Robt 407 Zarema Balgabekova HW 1

A) Problems from the TextBook

0. 1 Learning Exercises

1. Exercise 1. 1

a) Input space $\mathcal{X}$: medical history (diseases patients had in the past, vaccinations, methods of treatment ...), current symptoms (body temperature, blood pressure , cough, stomach ache ...), patient's personal information (for some diseases, it is necessary to know the person's age, gender ...)

Output space $\mathcal{Y}$: medical diagnoses (possible diseases or that a person is healthy)

Target function $f : \mathcal{X} \to \mathcal{Y}$: an ideal formula to make a diagnosis based on inputs

Specifics of the data set : medical history, symptoms and personal information should be reliable, detailed and correctly matched with medical diagnosis

b) Input space $\mathcal{X}$: set of handwritten digits by different people (images)

Output space $\mathcal{Y}$: combinations of digits (0-9)

Target function $f: X \to \mathcal{Y}$: an ideal formula for handwritten digit recognition based on their features

Specifics of the data set: handwritten digits should be clearly written and cover different handwriting styles to recognize them

c) Input space X: useful information about emails (whether they contain specific words, who has sent them, at what time they were sent...)

Output space $\mathcal{Y}$: spam/not spam (yes/no, +1/-1)

Target function $f: X \to \mathcal{Y}$: an ideal formula to determine if the email is spam or not based on inputs

Specifics of the data set: spam and not spam emails should be identified by human beings; some information about emails should be updated (for example, senders of spam emails)

d) Input space $\mathcal{X}$: set of prices of electric load, temperatures, and days of the week

Output space $\mathcal{Y}$: possible values of electric load

Target function $f: \mathcal{X} \to \mathcal{Y}$: ideal formula that predicts how an electric load varies with price, temperature, and day of the week

Specifics of the data set: the values of an electric load should be accurately measured for different prices, temperatures, and days of the week

e) Input space $\mathcal{X}$: all available relevant to the problem data

Output space $\mathcal{Y}$: possible results in the appropriate to the problem form

Target function $f: X \to \mathcal{Y}$: an ideal formula to construct an empirical solution based on available data

Specifics of the data set: data should be relevant to the problem and accurate, as well as sufficient to construct an empirical solution

2, Exercise 1.2

a) Keywords that will end up with a large positive weight in the perceptron are the words that often occur in spam messages.
For example, 'free', 'Bonus', 'bargain', 'prize', 'order', 'now' ...

b) Keywords that will get a negative weight do not often occur in spam messages.
For example, 'university', 'schedule', 'meeting', 'information', 'office', 'RDBT 407' ...

c) The parameter in the perceptron that directly affects how many border-line messages end up being classified as spam is a threshold, according to which messages are classified as spam and not spam

3. Exercise 1.5

a) Learning approach because the target function is unknown and we need data to match medical tests with ages

b) Design approach because the problem is well specified and there is an analytical way to derive a target function

c) Learning approach because the problem is less specified, the target function is unknown => we need data of credit card charges to detect potential fraud

d) Design approach because the task is well specified and we know the formula to determine the time

e) Design approach if the problem is well specified (we are given the number of cars, the number of people passing through the intersection in a particular time interval, their velocities ...)
It also could be learning approach if we are given a bunch of data with some inputs and corresponding optimal cycles

4. Exercise 1.6

a) Supervised learning because we need to know previous book ratings by a user as output. Training data may include characteristics of books as input and ratings by a user as output

b) Reinforcement learning because it is especially useful for learning how to play a game. Training data may be some actions and how well they are.
Unsupervised learning if you just watch how other people play this game. Training data is their actions.

c) Supervised learning if types of movies are given as labels. Training data may include features of movies as inputs and their types as outputs.

Unsupervised learning if types of movies are not provided. Training data may be just different characteristics of movies as inputs

d) Reinforcement learning if there is a measure of how good your music is. Training data may include your music and its grades by listeners.

Unsupervised learning if you listen to or watch how other people play music. Training data is the sound of music or musicians' actions

e) Supervised learning since we need to know the relationship between customers' information and credit limits in the past. Training data is customers' profiles as inputs and credit limits as outputs.

0.2 Perceptron learning Algorithm

1. Exercise 1.3-a

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

From hint, we see that if $x(t)$ is misclassified, we have two possible cases:

Case 1

if $y(t) = +1$, $\text{sign}(w^T(t)\, x(t)) = -1$ $(\neq y(t))$

$\Rightarrow y(t)\, w^T(t)\, x(t)$ = 'positive'·'negative' = 'negative' < 0

Case 2

if $y(t) = -1$, $\text{sign}(w^T(t)\, x(t)) = +1$

$\Rightarrow y(t)\, w^T(t)\, x(t)$ = 'negative'·'positive' = 'negative' < 0

2. Exercise 1.3-b

Show that $y(t)\, w^T(t+1)\, x(t) > y(t)\, w^T(t)\, x(t)$ [Hint: use (1.3)]

$$w(t+1) = w(t) + y(t)\, x(t) \qquad (1.3)$$

$$y(t)\, w^T(t+1)\, x(t) = y(t)\,[\, w(t) + y(t)\, x(t)\,]^T\, x(t)$$

Since transpose is a linear operator,

$$y(t)\, w^T(t+1)\, x(t) = y(t)\,[\, w^T(t) + y(t)\, x^T(t)\,]\, x(t) =$$

$$y(t)\, w^T(t)\, x(t) + y^2(t)\, x^T(t)\, x(t)$$

As $y(t) = 1$ or -1,

$$y(t)\, w^T(t+1)\, x(t) = y(t)\, w^T(t)\, x(t) + \underbrace{x^T(t)\, x(t)}_{\text{positive}}$$ (x_0 is always one)

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

3. Exercise 1.3-C

Argue that the move from $w(t)$ to $w(t+1)$ is a move 'in the right direction'

We can use geometrical explanation.

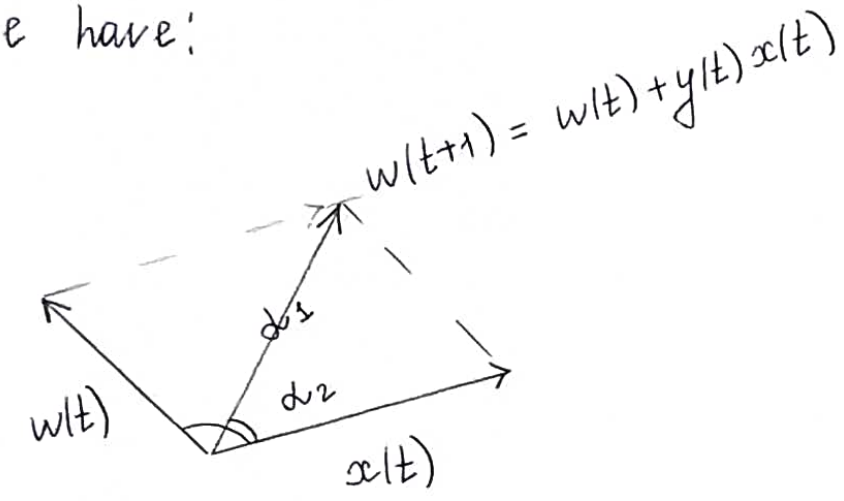
Dot product : $w^T x = \|w^T\|\,\|x\| \cdot \cos\alpha$

$\Rightarrow$ sign $(w^T x)$ depends on $\cos\alpha$ (α is the angle between w and x)

We know that $\cos\alpha$ is positive if $0 \leqslant \alpha < 90°$

$\cos\alpha$ is negative if $90° < \alpha \leqslant 180°$

If $y(t) = +1$ and $x(t)$ is misclassified by $w(t)$, we have:



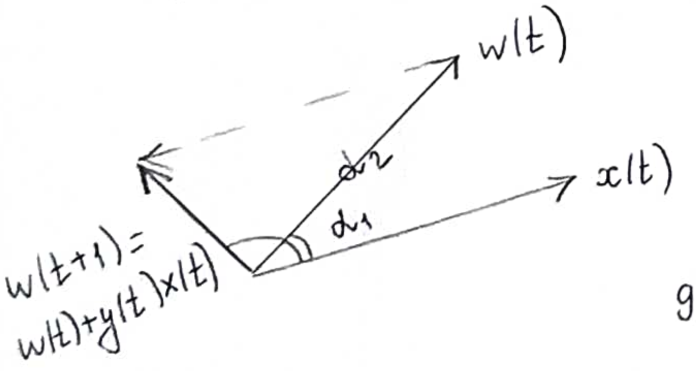
$90° < \alpha_1 < 180°$ => $\cos(\alpha_1)$ is negative =>

$\text{sign}(w^T(t) \cdot x(t)) = -1 \neq y(t)$

$0 < \alpha_2 < 90°$ => $\cos(\alpha_2)$ is positive =>

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

If $y(t) = -1$ and $x(t)$ is misclassified by $w(t)$, we have:



$0 < \alpha_1 < 90°$ => $\cos(\alpha_1)$ is positive =>

$\text{sign}(w^T(t)\, x(t)) = 1 \neq y(t)$

$90° < \alpha_2 < 180°$ => $\cos(\alpha_2)$ is negative =>

$\text{sign}(w^T(t+1)\, x(t)) = -1 = y(t)$

=> The move is 'in the right direction'

0.3 Experiments with Perceptron Learning Algorithm

1.

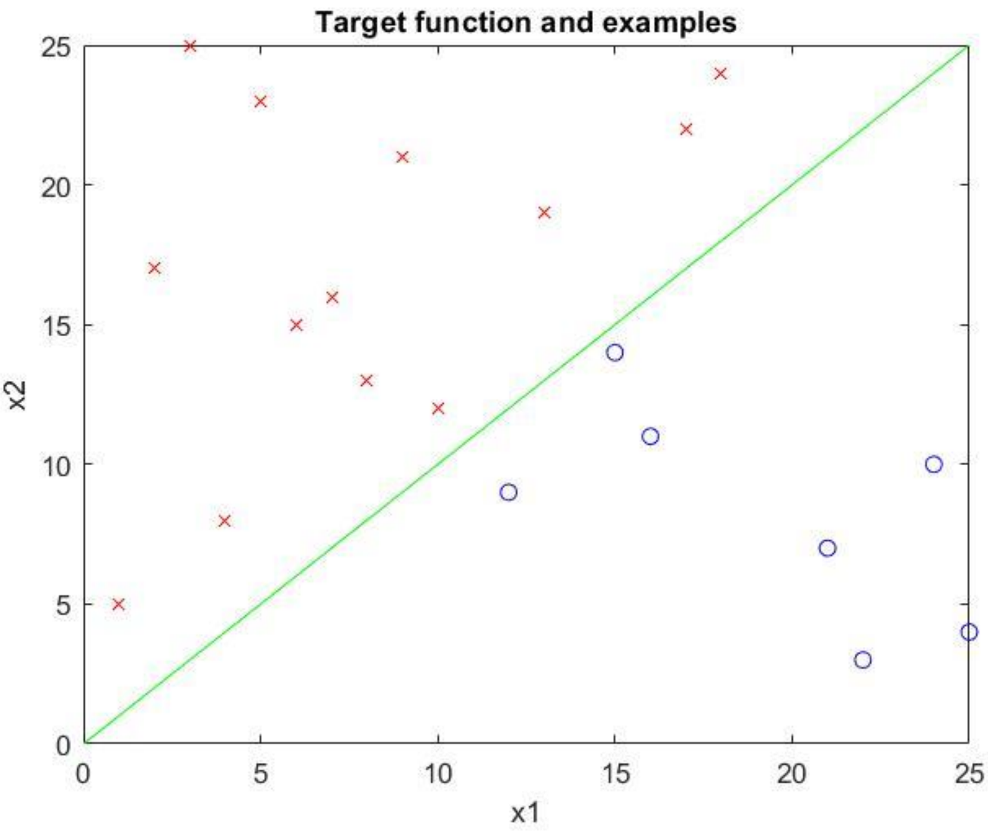


Figure 1: Target function and 20 marked examples

In this case, the target function is chosen to be $x_2 = x_1$ (a = 1, b = 0). Examples above the target line are marked with red circles (label = -1) while examples below the line are marked as blue circles (label = +1). If there are examples that lie on the target line, they are also marked as blue circles by default. There are no such examples in Figure 1.

2.

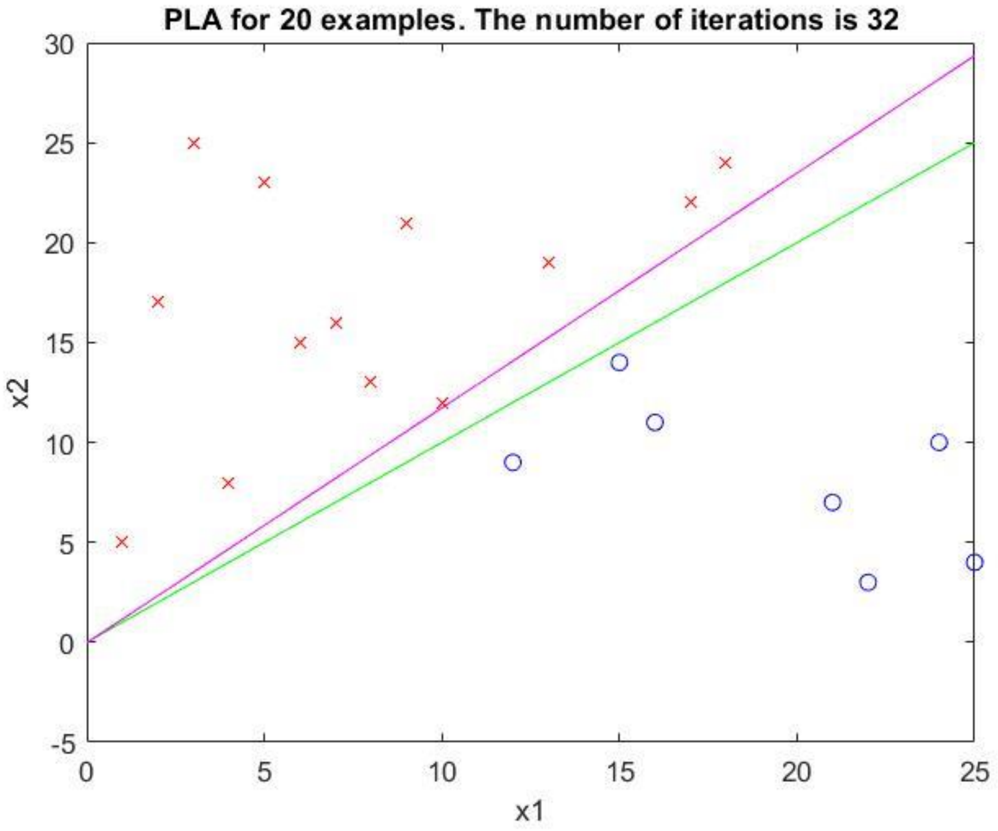


Figure 2: PLA for 20 examples (Trial 1)

Figure 2 presents PLA for examples in Figure 1. The number of updates that PLA takes before converging is 32. Final hypothesis g is colored with magenta. It is seen that though it slightly differs from the target function (green), it separates labelled examples correctly.

3.

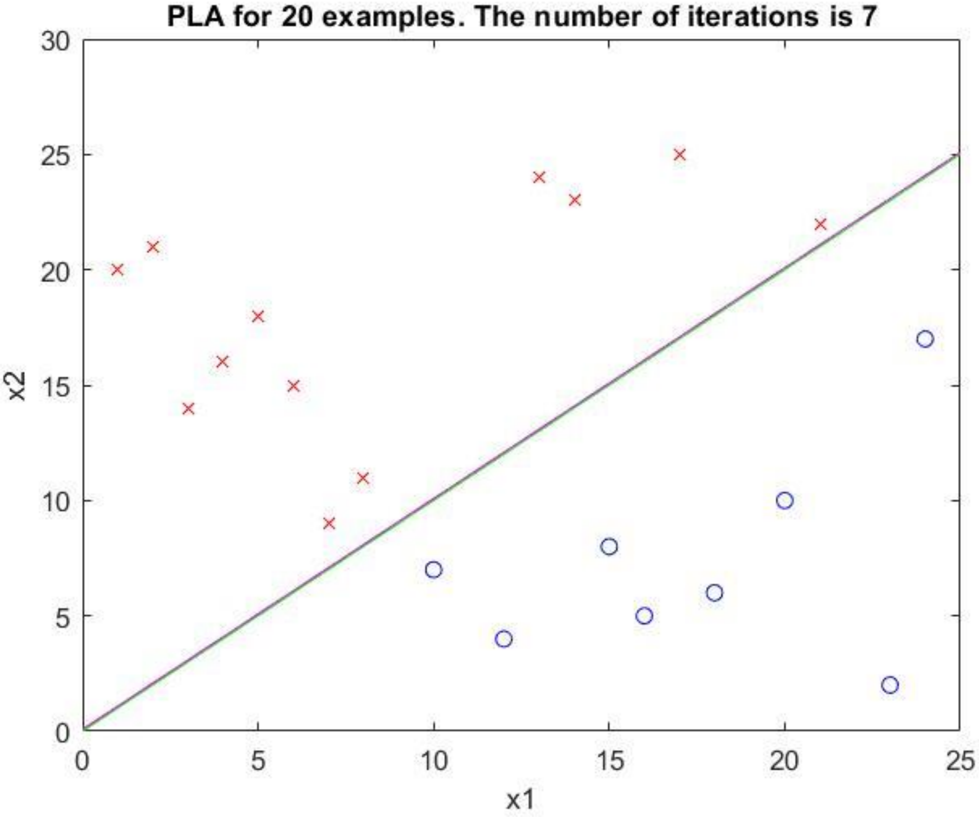


Figure 3: PLA for 20 examples (Trial 2)

Running PLA with another data set of 20 examples, we obtain a quite different result from Trial 1. This time, the number of iterations reduced to 7. Furthermore, the final hypothesis and the target function seem to coincide. This means that the performance of PLA depends on a data set.

4.

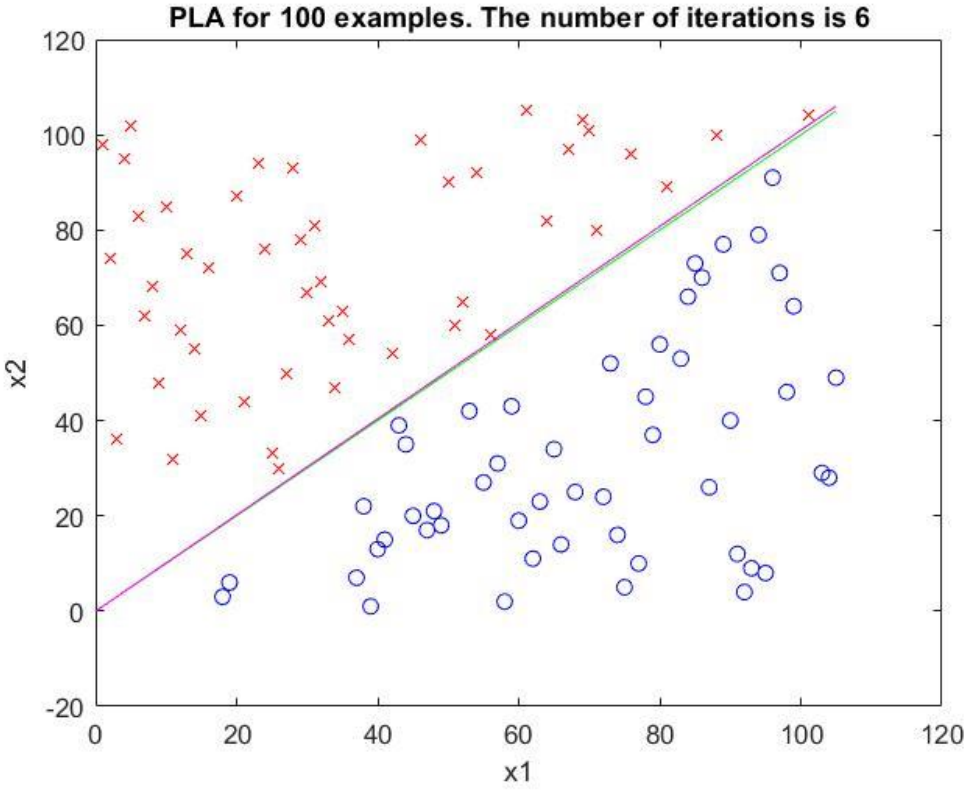


Figure 4: PLA for 100 examples

Surprisingly, after increasing the number of examples, we again obtain a better result than in Figure 2. The number of updates before converging is 6, which is even smaller by 1 than for Trial 2 with 20 examples. Likewise, the final hypothesis and the target function almost coincide. This means that PLA does not depend on the number of examples but on the distribution of data itself.

5.

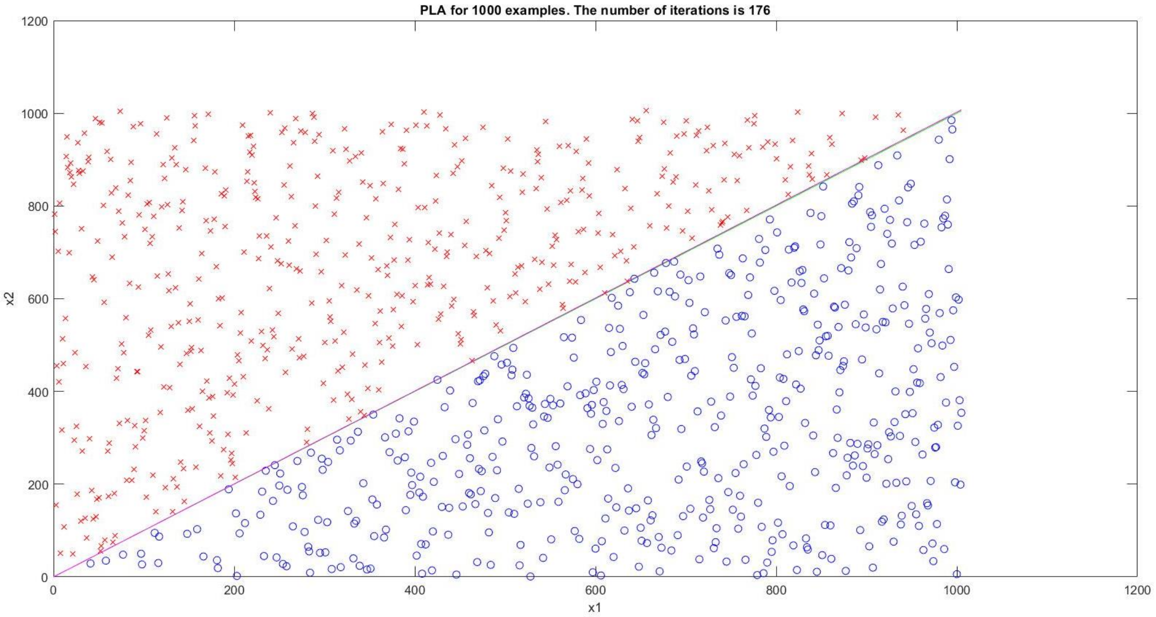


Figure 5: PLA for 1000 examples

When the number of examples increased to 1000, the number of updates became 176, which is larger than for 20 examples (Trial 1). However, the final hypothesis and the target function are almost the same. The reason for the large number of updates can be the crowdedness of points near the target line.

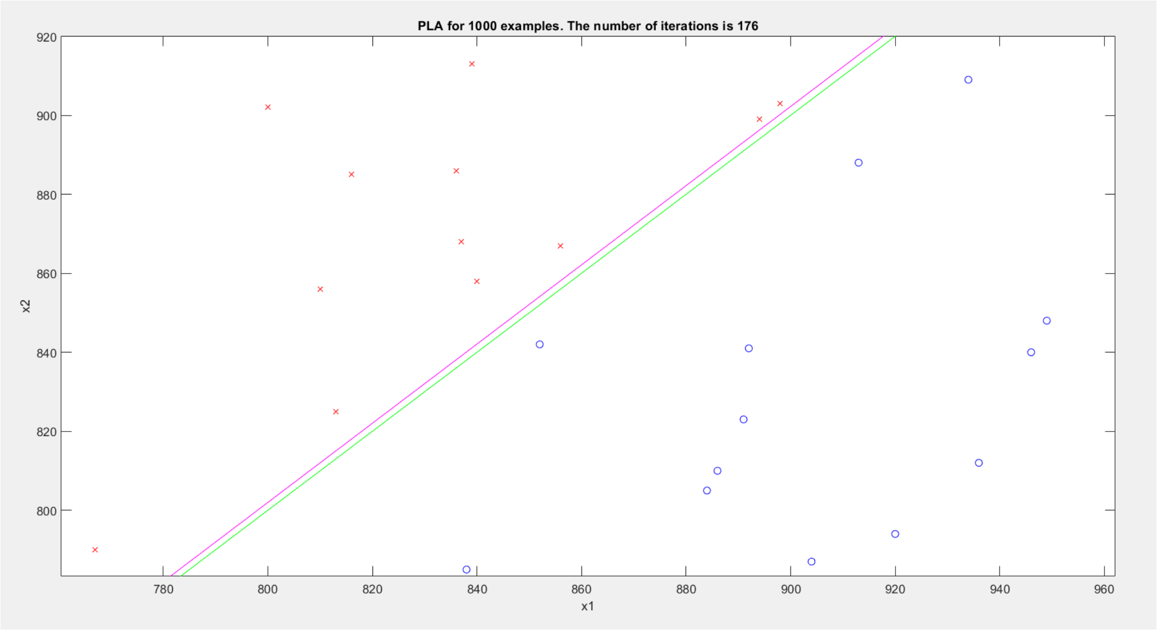


Figure 6: A closer look at Figure 5

Figure 6 demonstrates that data is indeed linear separable and there is no noise.

B) Probability Related Problems

D.4 Independence

If X and Y are discrete r.v, independence means that

$$\mathbb{P}(X=x; Y=y) = \mathbb{P}(X=x)\mathbb{P}(Y=y)$$

If they are continuous, $f(X=x; Y=y) = f(X=x)\, f(Y=y)$

Show: for independent random variables X any Y,

$$\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$$ for discrete and continuous cases

a) Discrete case:

Expectation of X: $\mathbb{E}(X) = \sum_{x} x \cdot \mathbb{P}(X=x)$

Expectation of y: $\mathbb{E}(Y) = \sum_{y} y \cdot \mathbb{P}(Y=y)$

$$\mathbb{E}[XY] = \sum_{x,y} xy\,\mathbb{P}(X=x; Y=y) = \sum_{y}\sum_{x} xy\,\mathbb{P}(X=x; Y=y) =$$

$$\sum_{y}\sum_{x} xy\,\mathbb{P}(X=x)\,\mathbb{P}(Y=y) = \sum_{y} y\,\mathbb{P}(Y=y)\Big(\sum_{x} x\,\mathbb{P}(X=x)\Big) =$$

$$\sum_{y} y\,\mathbb{P}(Y=y)\,\mathbb{E}[X] = \mathbb{E}[X]\sum_{y} y\,\mathbb{P}(Y=y) = \mathbb{E}[X]\,\mathbb{E}[Y]$$

Continuous case:

Expectation of X: $\mathbb{E}[X] = \int_x x f(X=x) dx$

Expectation of Y: $\mathbb{E}[Y] = \int_y y f(Y=y) dy$

$$\mathbb{E}[XY] = \int_x \int_y xy \, f(X=x; Y=y) \, dy \, dx = \int_x \int_y xy \, f(X=x) f(Y=y) dy dx$$

$$= \int_x x f(X=x) \left(\int_y y f(Y=y) dy \right) dx = \int_x x f(X=x) \, \mathbb{E}[Y] \, dx =$$

$$\mathbb{E}[Y] \int_x x f(X=x) dx = \mathbb{E}[Y] \mathbb{E}[X] = \mathbb{E}[X] \mathbb{E}[Y]$$

0.5 I.I.D. assumption in spam filters

I.I.D assumption can be violated if

1. There are some rare words that are not identically distributed in spam messages. (identity is violated)

2. The change of the order of words is not taken into account by filtering (distribution changes)

3. Some words have dependent relationships between each other. For example, adjacent words (independence is violated)

4. Variations of syntax is not considered by filtering. (distribution changes)

⇒ If spammers use rare words, change word order, use different syntactic structures or make syntactic errors, and write dependent words in their messages, they can deceive spam filters.

0.6 Spam filtering equation

Pr (S|W) : the probability that a message is a spam given the word 'replica' appears in it

1. Pr (S): the overall probability that any given message is spam

2. Pr (W|S) : the probability the word 'replica' appears in spam messages

3. Pr (H): the overall probability that any given message is not spam (is ham)

4. Pr (W|H): the probability that the word 'replica' appears in ham messages

Express $\Pr(S|W)$ in terms of 1-4:

As events S and H are a partition of a sample space ($S \cap H = \phi$, $S \cup H = \Omega$), Bayes' theorem can be used.

$$\Pr(S|W) = \frac{\Pr(W|S) \cdot \Pr(S)}{\Pr(W|S) \cdot \Pr(S) + \Pr(W|H) \cdot \Pr(H)}$$

(Conditional probability:

$$\Pr(S|W) = \frac{\Pr(S \cap W)}{\Pr(W)} = \frac{\Pr(W|S) \Pr(S)}{\Pr(W|S) \Pr(S) + \Pr(W|H) \Pr(H)}$$

- Total probability rule is used in the denominator)