Machine Learning WS 2018

Solution to Assignment 6

Submitted By : Vindhya Singh **Enrollment Number :** 03693296
Colaborated By : Wasiq Rumaney **Enrollment Number :** 03694978

Solution 1 :

minimise $f_o(\theta)=\theta_1-\sqrt{3}\,\theta_2$

subject to $f_1(\theta)=\theta_1^2+\theta_2^2-4\leq 0$

Using the recipe described in slide 17,

Step 1 : Calculating the Lagrangian Function :

$$L(\theta,\alpha)=\theta_1-\sqrt{3}\,\theta_2+\alpha(\theta_1^2+\theta_2^2-4) \tag{1}$$

Taking the derivative of (1) and setting it to zero, we get,

$$\nabla_{theta} L(\theta,\alpha)=\begin{pmatrix}1\\-\sqrt{3}\end{pmatrix}+2\alpha\theta=0 \tag{2}$$

Thus, solving (2), we get,

$$\theta_1=\frac{-1}{2\alpha} \quad \text{and} \quad \theta_1=\frac{\sqrt{3}}{2\alpha} \tag{A}$$

Step 2 : Calculating the Dual Function :

Putting the value of θ_1 and θ_2 in (1) and getting the dual function $g(\alpha)$

$$g(\alpha)=L(\theta,\alpha)=\frac{-1}{2\alpha}-\sqrt{3}\left(\frac{\sqrt{3}}{2\alpha}\right)+\alpha\left(\frac{1}{4\alpha^2}+\frac{3}{4\alpha^2}-\frac{16\alpha^2}{4\alpha^2}\right)$$

$$= \frac{-1}{2\alpha}-\frac{3}{2\alpha}+\frac{4}{4\alpha}-\frac{16\alpha^2}{4\alpha}$$

$$= \frac{-2-6+4-16\alpha^2}{4\alpha} = \frac{-1}{\alpha}-4\alpha$$

Thus, the dual function is given by

$$g(\alpha)=\frac{-1}{\alpha}-4\alpha \tag{3}$$

Step 3 : Setting the derivative of dual to zero subject to the condition that $g(\alpha)\geq 0$

$\frac{d\,g(\alpha)}{d\,\alpha}=0$ => $\frac{1}{\alpha^2}-4=0$ Thus, subject to the above condition of Step 3, $\alpha=\frac{1}{2}$ (4)

As Slater's Conditions hold, the minimal value of $f_o(\theta)=g(\alpha)=g(\frac{1}{2})=\frac{-1}{\frac{1}{2}}-4(\frac{1}{2})=-4$

[by substituting the value of (4) in (3)]

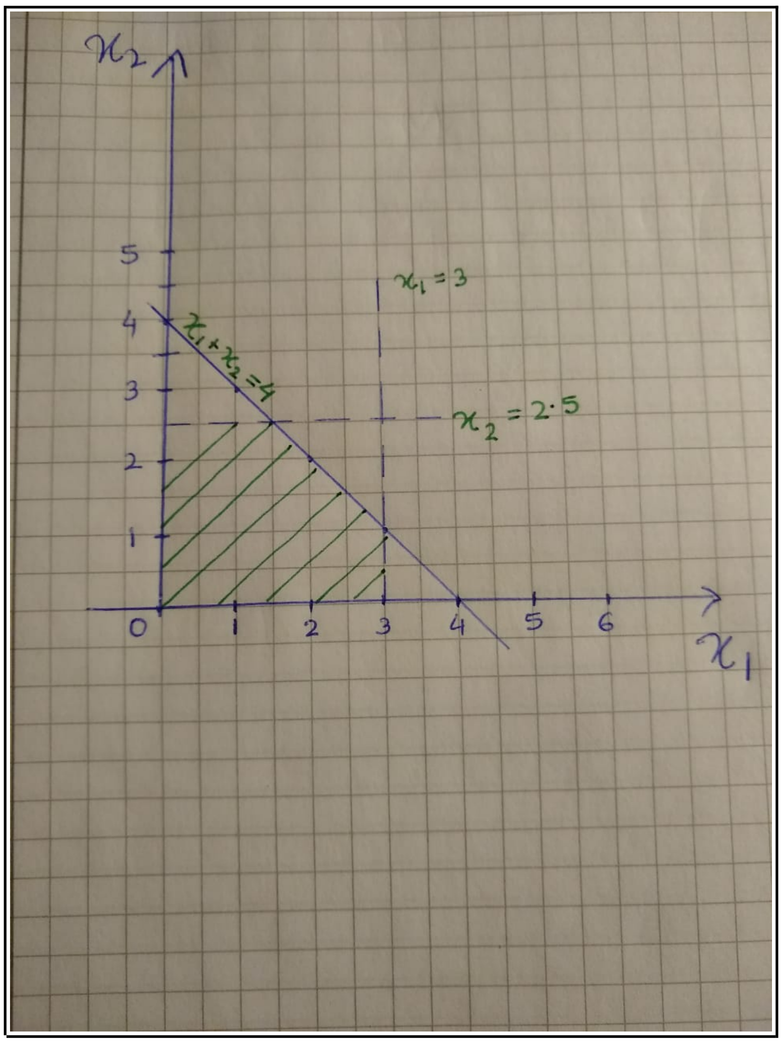
Thus, putting value of (4) in (A), we get,

$\theta_1=-1\,while\;\theta_2=\sqrt{3}$

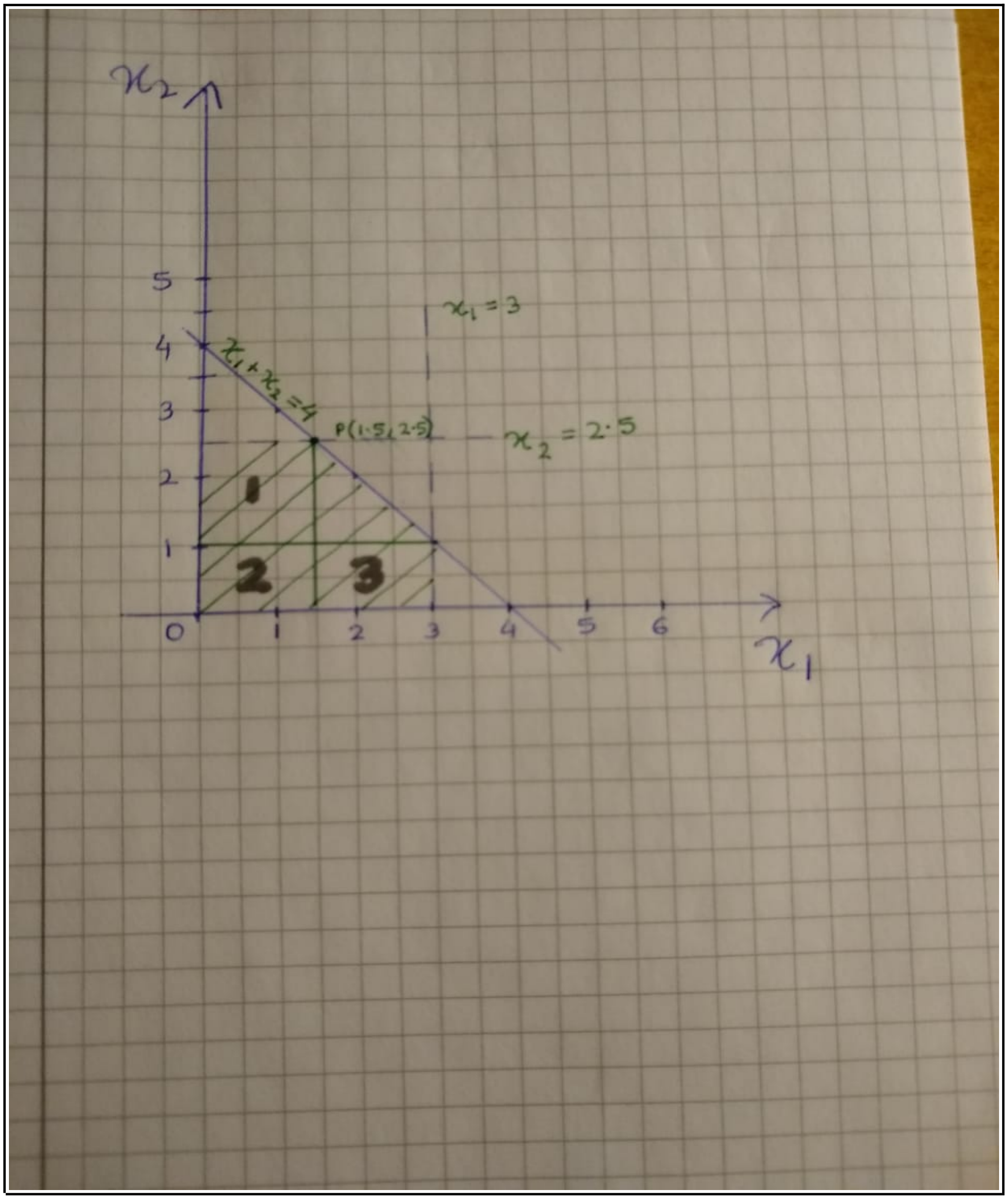
Solution 2 :

a)

The set **X** is visualized as follows :



b)



The above set **X,** as visualised in part (a) can be divided into 3 boxes upon taking into consideration the constraints on x_1 and x_2. In the figure above these regions where box constraints apply are depicted as **Box 1, Box 2 and Box 3.** In the above box regions the box constraints hold . From the lecture slides, we know that,

$\pi_x(p) = min(max(p, l), u)$ where l = lower and u = upper and p = point in set (1)

Thus the vector p, if it falls under the box constraints region, the distance can be evaluated as :

Box 1	Box 2	Box 3
$l_{x1}=0\,;u_{x1}=1.5$	$l_{x1}=0\,;u_{x1}=1.5$	$l_{x1}=1.5\,;u_{x1}=3$
$l_{x2}=1\,;u_{x2}=2.5$	$l_{x2}=0\,;u_{x2}=1$	$l_{x2}=0\,;u_{x2}=1$

where l = lower and u = upper

For regions other than these 3, where the vector is closer to the line,

$\pi_{line}(p)=a+\frac{(p-a)^T(b-a)}{\|b-a\|_2^2}(b-a)$ a and b are defined as (3,1) and (1.5,2.5) respectively. (A)

c)

Given : $minimize\,(x_1-2)^2+(2x_2-7)^2\,such\,that\,,x\,\epsilon\,\mathrm{X}$

Taking the derivative of the above given function wrt x_1 and x_2, we get,

$$\frac{\partial f_o}{\partial x_1}=2x_1-4\,;\frac{\partial f_o}{\partial x_2}=4x_2-28 \qquad (1)$$

Given point $x_o=[2.5,1]^T$, t = learning rate/ step size.

$$x_i^{(s)}=x_i^{(s-1)}-t\frac{\partial f_o}{\partial x_i} \quad \text{where} \quad x_i^s=value\,of\,x_i\,at\,step\,s \qquad (2)$$

From (1) and (2), we get,

$x_1^1=2.5-0.05*1=2.45$ and $x_2^1=1-0.05*4*-5=2$ => point x_1 becomes (2.45,2)

But this point is not in the defined valid region as mentioned in part (b) of the problem. Therefore, using (A) from part b of this question, and substituting the values of a(3,1), b(1.5,2.5) and p(-0.55,1), we get,

x^1 = [1.36,2.63]

For step 2,

$$\frac{\partial f}{\partial x_1}=2x_1-4=0.9 \quad \text{and} \quad \frac{\partial f}{\partial x_2}=4x_2-28=-6.96$$

From (2),

$$x_1^{(2)}=1.36-0.05*0.9=1.32 \quad \text{and} \quad x_2^{(2)}=2.63+0.05*6.96=2.98$$

Thus, the point becomes (1.32,2.98). This value is nearer to the edge of box 1. Therefore, using (A) from part (b) of the question and substituting the values of a(1.5,2.5) , b(0,2.5) and p(-0.18,0.48), we get,

$x^{(2)} = (1.23,2.5)$

Solution 3 :

Similarities in SVM and Perceptron Algorithms : Both separate two classes from each other. Both use hyperplanes. *[Bishop, Section 4.1.7 and 7.1]*

Differences between SVM and Perceptron Algorithms : The decision boundary chosen in an SVM is chosen with the aim of maximizing the margin. A perceptron is guaranteed to find a solution in a finite number of steps, dependent on the intial values for **w** and *b* and the order of the data. Also, SVM uses the concept of *margins* unlike the Perceptron Algorithm. *[Bishop, Section 7.1.5]*

Solution 4 :

From Lecture 06, slide 13, we know that for strong duality to hold,

$$d^* = p^*$$

i.e. the solution to the Lagrange dual problem is a solution of the original (primal) constrained optimization problem.

Also, for an optimization problem such as

Minimize $f_o(x)$

Subject to $f_i(x) \le 0\,; i=1,2,\ldots\ldots,m$

$Ax=b$ where $f_o, f_{1,\ldots}, f_m$ are convex functions.

Slater's condition for convex programming states that strong duality holds if there exists an x such that all constraints are satisfied and the nonlinear constraints are satisfied with strict inequalities.

Since, in an SVM, the objective function is convex (from the Lecture slides), the constraints are affine therefore the duality gap is zero in an SVM as the Slater's conditions hold.

Solution 5 :

a)

$$g(\alpha)=\frac{1}{2}\alpha^T Q\,\alpha+\alpha^T N$$

$$g(\alpha)=\sum_i \alpha_i-\frac{1}{2}\sum_j\sum_i \alpha_i\alpha_j x_i^T x_j y_i y_j=\sum_i \alpha_i-\frac{1}{2}\sum_i\sum_j \alpha_i y_i x_i^T y_j x_j \alpha_j$$

$\frac{-1}{2}\alpha^T(yy^T.xx^T)\alpha$ => $Q=-yy^T\odot xx^T$ [According to the Hadamard Product]

b)

We know that, Q = -M. Thus, if we prove that M is a Positive semi definitive (PSD) matrix then Q is a Negative Semi definitive matrix. (A)

Thus, for M to be PSD, the following condition should hold,

$z^T Mz\geq 0\,\forall z$ i.e.

$z^T(yy^T\odot xx^T)z\geq 0$

So,

$$\sum_i\sum_j z_i z_j y_i y_j x_j x^T=\sum_i\sum_j (z_i y_i x_i)^T(z_j x_j y_j)=(z\odot y)^T xx^T(z\odot y)=\langle(z\odot y)^T x,(z\odot y)^T\rangle\geq 0$$

[positive definite property as mentioned here.]

Thus from the above, M is a PSD => Q is a Negative Semi definite matrix. [From A]

c)

From the result in the b part, we can infer that the maximisation problem will be concave. As in every concave function, every local maxima is a global maxima, thus this property will also hold for the stated problem.