CSCE 636: Neural Networks (Fall 2018).

Assignment #1.

Niraj Goel.

226009509.

Q1.

a). The function train_valid_split splits the training data into training and validation set. The validation set is what is used for hyperparameter tuning. Using the test set would be cheating as test-set should not be used at any point during training. Using the training data by itself will lead to overfitting, Hence the Validation split is required.

b) Yes, ~~#~~ Retraining the model on the whole training set usually leads to better generalization of the model i.e it will perform relatively better with unseen/test data. The reason for this is the more amount of data, the model is trained with when validation set is also included. However, In case of deep networks with large data set retraining might not be feasible, In such situation it might not be a good idea to retrain.

d) The third feature which is always 1 is ~~bias~~ used to introduce bias in the model. The model is given by

$$y = wx + b.$$

if bias is zero, the ~~d~~. Separation line will always pass through the origin, which might not be the case always. Hence, the bias term is required for the shifting of the decision boundry. ie it gives the model the ability to have decision boundary's which does not necessarily pass through origin.

f). Figure included in last page.

Q2.

c) Figure Included in last page.

d) Test accuracy obtained with best model :- ~~95.28~~%. 96.22%

The best model for me was the one with 200 iterations in this case.

Q3.

a) The loss (cross entropy) function for one training data Sample (x,y) is given by

$$e(x,y) = \ln\left(1 + e^{-y \cdot w^T x}\right)$$

b) $\nabla E(w)$.

$$= \frac{d\left(\ln\left(1+e^{-y \cdot w^T x}\right)\right)}{dw}$$

$$= \frac{1}{1+e^{-y w^T x}} \cdot \frac{d\left(1+e^{-y w^T x}\right)}{dw}$$

$$= \frac{e^{-y w^T x}(-y \cdot x)}{1+e^{-y w^T x}} = -\frac{y \cdot x}{1+e^{y w^T x}}$$

$$\therefore \nabla E(w) = -\frac{y \cdot x}{1+e^{y w^T x}}$$ { for one training example.

c) Since the decision boundary is linear, Using the Sigmoid function followed by the threshold to predict the class is not very efficient. The class prediction can also be done by some simpler function like "Sign" and simpler algorithms like PLA or Pocket.

The Utility of Sigmoid function comes from the fact that it gives probability as output and then the class can be assigned on the basis of some threshold value on that probability. This comes in very handy in applications where just giving a positive or negative (PLA) won't make much sense eg. to predict if a person will get cancer or not, it makes more sense to get ~~a~~ the probability of getting a cancer as the accuracy will be terrible if we directly predict just +1 or -1. In other words, Sigmoid function helps in getting the measure of Uncertainity.

d). Yes the decision boundary is still linear.

$z = w^T x.$ we have Sigmoid $= \frac{1}{1+e^{-z}}.$

with threshold = 0.9 → $\frac{1}{1+e^{-z}} = 0.9$

$\therefore \quad 1+e^{-z} = \frac{10}{9}$

$\Rightarrow e^{-z} = \frac{1}{9}$

$\Rightarrow e^{z} = 9$

or $z = \log_e 9$

$\Rightarrow w^T x = \log_e 9$ → ~~hence the~~ ~~a~~ this is a linear function in x, therefore the decision boundary is linear.

e). The essential properties which leads to linear decision boundaries are :-

i) hypothesis, $h(x) = \theta(w^T x)$ where $\theta = \frac{1}{1+e^{-z}}$

ii) Prediction is given by : $h(x) \geq 0.5$ and $h(x) < 0.5$

from i) and ii) we get .

$y = 1$ when $h(x) \geq 0.5$.

or $\theta(w^T x) \geq 0.5$.

ie $w^T x \geq 0$. — Ⓐ

Similarly $y = -1$ when $h(x) < 0.5$

or $\theta(w^T x) < 0.5$

or $w^T x < 0$. — Ⓑ

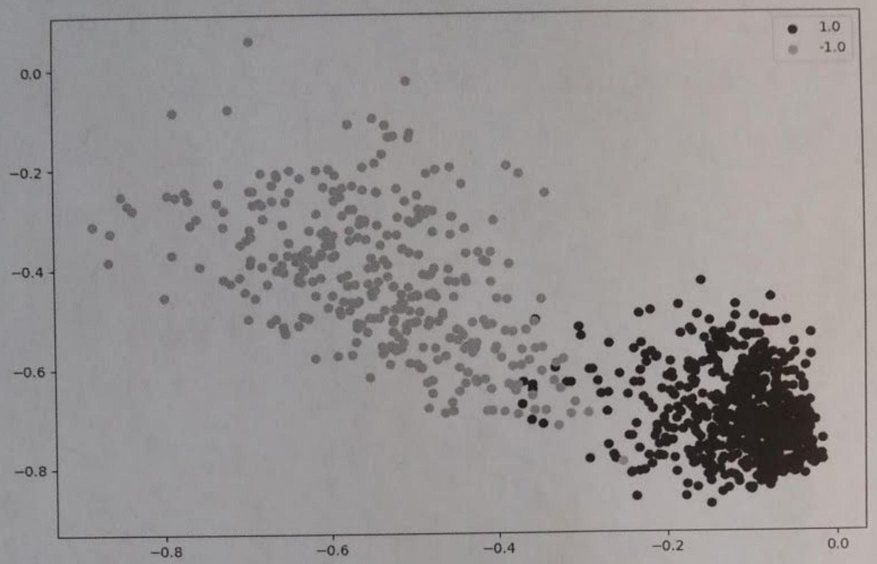
Ⓐ & Ⓑ is what leads to the decision boundry, Since they are linear ~~for~~ combination of x, the decision boundary comes out as linear.

Q4) d) Figure Included in last page.

e) Test Accuracy with best model :- 95.28 %

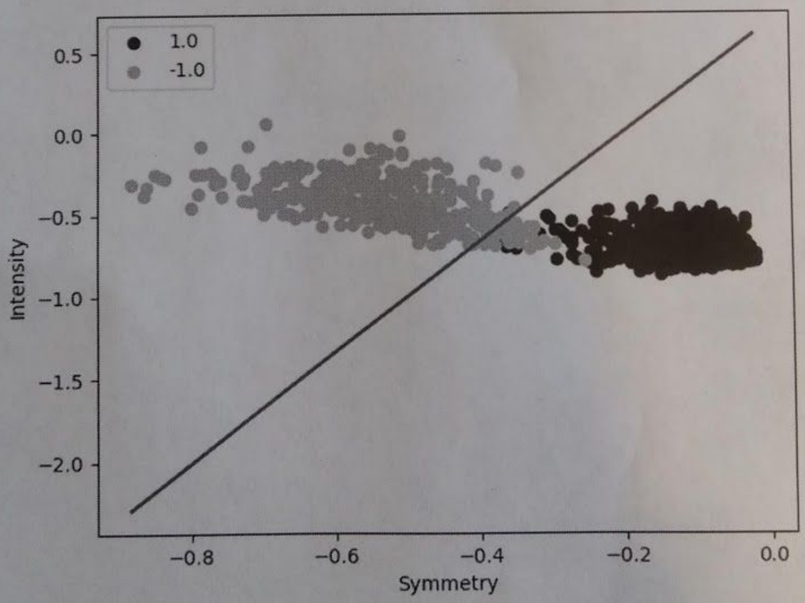
Best model parameters :- Learning Rate : 0.8
Max_iters : 500
Batch Size :- 1

Q1) f)



● 1.0
● -1.0
0.0
−0.2
−0.4
−0.6
−0.8
−0.8
−0.6
−0.4
−0.2
0.0

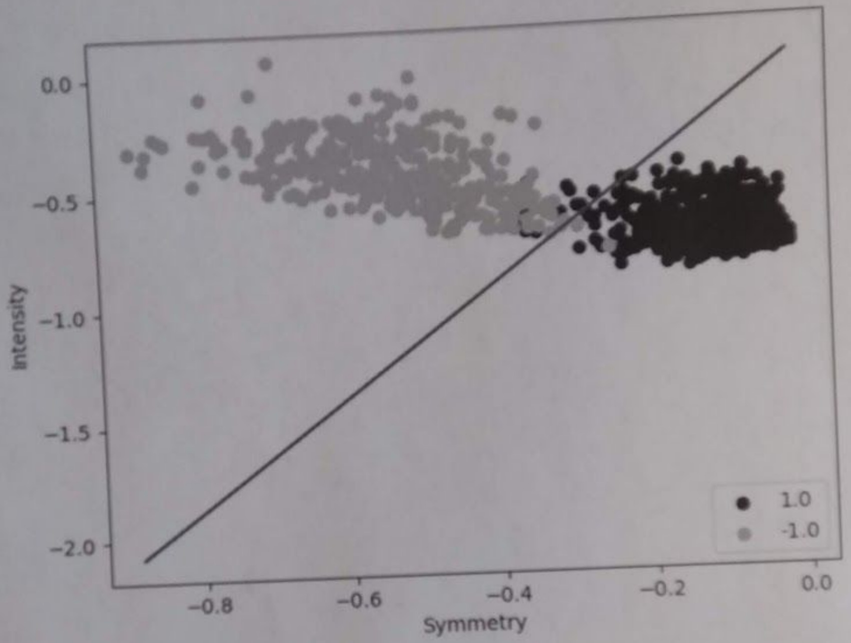
Q2c)



● 1.0
● -1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
Intensity
−0.8
−0.6
−0.4
−0.2
0.0
Symmetry

Q4

d)


Intensity
Symmetry
● 1.0
● -1.0