_h = -\eta \frac{\partial E}{\partial y^t} \frac{dy^t}{\partial z_h^t} \frac{\partial z_h^t}{\partial w_h} = -\eta \frac{\partial E}{\partial w_h}$$

$$= +\eta \sum_t \left[\frac{r^t}{y^t} - \frac{(1-r^t)}{1-y^t} \right] (1-y^{t2}) v_h \times LReLU'(w_h^T x) \cdot x^t + 2 \sum_h w_h$$

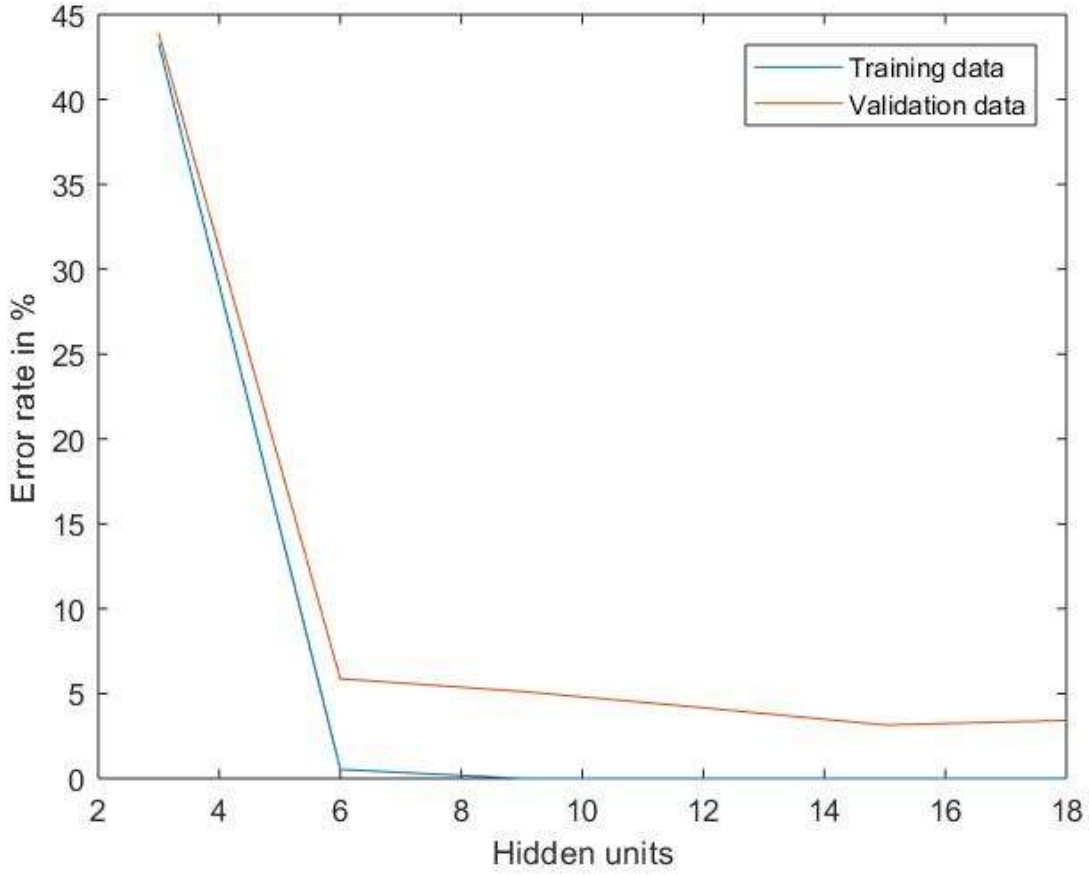
$$= \eta \sum_t \frac{(r^t - y^t)(1+y^t)}{y^t} v_h \times LReLU'(w_h^T x^t) x^t + 2 \sum_h w_h$$

$$LReLU'(w_h^T x^t) = \begin{cases} 0.01 & \text{if } w_h^T x < 0 \\ 1 & \text{otherwise} \end{cases}$$

2a) The training error rate and Validation error rates are tabulated after simulations.

Number of Hidden Units	Error rate on Training	Error rate on validation
3	43.2995	43.8868
6	0.5339	5.8729
9	0	5.1255
12	0	4.1644
15	0	3.1500
18	0	3.4170

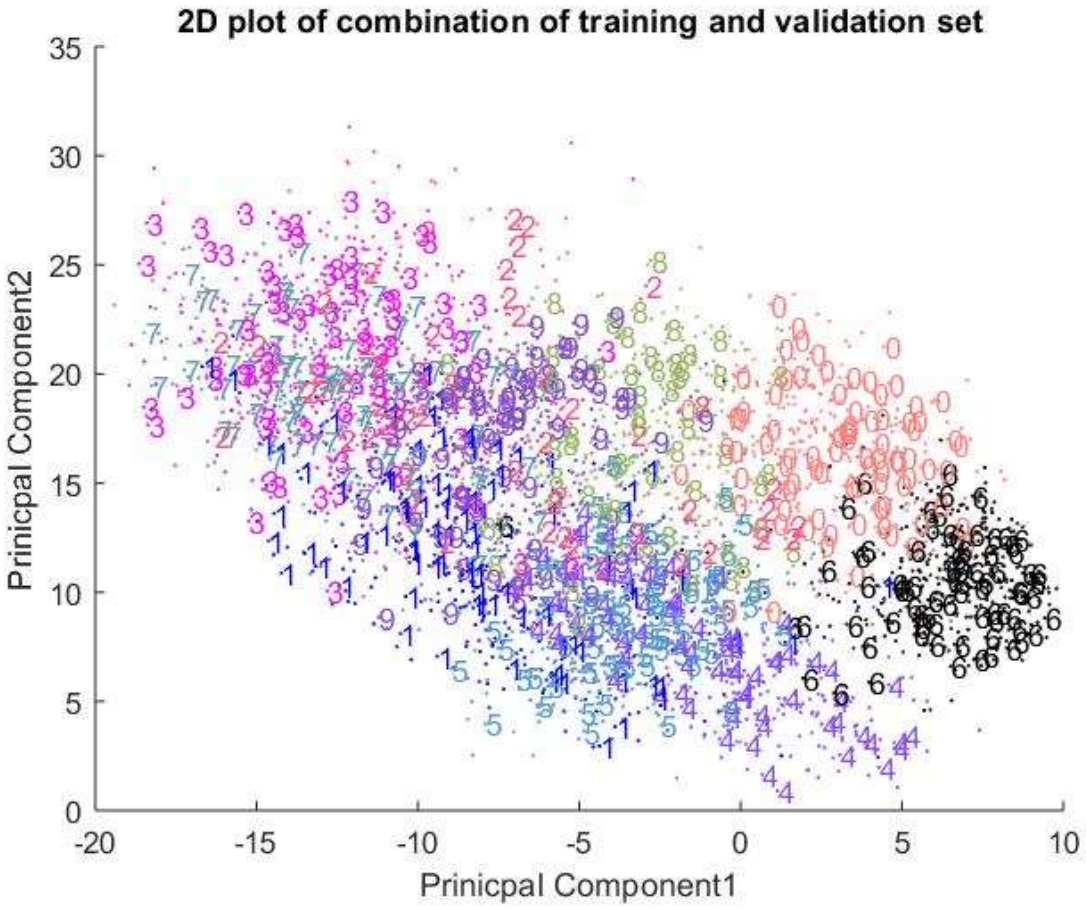
Plot of the Training and Validation error rates by the number of hidden units is depicted below

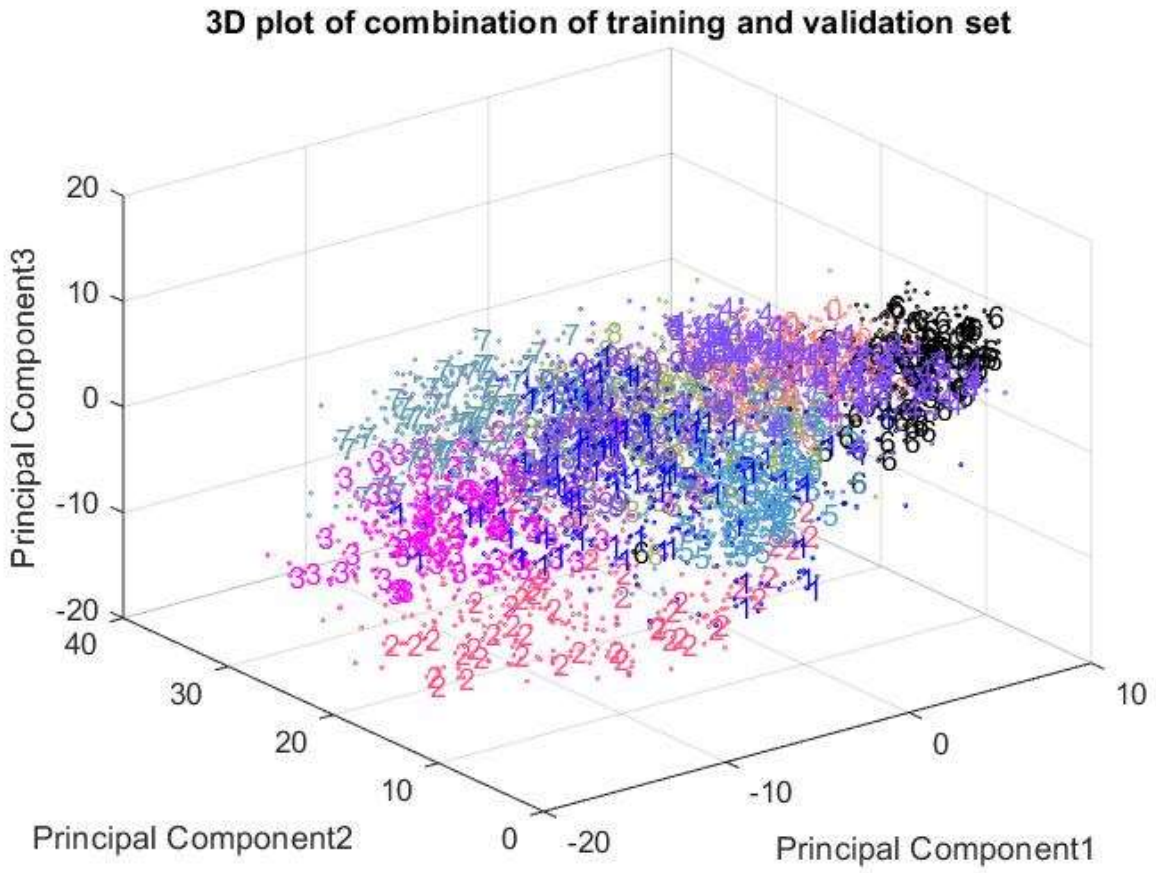


We can observe that the minimum error rate is obtained for m= 15 hidden units. The error rate is 0 for training data and 3.15% for validation data. Hence, I have chosen to use 15 hidden units.

The error rate on the test data set with m=15 number of hidden units is found to **be 3.5752%.**

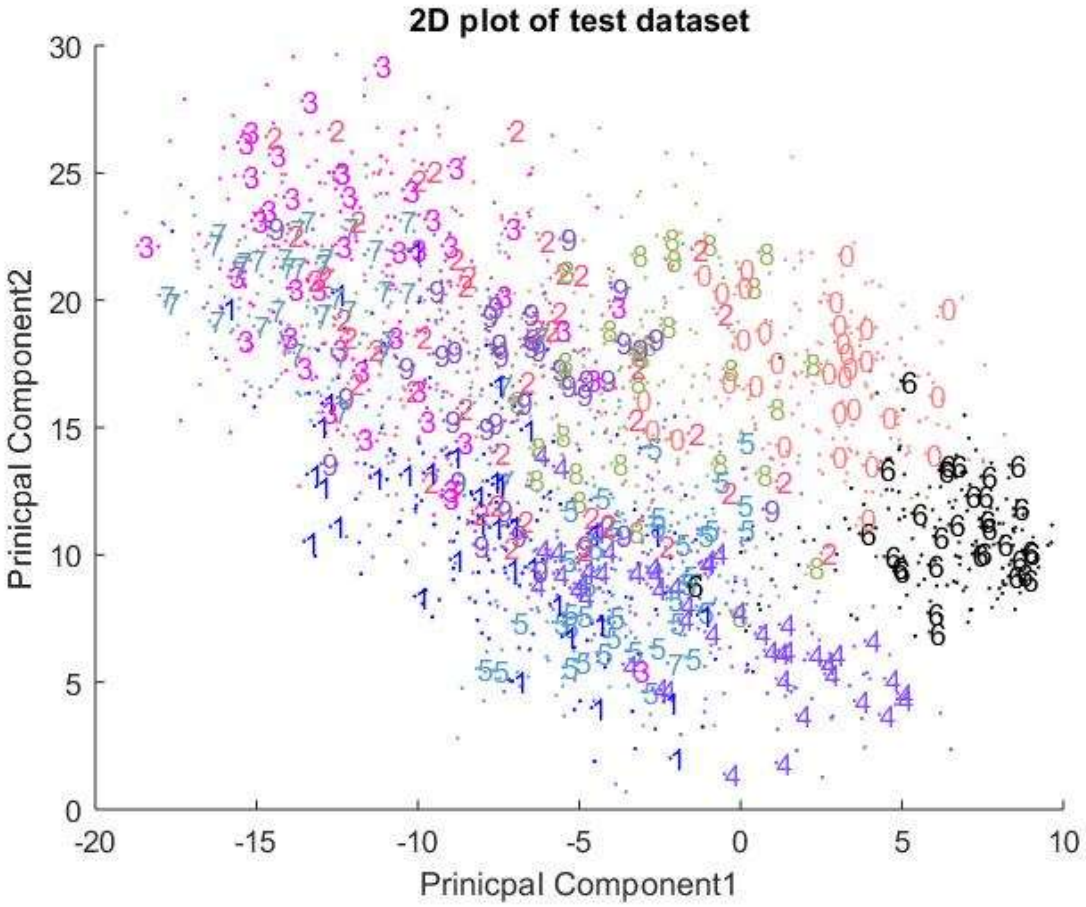
2b) 2D and 3D plots of combination of training and validation data are shown.

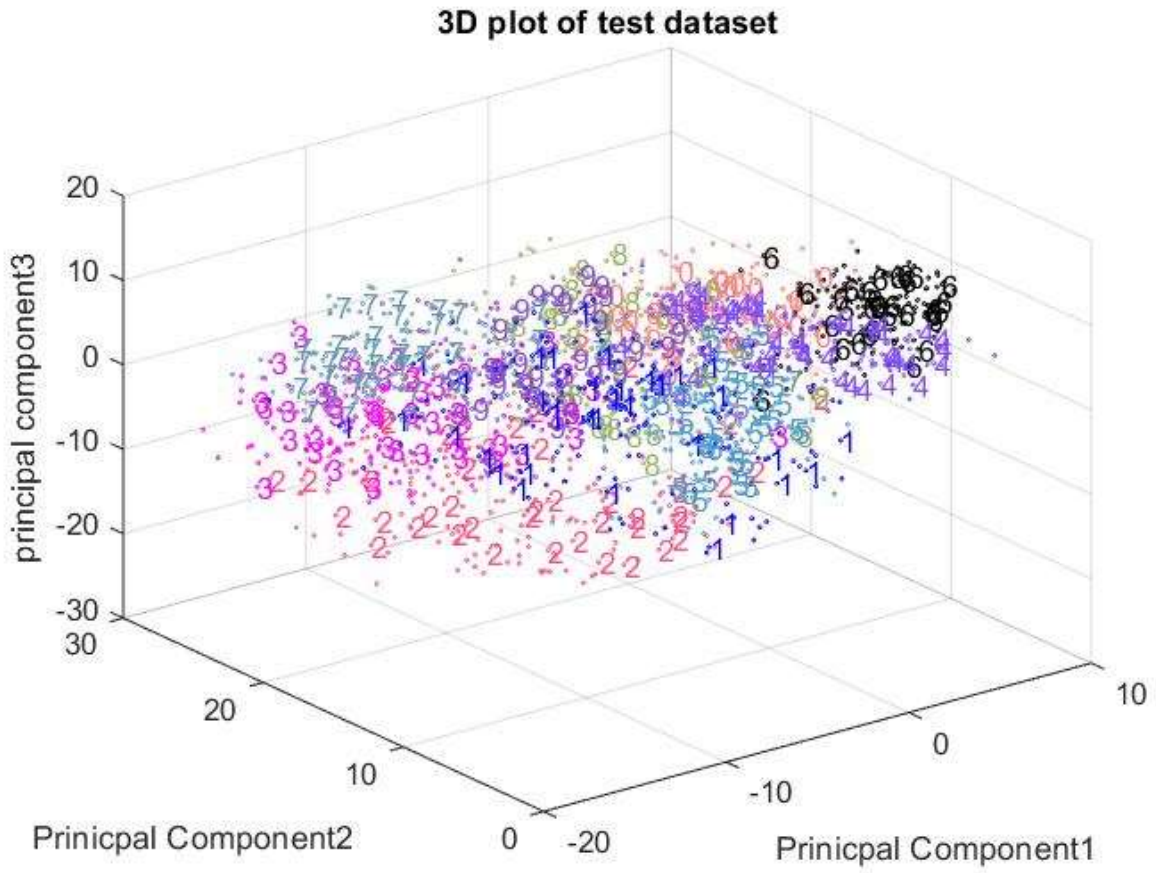




The ‘d’ dimensions of data are reduced to 15 (# of hidden layers) dimensions in this case by the MLP implementation. The implementation of PCA sorts the eigenvectors which have the highest variance. Thus, we can see clear class separation in the 3D plot. The 2D plot clearly depicts clustering in a way that minimizes mis-classification error.

2c) 2D and 3D plots of test data are shown.





The weights learnt through the backpropagation (gradient descent) in the training data reduces the mis-classification error rate for the training data. When we use the combination on the train and valid data, the error rate (1.5750) so obtained by using the learnt parameters and optimal hidden units is comparatively lesser than the error rate obtained on the forward step in the test data (3.5752%). We can visually inspect that the 2D and 3D plot of combination of training and validation data shows less misclassification than the 2D and 3D plot of test data.