University of Stavanger

Stavanger, December 17, 2024

Solutions to theoretical exercise 3

ELE520 Machine learning

Problem 1

We shall design a pattern recognition system to classify 2-dimensional feature vectors $\{\boldsymbol{x}\}$ to class ω_1 ('×'), or class ω_2 ('○'). As a basis for this work, we have performed measurements so that 7 feature vectors are available as training vectors:

$$\mathcal{X}_1 = \left\{ \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 2 \\ 0 \end{pmatrix} \right\} \tag{1}$$

for class ω_1, and

$$\mathcal{X}_2 = \left\{ \begin{pmatrix} 3 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 \\ 3 \end{pmatrix} \right\} \tag{2}$$

for class ω_2.

a) Figure 1 plots the training vectors in the x_1x_2–plane. Figure 2 plots the training vectors in the augmented feature space.

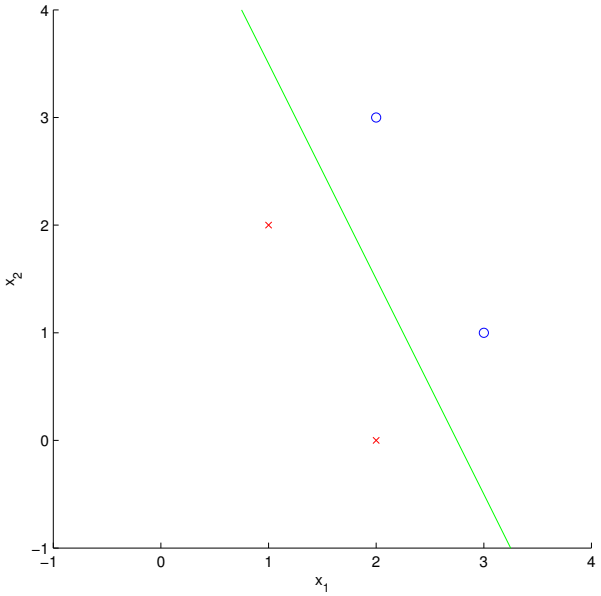

Figure 1: Training vectors and decision boundary (MSE) for the two-class problem.

Trygve Eftestøl, Professor
Faculty of Science and Technology
Department of Electrical and Computer Engineering
Kjølv Egelands hus
University of Stavanger
N-4036 Stavanger.
Telephone: +47 51 83 10 00.
Telefax: +47 51 83 17 50.
E-mail: trygve.eftestol@uis.no
www.ux.his.no/~trygve-e

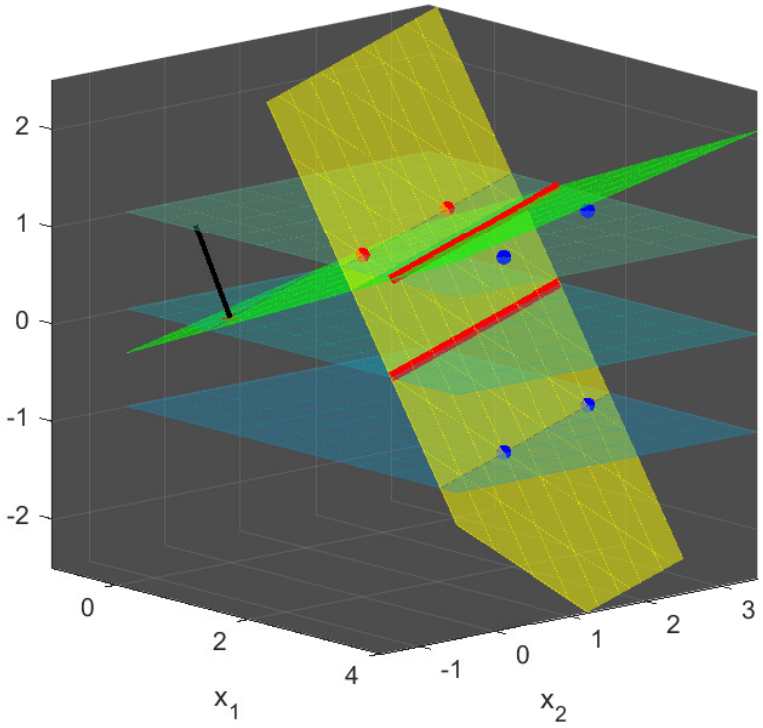

Figure 2: Training vectors and decision boundary (MSE) for the two-class problem in augmented feature space.

b) Determine $\boldsymbol{\theta}$ according to the MSE-method:

$$\boldsymbol{\theta} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\mathbf{b} \tag{3}$$

where $\boldsymbol{y} = (1\ 1\ -1\ -1)^T$ and $\boldsymbol{X} = (\boldsymbol{x}_1 ... \boldsymbol{x}_4)^T$ are the training vectors augmented to to $\boldsymbol{x} = (x_1\ x_2\ 1)^T$.

$$\boldsymbol{X} = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 0 & 1 \\ 3 & 1 & 1 \\ 2 & 3 & 1 \end{pmatrix} \tag{4}$$

Substituted into equation 3 we get

$$\begin{aligned} \boldsymbol{\theta} &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \\ &= \begin{pmatrix} 18 & 11 & 8 \\ 11 & 14 & 6 \\ 8 & 6 & 4 \end{pmatrix}^{-1} \begin{pmatrix} 1 & 2 & 1 \\ 2 & 0 & 1 \\ 3 & 1 & 1 \\ 2 & 3 & 1 \end{pmatrix}^T \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix} \\ &= \begin{pmatrix} -4/3 & -2/3 & 11/3 \end{pmatrix}^T \end{aligned} \tag{5}$$

We find the expression for the decision boundary by solving $g(\boldsymbol{x}) = 0$

$$\begin{aligned} g(\boldsymbol{x}) &= 0 \\ \Leftrightarrow \boldsymbol{\theta}^T\boldsymbol{x} &= 0 \\ \Leftrightarrow (-4/3\ -2/3\ 11/3)(x_1\ x_2\ 1)^T &= 0 \\ \Leftrightarrow -4/3x_1 - 2/3x_2 + 11/3 &= 0 \\ \Leftrightarrow x_2 &= (11/3 - 4/3x_1)/(2/3) \\ \Leftrightarrow x_2 &= (11 - 4x_1)/2 \\ \Leftrightarrow x_2 &= 11/2 - 2x_1 \end{aligned} \tag{6}$$

The decision boundary is shown in figure 1 (solid line).

c) We use the LMS-method and compute $\boldsymbol{\theta}^{(i)}$ for $i = 2, 3...$ with $\boldsymbol{\theta}^{(1)} = (1\ 1\ 1)^T$, threshold value $\theta = 1$ and learning rate $\mu = 0.5$.

The training set $\boldsymbol{x}_k$ is cycled $k = 1, 2, ..., 4, 1, 2, ..., 4, ...$ For each iteration i is picked $\boldsymbol{x}^{(i)}$ corresponding to the next $\boldsymbol{x}_k$ that does not satisfy the criterion $\|\mu(i)(y_k - \boldsymbol{\theta}^{(i)T}\boldsymbol{x}_k)\| < \theta$ (Correspondingly, we set $y^{(i)} = y_k$).
First iteration:

i	k	$\boldsymbol{\theta}^{(i)}$	$\mu(i)$	$\|\mu(i)(y^{(i)} - \boldsymbol{\theta}^T\boldsymbol{x}^{(i)})\boldsymbol{x}^{(i)}\|$	$\mu(i)(y^{(i)} - \boldsymbol{\theta}^T\boldsymbol{x}^{(i)})\boldsymbol{x}^{(i)}$
1	1	$(1\ 1\ 1)^t$	0.5	3.6742	$(-1.5\ -3\ -1.5)^t$
2	2	$(-0.5\ -2\ -0.5)^t$	0.25	1.3975	$(1.25\ 0\ 0.63)^t$
3	4	$(0.75\ -2\ 0.13)^t$	0.16667	2.1047	$(1.13\ 1.69\ 0.56)^t$
4	3	$(1.88\ -0.31\ 0.69)^t$	0.125	2.902	$(-2.63\ -0.88\ -0.88)^t$
5	4	$(-0.75\ -1.19\ -0.19)^t$	0.1	1.5902	$(0.85\ 1.28\ 0.43)^t$

Table 1: Results from the LMS-algorithm-iterations with $\mu = 0.5, \theta = 1$

$$
\begin{aligned}
i &= 1 \\
\mu(1) &= \mu/1 = 0.5 \\
\theta &< \|\mu(i)(y_1 - \boldsymbol{\theta}^T\boldsymbol{x}_1)\boldsymbol{x}_1\| \\
&= \|0.5(1 - (1\ 1\ 1)(1\ 2\ 1)^T)(1\ 2\ 1)^T\| = 3.6742 \\
\Rightarrow \boldsymbol{x}^{(1)} &= \boldsymbol{x}_1, \quad y^{(1)} = y_1 \\
\Rightarrow \boldsymbol{\theta}^{(2)} &= \boldsymbol{\theta}^{(1)} + \mu(1)(y^{(1)} - \boldsymbol{\theta}^{(1)T}\boldsymbol{x}^{(1)})\boldsymbol{x}^{(1)} \\
&= (1\ 1\ 1)^T + 0.5(1 - (1\ 1\ 1)(1\ 2\ 1)^T)(1\ 2\ 1)^T \\
&= (-0.5\ -2\ -0.5)^T
\end{aligned}
\tag{7}
$$

We do the following iterations $i = 2, 3, 4, ...$ until convergence as shown in table 1. Note that only five iterations are required, but the converged decision boundary is placed far away from the data and is definitively not a success.

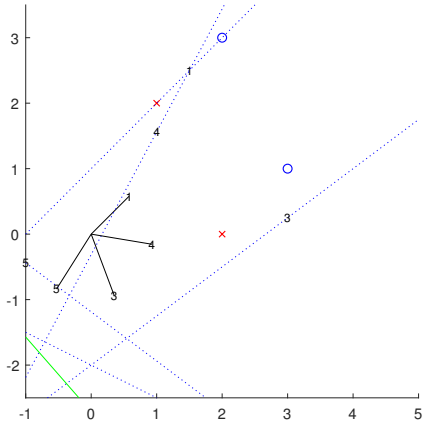

Figure 3: Plot of the training vectors and decision boundaries during the iterations of the LMS-algorithm with $\mu = 0.5, \theta = 1$.

We change the threshold value to $\theta = 0.5$ and learning rate to $\mu = 1$ and make a new attempt.

i	k	$\boldsymbol{\theta}^{(i)}$	$\mu(i)$	$\|\mu(i)(y^{(i)} - \boldsymbol{\theta}^T\boldsymbol{x}^{(i)})\boldsymbol{x}^{(i)}\|$	$\mu(i)(y^{(i)} - \boldsymbol{\theta}^T\boldsymbol{x}^{(i)})\boldsymbol{x}^{(i)}$
1	1	$(1\ 1\ 1)^t$	1	7.3485	$(-3\ -6\ -3)^t$
2	2	$(-2\ -5\ -2)^t$	0.5	7.8262	$(7\ 0\ 3.5)^t$
3	3	$(5\ -5\ 1.5)^t$	0.33333	13.8193	$(-12.5\ -4.17\ -4.17)^t$
4	4	$(-7.5\ -9.17\ -2.67)^t$	0.25	41.3141	$(22.08\ 33.13\ 11.04)^t$
5	1	$(14.58\ 23.96\ 8.38)^t$	0.2	34.2316	$(-13.98\ -27.95\ -13.98)^t$
6	2	$(0.61\ -3.99\ -5.6)^t$	0.16667	2.0062	$(1.79\ 0\ 0.9)^t$
7	4	$(2.4\ -3.99\ -4.7)^t$	0.14286	5.8114	$(3.11\ 4.66\ 1.55)^t$
8	1	$(5.51\ 0.67\ -3.15)^t$	0.125	0.82524	$(-0.34\ -0.67\ -0.34)^t$
9	2	$(5.17\ -0.01\ -3.49)^t$	0.11111	1.4554	$(-1.3\ 0\ -0.65)^t$
10	3	$(3.87\ -0.01\ -4.14)^t$	0.1	2.8085	$(-2.54\ -0.85\ -0.85)^t$
11	4	$(1.33\ -0.85\ -4.98)^t$	0.090909	1.3206	$(0.71\ 1.06\ 0.35)^t$
12	1	$(2.04\ 0.21\ -4.63)^t$	0.083333	0.64976	$(0.27\ 0.53\ 0.27)^t$
13	3	$(2.3\ 0.74\ -4.37)^t$	0.076923	1.0905	$(-0.99\ -0.33\ -0.33)^t$
14	1	$(1.32\ 0.41\ -4.7)^t$	0.071429	0.62362	$(0.26\ 0.51\ 0.26)^t$
15	4	$(1.57\ 0.92\ -4.44)^t$	0.066667	0.61106	$(-0.33\ -0.49\ -0.16)^t$
16	1	$(1.24\ 0.43\ -4.6)^t$	0.0625	0.53683	$(0.22\ 0.44\ 0.22)^t$

Table 2: Results from the LMS-algorithm-iterations with $\theta = 0.5$ and $\mu = 1$.

First iteration:

$$\begin{aligned} i &= 1 \\ \mu(1) &= \mu/1 = 1 \\ \theta &< \|\mu(i)(y_1 - \boldsymbol{\theta}^T\boldsymbol{x}_1)\boldsymbol{x}_1\| \\ &= \|1(1 - (1\ 1\ 1)(1\ 2\ 1)^T)(1\ 2\ 1)^T\| = 7.3485 \\ \Rightarrow \boldsymbol{x}^{(1)} &= \boldsymbol{x}_1,\ \ y^{(1)} = y_1 \\ \Rightarrow \boldsymbol{\theta}^{(2)} &= \boldsymbol{\theta}^{(1)} + \mu(1)(y^{(1)} - \boldsymbol{\theta}^{(1)T}\boldsymbol{x}^{(1)})\boldsymbol{x}^{(1)} \\ &= (1\ 1\ 1)^T + 1(1 - (1\ 1\ 1)(1\ 2\ 1)^T)(1\ 2\ 1)^T \\ &= (-2\ -5\ -2)^T \end{aligned} \tag{8}$$

This time we have to do 16 iterations before convergence as shown in table 2. The decision boundary for $\boldsymbol{\theta}(k)$, is computed according to the procedure shown in 6 and shown in figure 4.

The following table 3 shows the results for the iterations with the initial learning rate set to $\mu = 0.5$ and $\theta = 0.5$. Note that only six iterations are required, but the converged decision boundary is definitively not a success but a blit closer to the data than the first attempt.

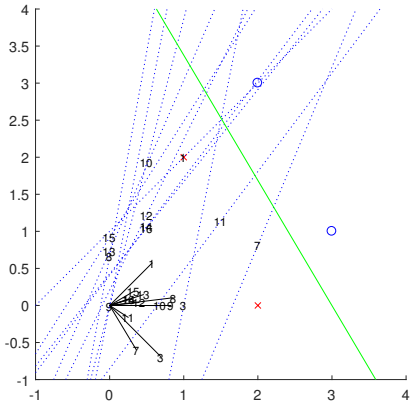

Figure 4: Plot of the training vectors and decision boundaries during the iterations of the LMS-algorithm.

i	k	$\boldsymbol{\theta}^{(i)}$	$\mu(i)$	$\|\mu(i)(y^{(i)} - \boldsymbol{\theta}^T \boldsymbol{x}^{(i)})\boldsymbol{x}^{(i)}\|$	$\mu(i)(y^{(i)} - \boldsymbol{\theta}^T \boldsymbol{x}^{(i)})\boldsymbol{x}^{(i)}$
1	1	$(1\ 1\ 1)^t$	0.5	3.6742	$(-1.5\ \ -3\ \ -1.5)^t$
2	2	$(-0.5\ \ -2\ \ -0.5)^t$	0.25	1.3975	$(1.25\ 0\ 0.63)^t$
3	3	$(0.75\ \ -2\ 0.13)^t$	0.16667	0.76006	$(-0.69\ \ -0.23\ \ -0.23)^t$
4	4	$(0.06\ \ -2.23\ \ -0.1)^t$	0.125	2.6503	$(1.42\ 2.13\ 0.71)^t$
5	2	$(1.48\ \ -0.1\ 0.6)^t$	0.1	0.57299	$(-0.51\ 0\ \ -0.26)^t$
6	3	$(0.97\ \ -0.1\ 0.35)^t$	0.083333	1.1453	$(-1.04\ \ -0.35\ \ -0.35)^t$

Table 3: Results from the LMS-algorithm-iterations with $\mu = 0.5$ and $\theta = 0.5$

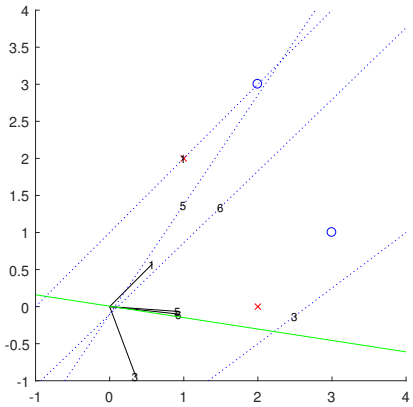

Figure 5: Plot of the training vectors and decision boundaries during the iterations of the LMS-algorithm with $\mu = 0.5$ and $\theta = 0.5$.

Problem 2

a) Given the equation

$$\boldsymbol{X\theta} = \boldsymbol{y} \tag{1}$$

where $\boldsymbol{X}$ has N (number of samples) rows and $(l+1)$ (feature vector dimension) columns ($\boldsymbol{X}$ and thus the dimension $N \times l+1$).

As generally $N \gg l$, $\boldsymbol{X}$ will be rectangular; it ha no inverse to give a solution on the form $\boldsymbol{X}^{-1}\boldsymbol{y}$. That $n \gg d$ also means that we have more side conditions than free variables, which is generally not possible to satisfy. (If it had been opposite, $d > n$, we would have had many solutions, now we have none.)

b) We will minimise

$$\begin{aligned} J_s(\boldsymbol{\theta}) &= \|\boldsymbol{X\theta} - \boldsymbol{y}\|^2 \\ &= (\boldsymbol{X\theta} - \boldsymbol{y})^T(\boldsymbol{X\theta} - \boldsymbol{y}) \\ &= \boldsymbol{\theta}^T\boldsymbol{X}^T\boldsymbol{X\theta} - \boldsymbol{y}^T\boldsymbol{X\theta} - \boldsymbol{\theta}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{y} \\ &= \boldsymbol{\theta}^T\boldsymbol{X}^T\boldsymbol{X\theta} - 2\boldsymbol{\theta}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{y} \end{aligned} \tag{2}$$

Differentiate with respect to $\boldsymbol{\theta}$ (se textbook, A.2.4 s 606-607) and set equal to zero, and solve for $\boldsymbol{\theta}$:

$$\begin{aligned} \boldsymbol{\Delta} J_s &= 2\boldsymbol{X}^T\boldsymbol{X\theta} - 2\boldsymbol{X}^T\boldsymbol{y} = 0 \\ \Rightarrow \boldsymbol{X}^T\boldsymbol{X\theta} &= \boldsymbol{X}^T\boldsymbol{y} \\ \Rightarrow \boldsymbol{\theta} &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \quad \text{q.e.d} \end{aligned} \tag{3}$$

c) This solution is substituted into the exppression for $J_s(\boldsymbol{\theta})$.

$$\begin{aligned} J_{s_{min}} &= \boldsymbol{\theta}^T\boldsymbol{X}^T\boldsymbol{X\theta} - 2\boldsymbol{\theta}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{y} \\ &= ((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y})^T\boldsymbol{X}^T\boldsymbol{X}((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}) - 2((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y})^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{y} \\ &= \boldsymbol{y}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} - 2\boldsymbol{y}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{y} \\ &= \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{I}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} - 2\boldsymbol{y}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{y} \\ &= -\boldsymbol{y}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{I}\boldsymbol{y} \\ &= \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T)\boldsymbol{y} \end{aligned} \tag{4}$$

Note that if $\boldsymbol{X}$ had been quadratic and invertible, $\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = \boldsymbol{I}$, $J_s(\boldsymbol{\theta}) \equiv 0$, og $\boldsymbol{\theta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{X}^{-1}\boldsymbol{y}$ would have been an exact solution.

d) A problem is linearly separable if it can be solved exactly with a linear discriminant, meaning that the data are separated by a hyyperplene with no error.

This is equivalent with $\boldsymbol{X\theta} = \boldsymbol{y}$ having an exact solution for a $\boldsymbol{y}$ with small elements (meaning that $\boldsymbol{X}$ has a special form so that a problem not expected to have a solution actually has one!)

Then $J_{s_{min}} = 0$, will be satisfied which it will be if

$$\begin{aligned} \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T &= \boldsymbol{0} \\ &\Downarrow \\ \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T &= \boldsymbol{I} \end{aligned} \tag{5}$$

(or more generally if $(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T)\boldsymbol{y} = \boldsymbol{0}$).

Problem 3

Consider the data sets with samples $\boldsymbol{y}_i$.

$$\mathcal{X}_1 = \left\{ \begin{pmatrix} -3 \\ -2 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix} \right\} \tag{1}$$

for class ω_1, and

$$\mathcal{X}_2 = \left\{ \begin{pmatrix} -4 \\ -1 \end{pmatrix} \right\} \tag{2}$$

for class ω_2. We want to find the discriminant function $g(\boldsymbol{x}) = \boldsymbol{\theta}^T\boldsymbol{x}$ that solves the inequalities $\boldsymbol{\theta}^T\boldsymbol{x}_i > (<)0$ for class $\omega 1(\omega_2)$.

a) The labels are $y_1 = 1, y_2 = 1, y_3 = -1$. This corresponds to changing the sign for the data in $\mathcal{X}_2$ so that

$$\mathcal{X}_2 = \left\{ \begin{pmatrix} 4 \\ 1 \end{pmatrix} \right\} \tag{3}$$

This is the way this sample is handled in the iterations. It is also placed with canged sign in the plot so it is easier to see how it affects the updates.

b) The samples and the lines defined by $\boldsymbol{\theta}^T\boldsymbol{x}_i = 0$ for all the samples are shown in figure 6.

c) The positive sides of the lines are indicated by (+) and the solution region which is the intersection of the positive sides is shown in black in figure 6.

d) The algorithm is applied letting the initial values be $\boldsymbol{\theta} = 0$,$\mu(1) = 1$, criterion $\theta = 0$. Let $\mu(i) = 1$.

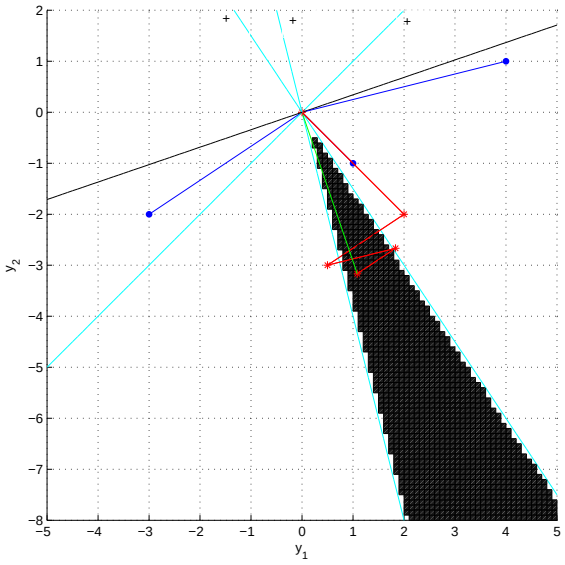

Figure 6: Plot of training vectors and decision boundary (LS) for the two classes.vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3$ (blue) with the corresponding lines defined by $\boldsymbol{\theta}^T \boldsymbol{x}_i = 0$ (cyan).

For the first iteration ($i = 1$)), $\boldsymbol{\theta}^T \boldsymbol{x}_i = 0, k = 1, \ldots, 3$, so that $\mathcal{X}_1 = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3\}$ and

$$\begin{aligned} \boldsymbol{\theta}^{(2)} &= \boldsymbol{\theta}^{(1)} + \mu(1) \sum_{\boldsymbol{x} \in \mathcal{X}_1} y_n \boldsymbol{x} \\ &= (0\ 0)^T + (2\ -2)^T = (2\ -2)^T \end{aligned} \tag{4}$$

The iterations are completed as shown in table 4.

e) As above, but let let $\eta(k) = 1/k$. The iterations are completed as shown in table 5 and also illustrated in figure 6 with updates of $\mathbf{a}(k)$ in red, $\mathbf{a}(5)$ in green with the corresponding decision boundary in black.

i	$\mathcal{X}_i$	$\boldsymbol{\theta}^{(i)}$	$\mu(i)$	$\mu(i)\sum_{\boldsymbol{x}\in\mathcal{X}_i}\boldsymbol{x}$
1	$\{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3\}$	$(0\ 0)^T$	1	$(2\ -2)^T$
2	$\{\boldsymbol{x}_1\}$	$(2\ -2)^T$	1	$(3\ -2)^T$
3	$\{\boldsymbol{x}_3\}$	$(-1\ -4)^T$	1	$(4\ 1)^T$
4	$\{\boldsymbol{x}_1\}$	$(3\ -3)^T$	1	$(-3\ -2)^T$
5	$\{\boldsymbol{x}_3\}$	$(0\ -5)^T$	1	$(-4\ 1)^T$
6	$\{\boldsymbol{x}_1\}$	$(4\ -4)^T$	1	$(-3\ -2)^T$
7	$\{\boldsymbol{x}_3\}$	$(1\ -6)^T$	1	$(-4\ 1)^T$
8	$\{\boldsymbol{x}_1\}$	$(5\ -5)^T$	1	$(-3\ -2)^T$
9	$\{\ \}$	$(2\ -7)^T$	NA	NA

Table 4: Results from the iterations of the batch perceptron algorithm

i	$\mathcal{X}_i$	$\boldsymbol{\theta}^{(i)}$	$\mu(i)$	$\mu(i)\sum_{\boldsymbol{x}\in\mathcal{X}_i}\boldsymbol{x}$
1	$\{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3\}$	$(0\ 0)^T$	1	$(2\ -2)^T$
2	$\{\boldsymbol{x}_1\}$	$(2\ -2)^T$	1/2	$(3\ -2)^T$
3	$\{\boldsymbol{x}_3\}$	$(0.5\ -3)^T$	1/3	$(4\ 1)^T$
4	$\{\boldsymbol{x}_1\}$	$(1.8\ -2.7)^T$	1/4	$(-3\ -2)^T$
5	$\{\ \}$	$(1.1\ -3.2)^T$	NA	NA

Table 5: Results from the iterations of the batch perceptron algorithm with decreasing learning rate.