Assignment# 01

Submitted by : Mohaiminul Al Nahian
RIN **:** 662026703
Course# **:** CSCI 6100

Exercise 1.3

a) If **x(t)** is misclassified by **w(t)**, then $w^T(t)x(t)$ is negative when y(t) is positive and vice-versa. So, we may conclude that the product $y(t)\mathbf{w}^{\mathbf{T}}\mathbf{(t)x(t)}<0$

b) We know from the weight update rule that

$\mathbf{w(t+1)} = \mathbf{w(t)} + y(t)x(t)$

or, $\mathbf{w}^{\mathbf{T}}\mathbf{(t+1)} = \mathbf{w}^{\mathbf{T}}\mathbf{(t)} + y(t)\mathbf{x}^{\mathbf{T}}\mathbf{(t)}$

or, $y(t)\mathbf{w}^{\mathbf{T}}\mathbf{(t+1)x(t)} = y(t)\mathbf{w}^{\mathbf{T}}\mathbf{(t)x(t)} + |y(t)|^2\mathbf{x}^{\mathbf{T}}\mathbf{(t)x(t)}$

The term $|y(t)|^2\mathbf{x}^{\mathbf{T}}\mathbf{(t)x(t)}$ is definitely a positive quantity.

So, we may write, $y(t)\mathbf{w}^{\mathbf{T}}\mathbf{(t+1)x(t)} > y(t)\mathbf{w}^{\mathbf{T}}\mathbf{(t)x(t)}$

c) From (b) we see that

$$y(t)\mathbf{w}^{\mathbf{T}}\mathbf{(t+1)x(t)} > y(t)\mathbf{w}^{\mathbf{T}}\mathbf{(t)x(t)} \ldots\ldots\ldots\mathbf{(i)}$$

so, $y(t)\mathbf{w}^{\mathbf{T}}\mathbf{(t)x(t)}$ is increasing in each step to become $y(t)\mathbf{w}^{\mathbf{T}}\mathbf{(t+1)x(t)}$

For a misclassification, if y(t) is positive, then $\mathbf{w}^{\mathbf{T}}\mathbf{(t)x(t)}$ is negative. And $\mathbf{w}^{\mathbf{T}}\mathbf{(t)x(t)}$ is **increased** by the algorithm to move it towards **positive value** so that $y(t)\mathbf{w}^{\mathbf{T}}\mathbf{(t)x(t)}$ increases to become $y(t)\mathbf{w}^{\mathbf{T}}\mathbf{(t+1)x(t)}$**.**

Again, if y(t) is negative, then $\mathbf{w}^{\mathbf{T}}\mathbf{(t)x(t)}$ is positive. And $\mathbf{w}^{\mathbf{T}}\mathbf{(t)x(t)}$ is **decreased** by the algorithm to move it towards **negative value** so that $y(t)\mathbf{w}^{\mathbf{T}}\mathbf{(t)x(t)}$ increases to become $y(t)\mathbf{w}^{\mathbf{T}}\mathbf{(t+1)x(t)}$**.**

So, as far as classifying **x(t)** is concerned, the move from **w(t)** to **w(t+1)** is a move towards the right direction.

Exercise 1.5

a) Determining the age at which a particular test is performed should be a ‘Learning Approach’
b) Classifying numbers into prime and non-prime is a ‘Design Approach’
c) Detecting potential fraud in credit card charges involves a ‘Learning Approach’
d) Calculating the time for a falling objection reaching ground is a ‘Design Approach’

e) Determining the optical cycle for traffic light in a busy intersection should be a ‘learning approach’

Exercise 1.6

a) Recommending a book to a user in an online bookstore may involve ‘Supervised Learning’. The training data can be Book ratings by various users, User’s age, Favorite genre, Number of copies of the book being sold, Primary language of user, Previous search history in the website etc.
b) Learning to play tic-tac-toe may involve ‘Reinforcement Learning’. For reinforcement learning, every move can be graded as good or bad so that the subject eventually learns to maximize the output reward, that is winning.
c) Categorizing movies into different types may involve ‘Unsupervised learning’. Different properties of the movie such as visual content, sound effect, motion etc. can be identified to cluster the movies with similar properties together.
It may also involve ‘Supervised Learning’. The training data could be imdb rating, keywords in user reviews, language of the movie, date released, cast and crew etc.
d) Learning to play music can be ‘Reinforcement Learning’. Different sound produced can be graded as good or bad to ultimately pick the best synthesized music by the algorithm
e) Setting up credit limit involves ‘Supervised Learning’. The inputs can be customer’s salary, age, previous debt, account balance, credit score if any etc.

Exercise 1.7

a) In the dataset, 3 of the 5 outputs are ‘●’. So, learning algorithm picks the hypothesis where all 3 new points have output g as ‘●’ as it matches the dataset the most. 1 function agrees with it in all 3 points (f8), 3 functions agree with it in 2 points (f7, f6, f4), 3 functions agree on 1 point (f2, f3, f5) and f1 does not agree with it in any point.
b) The algorithm picks the hypothesis where all 3 new points have output g as ‘○’ because it matches the dataset the least. 1 function agrees with it in all 3 points (f1), 3 functions agree with it in 2 points (f2, f3, f5), 3 functions agree on 1 point (f4, f6, f7) and f8 does not agree with it in any point.
c) The XOR output can be paired as (101, ○), (110, ○) and (111, ●). Here 1 function agrees with it in all 3 points (f2), 3 functions agree with it in 2 points (f1, f4, f6), 3 functions agree on 1 point (f3, f5, f8) and f7 does not agree with it in any point.
d) The target function that agrees with all the training examples and disagrees the most with the XOR hypothesis (○,○,●) on the 3 new points is f7, which has outputs (●,●,○) on the three new points. So, algorithm picks g such that the output pair becomes (101, ●), (110, ●) and (111, ○).
Here 1 function agrees with it in all 3 points (f7), 3 functions agree with it in 2 points (f3, f5, f8), 3 functions agree on 1 point (f1, f4, f6) and f2 does not agree with it in any point.

Problem 1.1

Let $Black_1$ and $Black_2$ be events of the first ball being black and the second ball being black respectively.

Let, Bag_1 contains two black balls and Bag_2 contains one black and one white ball.

We need to find, probability of second ball being black provided first ball being black or $P(Black_2|Black_1)$

From Baye's theorem, $P(Black_2|Black_1) = \frac{P\ (Black1 \cap\ Black2)}{P(Black1)}$

Probability of choosing Bag_1 or $P(Bag_1)$= Probability of choosing Bag_2 or $P(Bag_2)$= ½

From Total Probability theorem,

$P(Black_1)= P(Black_1|Bag_1)P(Bag_1)+ P(Black_1|Bag_2)P(Bag_2)$

$= 1 \text{ x } ½ + ½ \text{ x } ½$

$= ½ + ¼$

$= ¾$

$P\ (Black_1 \cap Black_2)$ = Probability of first ball being black and second ball being black, which is only possible if Bag_1 is chosen.

So, $P\ (Black_1 \cap Black_2)= P(Bag_1)=$ ½

So, $P(Black_2|Black_1) = \frac{P\ (Black1 \cap\ Black2)}{P(Black1)} = \frac{\frac{1}{2}}{\frac{3}{4}} = \frac{2}{3}$ (Answer)

Problem 1.2

a) We have h(x)= sign($\mathbf{w^T x}$), where $\mathbf{w^T}=[w_{0,} w_{1,} w_2]$ and $\mathbf{x}= [x_{1,} x_{2,} x_3]^T$

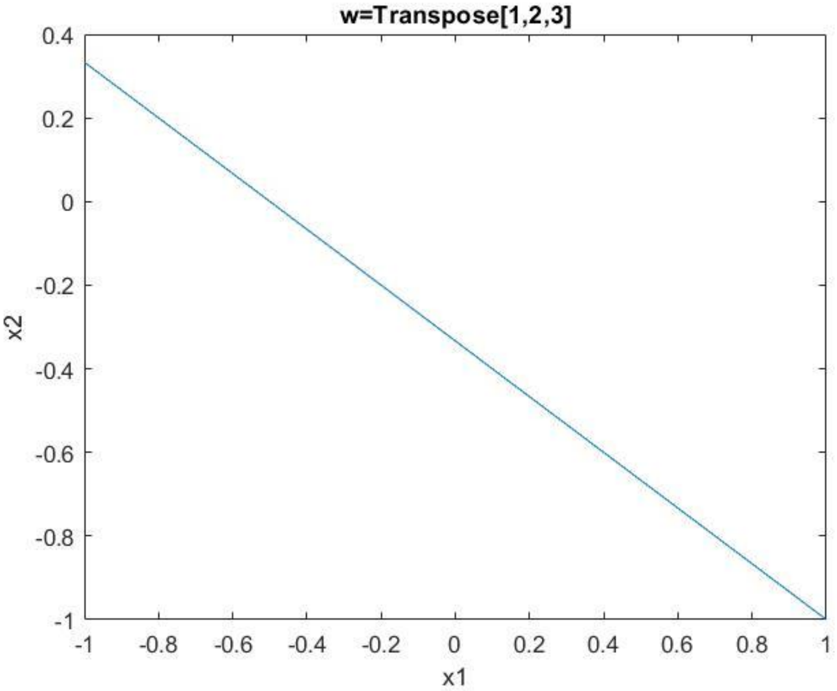
For, h(x)= +1, we have $\mathbf{w^T x}>0$ and for h(x)=-1, we have $\mathbf{w^T x}< 0$.
So, we may conclude that h(x) = +1 and h(x)= -1 are separated by a line $w^T x=0$
Or, $w_0 x_0 + w_1 x_1 + w_2 x_2 = 0$.

We may re-write, $w_2 x_2 = -w_1 x_1 - w_0 \text{ x } 1$

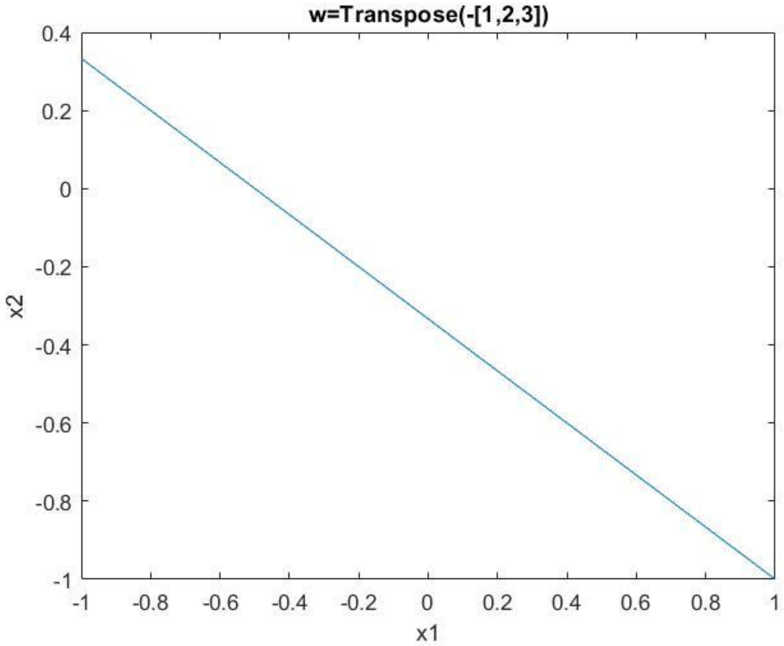
Or, $x_2 = (-w_1/w_2)\ x_1 + (-\ w_0/\ w_2) = ax_1 + b$

Where, a= (-w_1/w_2) and b = (- w_0/ w_2)

For w = $[1,2,3]^T$, we get



For w = -$[1,2,3]^T$, we get

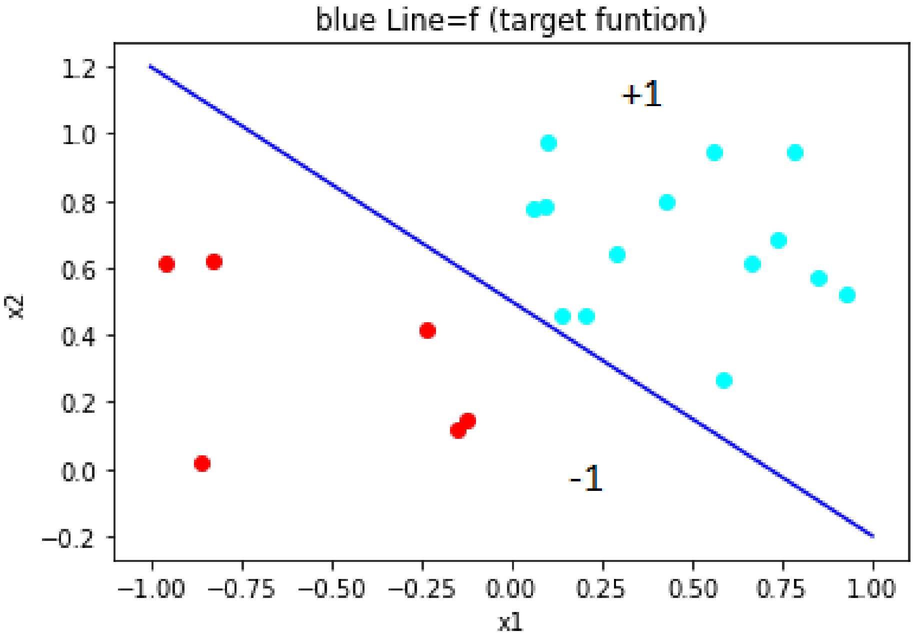


The graphs are identical, but for **w = [1, 2, 3]^T,** h(x)=+1 when $w_0x_0 + w_1x_1 + w_2x_2>0$, or $1+2x_1+3x_2 >0$. So, h(x)= +1 above the line and h(x)=-1 below the line.

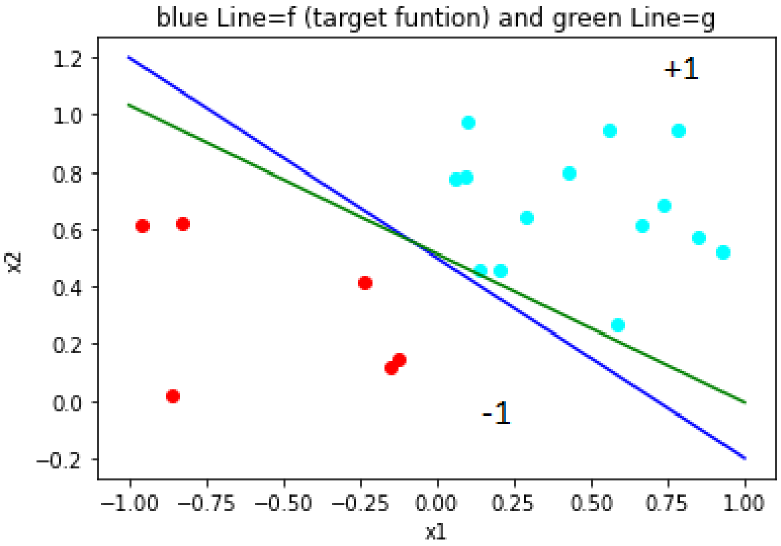
Whereas, for **w = -[1, 2, 3]^T,** h(x)=+1 when $w_0x_0 + w_1x_1 + w_2x_2>0$, or $-1-2x_1-3x_2 >0$ or $1+2x_1+3x_2 <0$. So, h(x)= +1 below the line and h(x)=-1 above the line.

Problem 1.4

a) For Dataset size of **20**, we get the following plot



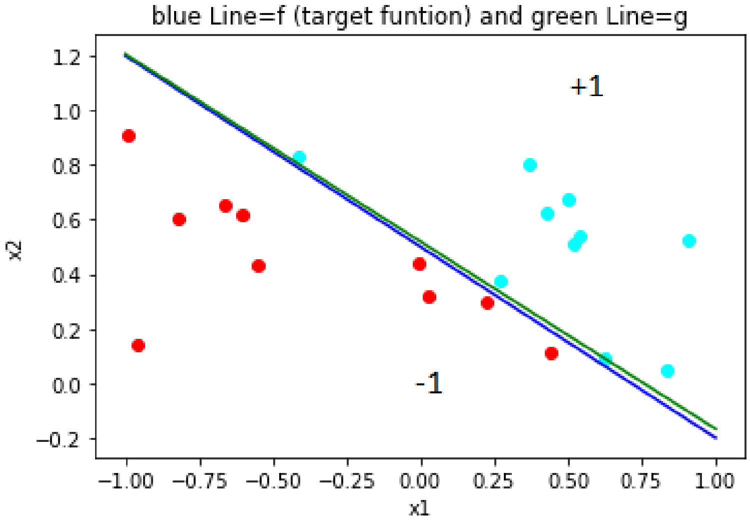
b) Perceptron Learning algorithm with dataset size 20 gives us the following plot



It took 7 iterations to converge with final **w** = `[-1, 1.01154237, 1.94561834]`

f is not very close to g

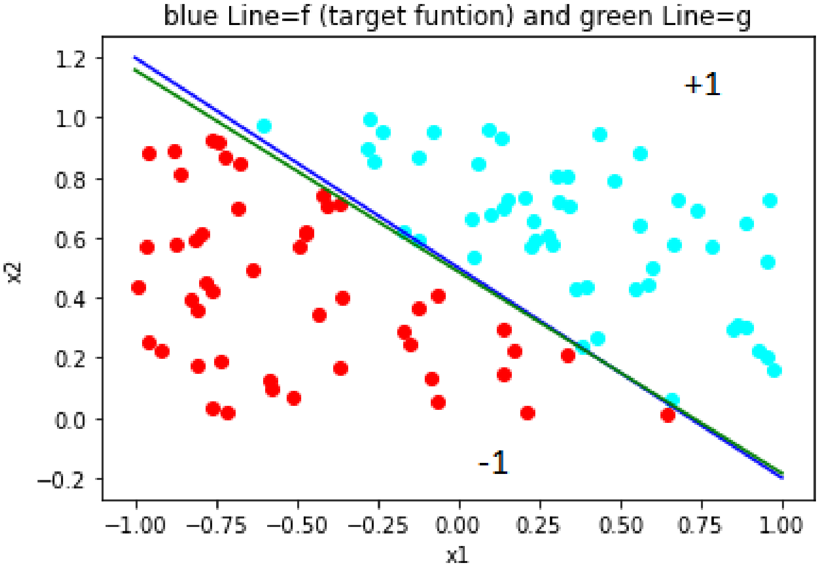
c) The PLA run on another random dataset of 20 gives the following plot



Now, some +1 and -1 valued points are very close to each other, so it took 141 iterations to converge with final w = `[-3, 3.96702165, 5.77749315]`

Here, g is closer to f than it was in b

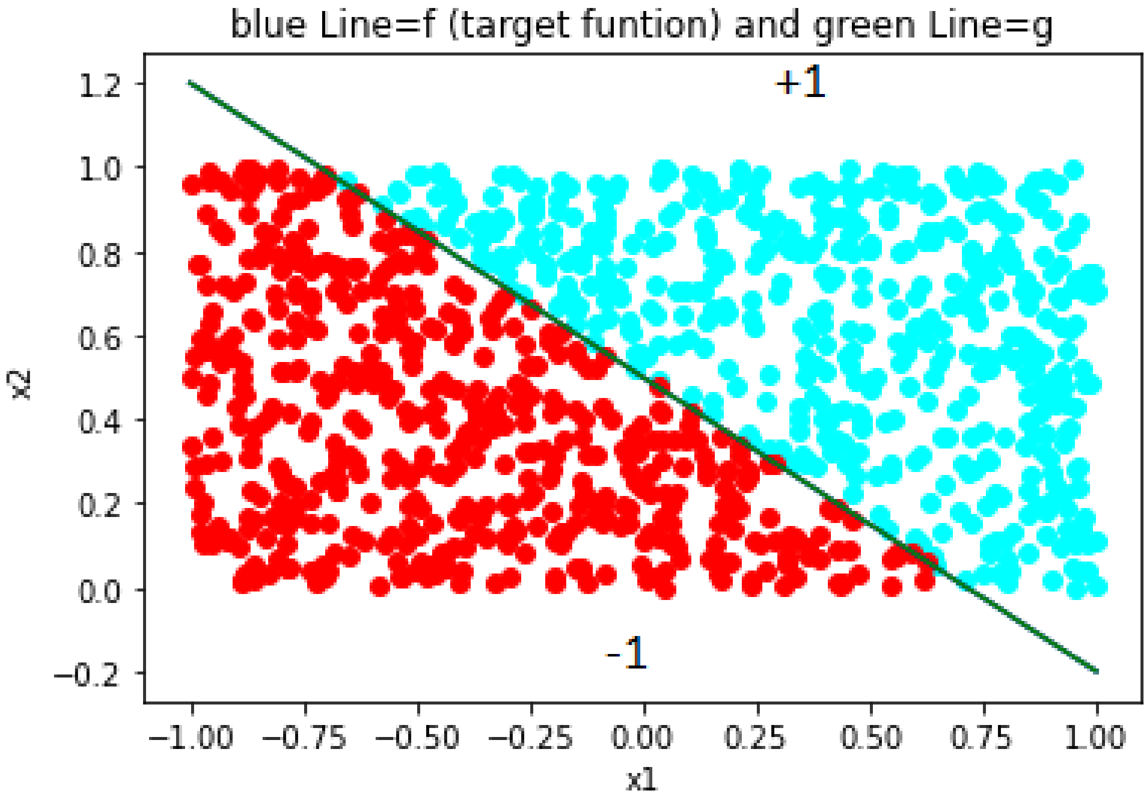
d) The PLA run on a random dataset of 100 gives the following plot



it took 83 iterations to converge with final w = `[-3.     4.13820767    6.16010824]`

Here, g is closer to f than it was in b

e) The PLA run on a random dataset of 1000 gives the following plot



it took 1207 iterations to converge with final w = `[-13.       18.1950969    25.99384041]`

Here, g is very close to f than it was in b. The two lines are so close that it is difficult to distinguish between f and g by looking.