LFD Problem Set 1

John Cohen

September 9, 2024

Exercise 1.3

(The weight update rule in (1.3) has the nice interpretation that is moves in the direction of classifying x(t) correctly.)

(a) Show that $y(t)w^T(t)x(t) < 0$ [*Hint: $x(t)$ is misclassified by $w(t)$*]

Proof. In the case of a misclassification, $Sign(w(t)^T x(t)) \neq y(t)$
The product of two opposite sign numbers, $Sign(w(t)^T x(t))$ and $y(t)$, is negative
Therefore: $y(t)w(t)^T x(t) < 0$ □

(b) Show that $y(t)w^T(t+t)x(t) > y(t)w^T(t)x(t)$ [*Hint: Use (1.3)*]

Proof. The update rule (1.3) gives:

$$w(t+1) = w(t) + y(t)x(t)$$

Dot product $x(t)$ on each side maintaining equivalence:

$$w^T(t+1)x(t) = w^T(t)x(t) + y(t)x^T(t)x(t)$$

Multiplying $y(t)$ on both sides:

$$y(t)w^T(t+1)x(t) = y(t)w^T(t)x(t) + y(t)^2 x^T(t)x(t)$$

Because $y(t)$ is either 1 or -1, $y(t)^2$ must be 1. Simplifying:

$$y(t)w^T(t+1)x(t) = y(t)w^T(t)x(t) + x^T(t)x(t)$$

Given x_0 is 1, $x^T(t)x(t) > 0$, we conclude that:

$$y(t)w^T(t+1)x(t) > y(t)w^T(t)x(t)$$

□

(c) As far as classifying $x(t)$ is concerned, argue that the move from $w(t)$ to $w(t+1)$ is a move 'in the right direction'.

Proof. As seen in part (a) a misclassification is negative:

$$y(t)w^T(t)x(t) < 0$$

As seen in part (b) each iteration moves the product of $w^T(t+1)x(t)$ in a more positive direction:

$$y(t)w^T(t+1)x(t) > y(t)w^T(t)x(t)$$

The update clearly increases $w^T(t+1)x(t)$ from $w^T(t)x(t)$. This improves the margin classifying $x(t)$. Each update is getting $w(t)$ closer to the optimal vector properly separating the Dataset. This is a move in the right direction. □

Exercise 1.5

(Which of the following problems are more suited for the learning approach and which are more suited for the design approach)

(a) Determining the age at which a particular medical test should be performed: Learning

(b) Classifying numbers into primes and non-primes: Design

(c) Detecting potential fraud in credit card charges: Learning

(d) Determining the time it would take a falling object to hit the ground: Design

(e) Determining the optimal cycle for traffic lights in a busy intersection: Learning

Exercise 1.6

(For each of the following tasks, identify which type of learning is involved (supervised, reinforcement, or unsupervised) and the training data to be used. If a task can fit more than one type, explain how and describe the training data for each type.)

(a) Recommending a book to a user in an online bookstore
Supervised or Unsupervised learning
Recommendation systems can be trained using either. Supervised with user raitings and feedback would label data. Unsupervised learning could be leveraged to cluster similar books and/or customers without and given labels.

(b) Playing tic-tac-toe
Reinforcement learning
Reinforcement learning is especially useful for learning to play games where it is hard to determine the exact outcome of choosing one move over another. All one needs to do is take execute some action and report how well things went. There is your training example. Trial and error, receiving rewards or penalties based on the outcome of each move. The training data would likely be the board situation with assessments on the outcome.

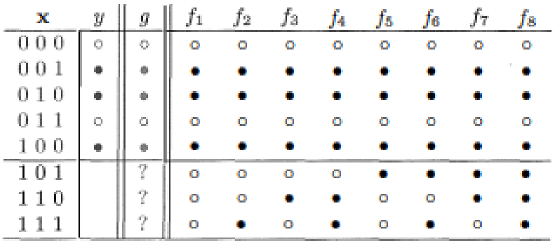
(c) Categorizing movies into different types
Unsupervised learning
This involves clustering movies into categories without any labels based on their features which could be genres, ratings, release date, references, actors, etc.

(d) Learning to play music
Reinforcement learning
Trial and error, receiving feedback based on the how well it is playing the music (practice and feedback). The training data will be the previous attempts and the feedback associated with each attempt.

(e) Credit limit: Deciding the maximum allowed debt for each bank customer
Supervised learning
The training data may include customer credit histories, limits, etc. labels with whether previous limits harmed or benefited the bank.

Exercise 1.7

(For each of the following learning scenarios in the problem, evaluate the performance of g on the three points in X outside D. To measure the performance, compute how many of the 8 possible target functions agree with g on all three points, on two of them, on one of them, and on none of them.)



x	y	g	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
0 0 0	○	○	○	○	○	○	○	○	○	○
0 0 1	●	●	●	●	●	●	●	●	●	●
0 1 0	●	●	●	●	●	●	●	●	●	●
0 1 1	○	○	○	○	○	○	○	○	○	○
1 0 0	●	●	●	●	●	●	●	●	●	●
1 0 1		?	○	○	○	○	●	●	●	●
1 1 0		?	○	○	●	●	○	○	●	●
1 1 1		?	○	●	○	●	○	●	○	●

Figure 1: Let empty circle be -1 and full circle be +1.

(a) H has only two hypotheses, one that always returns +1 and one that always returns -1. The learning algorithm picks the hypothesis that matches the data set the most.

Let $h_1 = +1$ and $h_2 = -1$
In other words $g \epsilon H = \{h_1, h_2\}$
y(x) gives 3's +1 and 2 -1's.
$E_{in}(h_1) = \frac{2}{5}$ that is 2 wrong, 3 right of the 5 examples.
Conversely $E_{in}(h_2) = \frac{3}{5}$ that is 3 wrong, 2 right of the 5 examples.
Given: the learning algorithm picks the hypothesis that matches the data set the most or in other-words, the learning algorithm picks the hypothesis with the lowest E_{in}
$E_{in}(h_1)$ is the lowest so by the given above statement, we conclude h_1 is selected as g, g = h_1.
With this and the table above:
g(101, 110, 111) = +1, +1, +1
1 f_8 agrees with g on all three points.
3 f_4, f_6, f_7 agree with g on two of them.
3 f_2, f_3, f_5 agree with g on one of them
1 f_1 agrees with g on none of them.

(b) The same H, but the learning algorithm now picks the hypothesis that matches the data set the least.

Let $h_1 = +1$ and $h_2 = -1$
In other words $g \epsilon H = \{h_1, h_2\}$
y(x) gives 3's +1 and 2 -1's.
$E_{in}(h_1) = \frac{3}{5}$ that is 3 wrong, 2 right of the 5 examples.
Conversely $E_{in}(h_2) = \frac{2}{5}$ that is 2 wrong, 3 right of the 5 examples.
Given: the learning algorithm picks the hypothesis that matches the data set the most or in other-words, the learning algorithm picks the hypothesis with the lowest E_{in}
$E_{in}(h_2)$ is the lowest so by the given above statement, we conclude h_2 is selected as g, g = h_2.
With this and the table above:

g(101, 110, 111) = -1, -1, -1
1 f_1 agrees with g on all three points.
3 f_2, f_3, f_5 agree with g on two of them.
3 f_4, f_6, f_7 agree with g on one of them
1 f_8 agrees with g on none of them.
With this and the table above:

(c) H = XOR (only one hypothesis which is always picked), where XOR is defined by XOR(x) = 1 if the number of 1's in x is odd and XOR(x) = 0 if the number is even.
Given: $g\epsilon H = \{XOR\}$
Thus $g = \{XOR\}$
g(101, 110, 111) = -1, -1, +1
1 f_2 agrees with g on all three points.
3 f_1, f_4, f_6 agree with g on two of them.
3 f_3, f_5, f_8 agree with g on one of them
1 f_7 agrees with g on none of them.

(d) H contains all possible hypotheses (all Boolean functions on three variables), and the learning algorithm picks the hypothesis that agrees with all training examples but otherwise disagrees the most with the XOR.
Let G be the set of all 3-digit binary strings that make up our hypothesis space. Each binary string corresponds to an input x such that g(x) = +1
H contains all possible hypotheses (all Boolean functions on three variables): $g\epsilon H = \{h_1, ..., h_256\}$
"The learning algorithm picks the hypothesis that agrees with all training examples but otherwise disagrees the most with the XOR." The way I interpret that is the goal is to choose a hypothesis g(x) such that:
1. Inside the training set D, g(x) = y(x) - meaning it perfectly fits the training data.
2. Outside the training set D, g(x) = -XOR(x) - meaning it disagrees with XOR as much as possible.
In constructing G, we add all three digit binary strings that disagree with XOR, remove those that disagree with y, and add those that agree with y. This asserts that no x disagrees with y and includes those remaining that disagree with XOR
G = x where XOR(x)=-1 - x where y(x)=-1 + x where y(x)=+1 .
G = 000,011,101,110 - 000,011 + 001,010,100 = 001,010,100,101,110
Selecting g(x) from H using the set G, we can now define the hypothesis g(x):
g(101) = +1
g(110) = +1
g(111) = -1 (since 111 disagrees with XOR outside of the training set).
1 f_7 agrees with g on all three points.
3 f_3, f_5, f_8 agree with g on two of them.

3 f_1, f_4, f_6 agree with g on one of them
1 f_2 agrees with g on none of them.

Proof. (Type your proof here.) □

Problem 1.1

(We have 2 opaque bags, each containing 2 balls. One bag has 2 black balls and the other has a black and a white ball. You pick a bag at random and then pick one of the balls in that bag at random. When you look at the ball it is black. You now pick the second ball from that same bag. What is the probability that the ball is also black? [*Hint: Use Bayes' Theorem:* $P(A \cap B) = P(A \mid B)P(B) = P(B \mid A)P(A)$])

Let A be the event the first ball picked is black
Let B_1 be the event the first bag was chosen
Let B_2 be the event the second bag was chosen

$$P[A \mid B_1] = 1$$

$$P[A \mid B_2] = \frac{1}{2}$$

$$P[B_2] = \frac{1}{2}$$

$$P[B_1] = \frac{1}{2}$$

$$P[blacksecond \mid B_1] = 1$$

$$P[blacksecond \mid B_2] = 0$$

$$P[A] = P(A \cap B_1) + P(A \cap B_2) = P[A \mid B_1] * P[B_1] + P[A \mid B_2]P[B_2] = 1 * \frac{1}{2} + \frac{1}{2} * \frac{1}{2} = \frac{3}{4}$$

$$P[A \cap B_1] = P[B_1 \mid A] * P[A] = P[A \mid B_1] * P[B_1]$$

$$P[B_1 \mid A] = \frac{P[A \mid B_1] * P[B_1]}{P[A]} = \frac{1 * \frac{1}{2}}{\frac{3}{4}} = \frac{2}{3}$$

$$P[A \cap B_2] = P[B_2 \mid A] * P[A] = P[A \mid B_2] * P[B_2]$$

$$P[B_2 \mid A] = \frac{P[A \mid B_2] * P[B_2]}{P[A]} = \frac{\frac{1}{2} * \frac{1}{2}}{\frac{3}{4}} = 1/3$$

$$P[blacksecond \mid A] = P[blacksecond \mid B_1]{*}P[B_1 \mid A]{+}P[blacksecond \mid B_2]{*}P[B_2 \mid A] = 1{*}\frac{2}{3}{+}0{*}\frac{1}{3} = \frac{2}{3}$$

Problem 1.2

(Consider the perceptron in two dimensions: $h(x) = \text{sign}(w^T x)$ where $w = [w_0, w_1, w_2]$ and $x = [1, x_1, x_2]^T$. Technically, x has three coordinates, but we call this perceptron two-dimensional because the first coordinate is fixed at 1.)

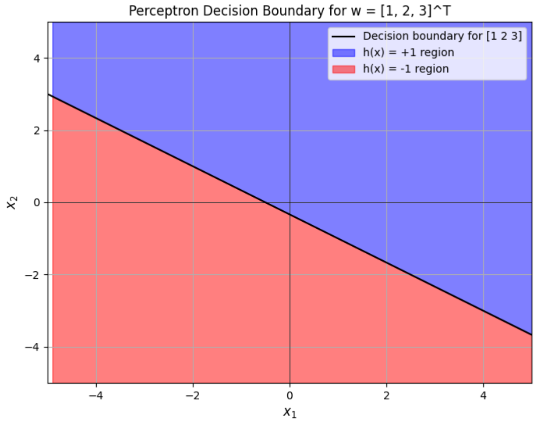
(a) Show that the regions on the plane where $h(x) = +1$ and $h(x) = -1$ are separated by a line. If we express this line by the equation $x_2 = ax_1 + b$, what are the slope a and the intercept b in terms of w_0, w_1, w_2?

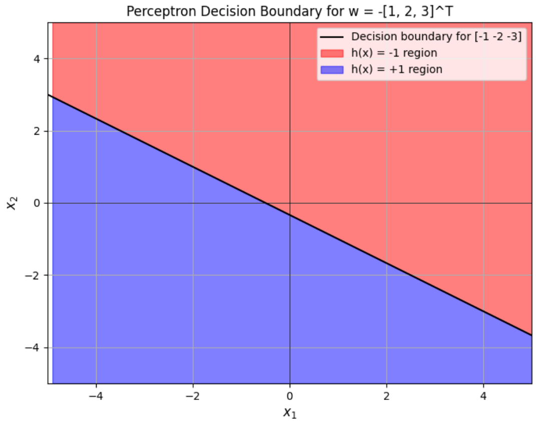
Proof. (The Regions where h(x) = +1 and h(x) = -1 Are Separated by a Line)
Perceptron Hypothesis: $h(x) = \text{sign}(w^T x)$
where $w = [w_0, w_1, w_2]^T$ is the weight vector, $x = [x_0, x_1, x_2]^T$ the input vector, and $w^T x$ is the dot product between the two.
h(x) = +1 when $w^T x > 0$ and h(x) = -1 when $w^T x < 0$
$w^T x = 0$ is the point where the perceptron switches classification from -1 to +1 and vice versa. That means the boundary between the regions of a two-dimensional space is represented by this $w^T x = 0$ which simplifies to $w_0 + w_1 x_1 + w_2 x_2 = 0$. This boundary is clearly a linear equation that represents a line in two-dimensional space. Therefore, we conclude that the regions where h(x) = +1 and h(x) = -1 are in fact separated by a line. □

$w_0 + w_1 x_1 + w_2 x_2 = 0$
$w_2 x_2 = -w_0 - w_1 x_1$
$x_2 = -\frac{w_1}{w_2} x_1 - \frac{w_0}{w_2}$

The slope a = $-\frac{w_1}{w_2}$
The intercept b = $-\frac{w_0}{w_2}$

(b) Draw a picture for the cases $w = [1, 2, 3]^T$ and $w = -[1, 2, 3]^T$.





Problem 1.4

(In exercise 1.4, we use an artificial data set to study the perceptron learning algorithm. This problem leads you to explore the algorithm further with data sets of different sizes and dimensions)

(a) Generate a linearly separable data set of size 20 as indicated in Exercise 1.4. Plot the examples $\{(x_n, y_n)\}$ as well as the target function f on a plane. Be sure to mark the examples from different classes differently and add labels to the axes of the plot.

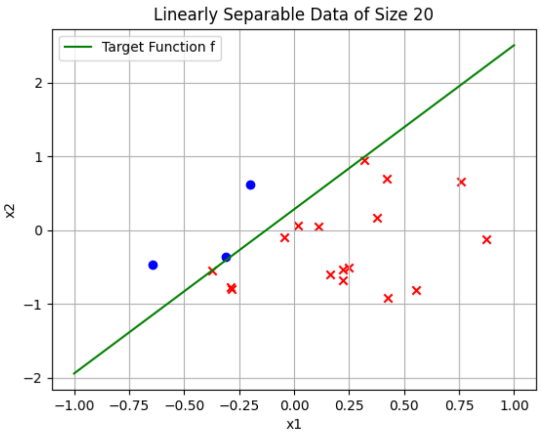


Figure 2: Target, accuracy is always 1

(b) Run the perceptron learning algorithm on the data set above. Report the number of updates that the algorithm takes before converging. Plot the examples $\{(x_n, y_n)\}$, the target function f, and the final hypothesis g in the same figure. Comment on whether f is close to g.

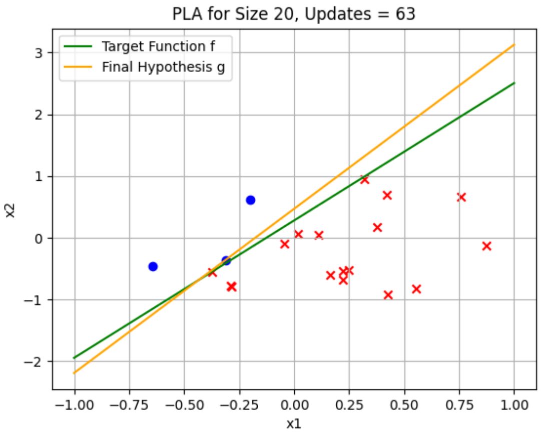


Figure 3: Accuracy of learned hypothesis g compared to target function f (size 20): 0.97

(c) Repeat everything in (b) with another randomly generated data set of size 20. Compare your results with (b).

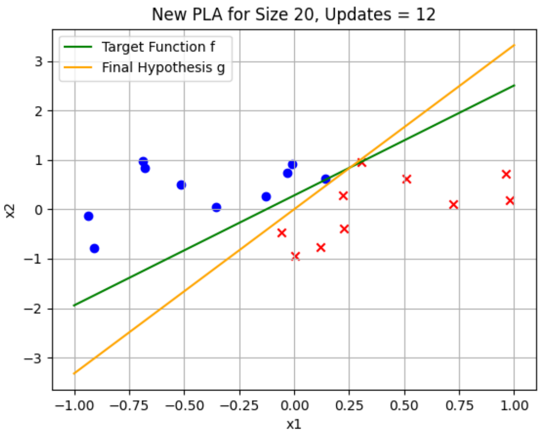


Figure 4: Accuracy of learned hypothesis g compared to target function f (size 20): 0.94

(d) Repeat everything in (b) with another randomly generated data set of size 100. Compare your results with (b).

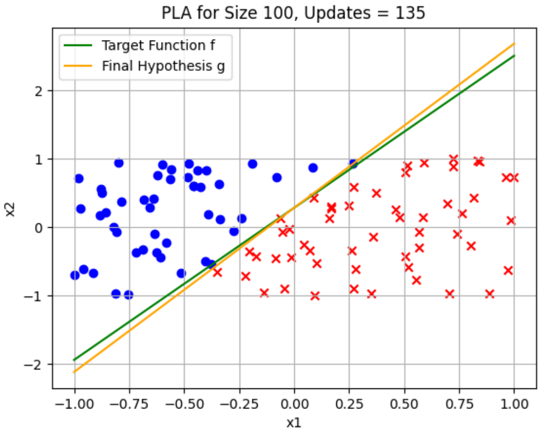


Figure 5: Accuracy of learned hypothesis g compared to target function f (size 100): 0.99

(e) Repeat everything in (b) with another randomly generated data set of size 1,000. Compare your results with (b).

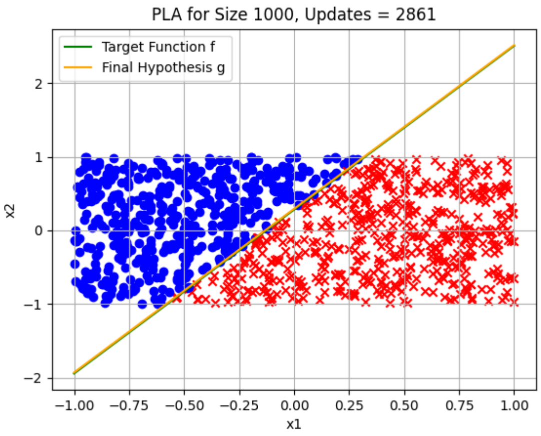


Figure 6: Accuracy of learned hypothesis g compared to target function f (size 1000): 1.00

NOTE: The closer the accuracy is to 1, the closer the hypothesis g is to the target function f. These accuracies are only tested accuracies, and may be different if tested again, however with a large set of 1000, it is exponentially rare for new accuracies to deviate by any significant amount. So, we continue with a degree of certainty. The accuracy of a learned hypothesis g compared to the target function f for each graph was calculated by randomly generating a testing dataset D of 1000 entries and dividing the total number of agreed data point classifications between the final hypothesis g and target function f by 1000 ($|D|$, the number of total data points). That is:

$$\frac{1}{|D|}\sum_{i=1}^{|D|}[[g(x_i) = f(x_i)]]$$