CS 446 / ECE 449 Homework 4

Naman Shukla

TOTAL POINTS

14 / 14

QUESTION 1

SVM Basics 10 pts

1.1 a) 2 / 2

✓ - **0 pts** Correct

- **0.5 pts** Incorrect w
- **0.5 pts** Incorrect b
- **2 pts** Incorrect
- **0 pts** Please select pages for questions
- **0.5 pts** incorrect, without reasoning

1.2 b) 2 / 2

✓ - **0 pts** Correct

- **2 pts** Incorrect
- **0.5 pts** 1 is missing
- **0.5 pts** 2 is missing
- **0.5 pts** 3 is missing
- **0.5 pts** 5 is missing

1.3 c) 4 / 4

✓ - **0 pts** Correct

- **0.5 pts** G is incorrect
- **0.5 pts** z is incorrect
- **0.5 pts** h is incorrect
- **0.5 pts** P is incorrect
- **0.5 pts** q is incorrect
- **4 pts** Incorrect

1.4 d) 2 / 2

✓ - **0 pts** Correct

- **0.5 pts** Typo
- **0.5 pts** Stated margin for C=\infty is minimized / always 0
- **1 pts** Inverted answers for \infty and 0
- **1 pts** Minor mistake
- **2 pts** Incorrect/No answer

QUESTION 2

Kernels 4 pts

2.1 a) 2 / 2

✓ - **0 pts** Correct

- **0.5 pts** Typo
- **1 pts** Minor mistake
- **2 pts** Incorrect

2.2 b) 2 / 2

✓ - **0 pts** Correct

- **0.5 pts** Typo
- **1 pts** Partial credit
- **0 pts** Incorrect

University of Illinois Spring 2018

CS 446: Machine Learning
Homework

Due on Tuesday, Feb 13, 2018, 11:59 a.m. Central Time

1. **[10 points]** SVM Basics
Consider the following dataset $\mathcal{D}$ in the two-dimensional space; $\mathbf{x}^{(i)} \in \mathbb{R}^2$ and $y^{(i)} \in \{1, -1\}$

i	$\mathbf{x}_1^{(i)}$	$\mathbf{x}_2^{(i)}$	$y^{(i)}$
1	-1	3	1
2	-2.5	-3	-1
3	2	-3	-1
4	4.7	5	1
5	4	3	1
6	-4.3	-4	-1

Recall a hard SVM is as follows:

$$\min_{w,b} \frac{1}{2} \|\mathbf{w}\|^2 \quad \text{s.t. } y^{(i)}(\mathbf{w}^\intercal \mathbf{x}^{(i)} + b \geq 1) \ \ , \forall (x^{(i)}, y^{(i)}) \in \mathcal{D} \tag{1}$$

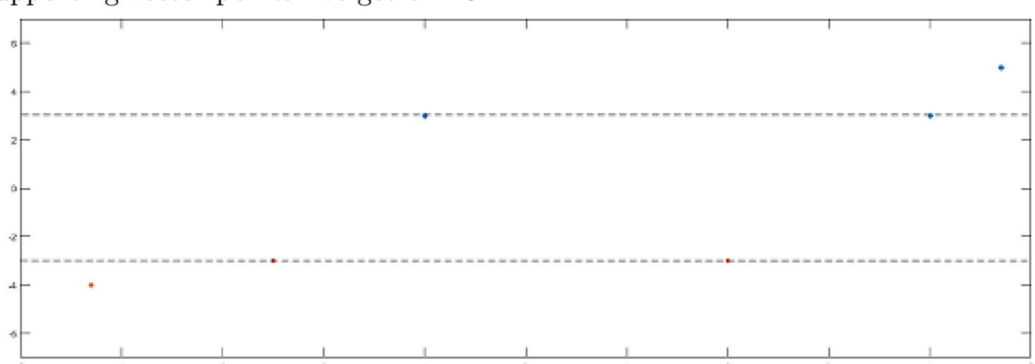
(a) What is the optimal $\mathbf{w}$ and b? Show all your work and reasoning. (Hint: Draw it out.)

Your answer: For the given dataset, the best separation is the two dotted lines in the figure.
Since $\mathbf{w}$ is always points perpendicular to the median of the decision boundaries, $\mathbf{w}$ have no component along x1 direction.
Also we know the length of the boundaries are given by $\frac{2}{\|\mathbf{w}\|}$. So we have,

$$\frac{2}{\|\mathbf{w}\|} = 6$$

$$\|\mathbf{w}\| = \frac{1}{3}$$

Hence, we got $\mathbf{w} = [0 \ \ \frac{1}{3}]$ and now substituting value of $\mathbf{w}$ in $y^{(i)}(\mathbf{w}^\intercal \mathbf{x}^{(i)} + b) = 1$ with supporting vector points. we get b = 0.



(b) Which of the examples are support vectors?

Your answer: The supporting vectors are : (-1,3), (-2.5,-3), (2,-3) and (4,3)

1.1 a) **2 / 2**

✓ - **0 pts** Correct

- **0.5 pts** Incorrect w

- **0.5 pts** Incorrect b

- **2 pts** Incorrect

- **0 pts** Please select pages for questions

- **0.5 pts** incorrect, without reasoning

University of Illinois Spring 2018

CS 446: Machine Learning
Homework

Due on Tuesday, Feb 13, 2018, 11:59 a.m. Central Time

1. **[10 points]** SVM Basics
Consider the following dataset $\mathcal{D}$ in the two-dimensional space; $\mathbf{x}^{(i)} \in \mathbb{R}^2$ and $y^{(i)} \in \{1, -1\}$

i	$\mathbf{x}_1^{(i)}$	$\mathbf{x}_2^{(i)}$	$y^{(i)}$
1	-1	3	1
2	-2.5	-3	-1
3	2	-3	-1
4	4.7	5	1
5	4	3	1
6	-4.3	-4	-1

Recall a hard SVM is as follows:

$$\min_{w,b} \frac{1}{2} \|\mathbf{w}\|^2 \quad \text{s.t. } y^{(i)}(\mathbf{w}^\intercal \mathbf{x}^{(i)} + b \geq 1) \ \ , \forall (x^{(i)}, y^{(i)}) \in \mathcal{D} \tag{1}$$

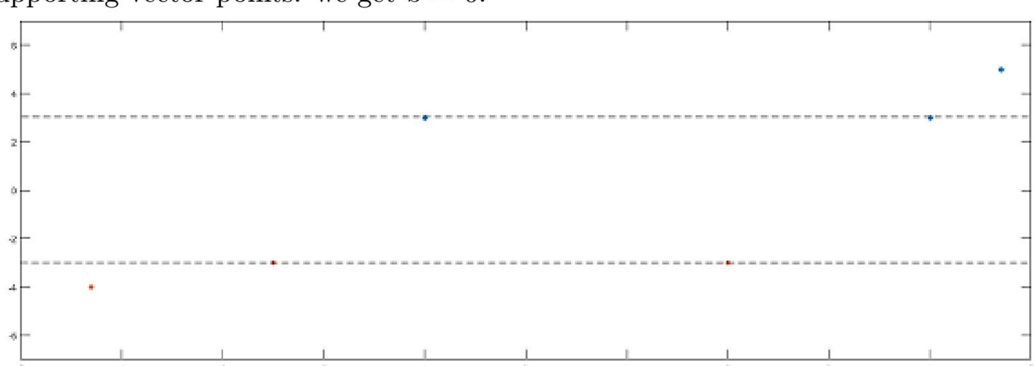
(a) What is the optimal $\mathbf{w}$ and b? Show all your work and reasoning. (Hint: Draw it out.)

Your answer: For the given dataset, the best separation is the two dotted lines in the figure.
Since $\mathbf{w}$ is always points perpendicular to the median of the decision boundaries, $\mathbf{w}$ have no component along x1 direction.
Also we know the length of the boundaries are given by $\frac{2}{\|\mathbf{w}\|}$. So we have,

$$\frac{2}{\|\mathbf{w}\|} = 6$$

$$\|\mathbf{w}\| = \frac{1}{3}$$

Hence, we got $\mathbf{w} = [0 \ \ \frac{1}{3}]$ and now substituting value of $\mathbf{w}$ in $y^{(i)}(\mathbf{w}^\intercal \mathbf{x}^{(i)} + b) = 1$ with supporting vector points. we get b = 0.



(b) Which of the examples are support vectors?

Your answer: The supporting vectors are : (-1,3), (-2.5,-3), (2,-3) and (4,3)

1.2 b) **2 / 2**

- ✓ **- 0 pts** Correct
- **- 2 pts** Incorrect
- **- 0.5 pts** 1 is missing
- **- 0.5 pts** 2 is missing
- **- 0.5 pts** 3 is missing
- **- 0.5 pts** 5 is missing

(c) A standard quadratic program is as follows,

$$\begin{aligned} \underset{\mathbf{z}}{\text{minimize}} \quad & \frac{1}{2}\mathbf{z}^\intercal P\mathbf{z} + \mathbf{q}^\intercal\mathbf{z} \\ \text{subject to} \quad & G\mathbf{z} \le \mathbf{h} \end{aligned}$$

Rewrite Equation (1) into the above form. (*i.e.* define $\mathbf{z}, P, \mathbf{q}, G, \mathbf{h}$ using $\mathbf{w}, b$ and values in $\mathcal{D}$). Write the constraints in the **same order** as provided in $\mathcal{D}$ and typeset it using `bmatrix`.

Your answer: Given:

$$\mathbf{x} = \begin{bmatrix} x_{(1,1)} & x_{(1,2)} & x_{(1,3)} \cdots & x_{(1,k)} \\ x_{(2,1)} & x_{(2,2)} & x_{(2,3)} \cdots & x_{(2,k)} \\ \vdots & & & \\ x_{(|\mathcal{D}|,1)} & x_{(|\mathcal{D}|,2)} & x_{(|\mathcal{D}|,3)} \cdots & x_{(|\mathcal{D}|,k)} \end{bmatrix} \mathbf{w} = [w_1 \ \ w_2 \ \ w_3 \ldots w_k]$$

including bias term as well, we get:

$$\mathbf{x}' = \begin{bmatrix} x_{(1,1)} & x_{(1,2)} & x_{(1,3)} \cdots & x_{(1,k)} & 1 \\ x_{(2,1)} & x_{(2,2)} & x_{(2,3)} \cdots & x_{(2,k)} & 1 \\ \vdots & & & & \\ x_{(|\mathcal{D}|,1)} & x_{(|\mathcal{D}|,2)} & x_{(|\mathcal{D}|,3)} \cdots & x_{(|\mathcal{D}|,k)} & 1 \end{bmatrix} \mathbf{w}' = [w_1 \ \ w_2 \ \ w_3 \ldots w_k \ \ b]$$

Comparing above equation with equation(1),

$$P_{(k+1,k+1)} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \ldots & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \ldots & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \ldots & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \ldots & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \ldots & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 \ldots & 0 \end{bmatrix}$$

where k is the number of dimensions of $x_{(i)}$

$$\mathbf{q}_{(k+1,1)} := Zeros$$
$$\mathbf{z} := \mathbf{w}'$$

Now rewriting the condition equation from equation (1),

$$\text{diag}(y_1, ..., y_{|\mathcal{D}|}) \cdot \mathbf{x}' \cdot \mathbf{w}' \geq \mathbb{1}$$

Where,

$$\text{diag}(y_1, ..., y_{|\mathcal{D}|}) := \begin{bmatrix} y_1 & & \\ & \ddots & \\ & & y_{|\mathcal{D}|} \end{bmatrix} \text{ and}$$

$$\mathbb{1} := \text{matrix with all ones with dimensions}(|\mathcal{D}| \times 1)$$

Now comparing above equation with $G\mathbf{z} \leq \mathbf{h}$ we get,

$$G := - \quad \text{diag}(y_1, ..., y_{|\mathcal{D}|}) \cdot \mathbf{x}'$$

and

$$\mathbf{h} = -\mathbb{1}$$

1.3 c) **4 / 4**

- ✓ **- 0 pts** Correct
- **- 0.5 pts** G is incorrect
- **- 0.5 pts** z is incorrect
- **- 0.5 pts** h is incorrect
- **- 0.5 pts** P is incorrect
- **- 0.5 pts** q is incorrect
- **- 4 pts** Incorrect

(d) Recall that for a soft-SVM we solve the following optimization problem.

$$\min_{w,b} \frac{1}{2} \|\mathbf{w}\|^2 + C \cdot \sum_{i=1}^{|D|} \xi^{(i)} \quad \text{s.t. } y^{(i)}(\mathbf{w}^\intercal \mathbf{x}^{(i)} + b \geq 1 - \xi^{(i)}), \xi^{(i)} \geq 0 \ \ , \forall (x^{(i)}, y^{(i)}) \in \mathcal{D} \tag{2}$$

Describe what happens to the margin when $C = \infty$ and $C = 0$.

Your answer: The C parameter tells the SVM optimization how much you want to avoid misclassifying each training example. For large values of C ($=\infty$), the optimization will choose the smallest margin which classify most of the points correctly. This is because the weight for the misclassification term is infinite. On the other hand, if C is 0, the optimization will give large margin even if it missclassify more number of points.

2. **[4 points]** Kernels

(a) If $K_1(\mathbf{x}, \mathbf{z})$ and $K_2(\mathbf{x}, \mathbf{z})$ are both valid kernel functions, and α and β are positive, prove that

$$\alpha K_1(\mathbf{x}, \mathbf{z}) + \beta K_2(\mathbf{x}, \mathbf{z})$$

is also a valid kernel function.

Your answer: A function is a valid kernel function if it is a real-valued positive definite function (A real-valued function K on X^2 is called a positive definite function if it is symmetric and follow the below equation)

$$\forall n \in \mathbb{N}^*, \ \forall \{x_i\}_{i=1}^n \in^n, \ \forall \{a_i\}_{i=1}^n \in^n, \quad \sum_{i=1}^n \sum_{j=1}^n a_i a_j K(x_i, x_j) \geq 0.$$

Now, proof:
By construction, the Gram matrix is given by

$$K = \alpha K_1(\mathbf{x}, \mathbf{z}) + \beta K_2(\mathbf{x}, \mathbf{z})$$

which implies that

$$\forall a \in^n, \quad a^T K a = \alpha(a^T K_1(\mathbf{x}, \mathbf{z}) a) + \beta(a^T K_2(\mathbf{x}, \mathbf{z}) a) \geq 0$$

due to the positivity of the α and β, hence the validity of the kernel K.
Another way:

$$k_1(x, y) = \langle \phi^{(1)}(x), \phi^{(1)}(y) \rangle$$
$$k_2(x, y) = \langle \phi^{(2)}(x), \phi^{(2)}(y) \rangle$$

Let us construct $\phi(x) = \langle \sqrt{a}\phi^{(1)}(x) \ , \ \sqrt{b}\phi^{(2)}(x) \rangle$
Clearly then,

$$k(x, y) = a\langle \phi^{(1)}(x), \phi^{(1)}(y) \rangle + b\langle \phi^{(2)}(x), \phi^{(2)}(y) \rangle = ak_1(x, y) + bk_2(x, y)$$

1.4 d) **2 / 2**

- ✓ **- 0 pts** Correct
- **- 0.5 pts** Typo
- **- 0.5 pts** Stated margin for C=\infty is minimized / always 0
- **- 1 pts** Inverted answers for \infty and 0
- **- 1 pts** Minor mistake
- **- 2 pts** Incorrect/No answer

(d) Recall that for a soft-SVM we solve the following optimization problem.

$$\min_{w,b} \frac{1}{2}\|\mathbf{w}\|^2 + C \cdot \sum_{i=1}^{|D|} \xi^{(i)} \quad \text{s.t. } y^{(i)}(\mathbf{w}^\intercal \mathbf{x}^{(i)} + b \geq 1 - \xi^{(i)}), \xi^{(i)} \geq 0 \ , \forall (x^{(i)}, y^{(i)}) \in \mathcal{D} \tag{2}$$

Describe what happens to the margin when $C = \infty$ and $C = 0$.

Your answer: The C parameter tells the SVM optimization how much you want to avoid misclassifying each training example. For large values of C ($=\infty$), the optimization will choose the smallest margin which classify most of the points correctly. This is because the weight for the misclassification term is infinite. On the other hand, if C is 0, the optimization will give large margin even if it missclassify more number of points.

2. **[4 points]** Kernels

(a) If $K_1(\mathbf{x}, \mathbf{z})$ and $K_2(\mathbf{x}, \mathbf{z})$ are both valid kernel functions, and α and β are positive, prove that

$$\alpha K_1(\mathbf{x}, \mathbf{z}) + \beta K_2(\mathbf{x}, \mathbf{z})$$

is also a valid kernel function.

Your answer: A function is a valid kernel function if it is a real-valued positive definite function (A real-valued function K on X^2 is called a positive definite function if it is symmetric and follow the below equation)

$$\forall n \in \mathbb{N}^*, \ \forall \{x_i\}_{i=1}^n \in^n, \ \forall \{a_i\}_{i=1}^n \in^n, \quad \sum_{i=1}^n \sum_{j=1}^n a_i a_j K(x_i, x_j) \geq 0.$$

Now, proof:
By construction, the Gram matrix is given by

$$K = \alpha K_1(\mathbf{x}, \mathbf{z}) + \beta K_2(\mathbf{x}, \mathbf{z})$$

which implies that

$$\forall a \in^n, \quad a^T K a = \alpha(a^T K_1(\mathbf{x}, \mathbf{z}) a) + \beta(a^T K_2(\mathbf{x}, \mathbf{z}) a) \geq 0$$

due to the positivity of the α and β, hence the validity of the kernel K.
Another way:

$$k_1(x, y) = \langle \phi^{(1)}(x), \phi^{(1)}(y) \rangle$$
$$k_2(x, y) = \langle \phi^{(2)}(x), \phi^{(2)}(y) \rangle$$

Let us construct $\phi(x) = \langle \sqrt{a}\phi^{(1)}(x) \ , \ \sqrt{b}\phi^{(2)}(x) \rangle$
Clearly then,

$$k(x, y) = a\langle \phi^{(1)}(x), \phi^{(1)}(y) \rangle + b\langle \phi^{(2)}(x), \phi^{(2)}(y) \rangle = ak_1(x, y) + bk_2(x, y)$$

2.1 a) **2 / 2**

- ✓ **- 0 pts** Correct
- **- 0.5 pts** Typo
- **- 1 pts** Minor mistake
- **- 2 pts** Incorrect

(b) Show that $K(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^\intercal \mathbf{z})^2$ is a valid kernel, for $\mathbf{x}, \mathbf{z} \in \mathbb{R}^2$.
($i.e.$ write out the $\Phi(\cdot)$, such that $K(\mathbf{x}, \mathbf{z}) = \Phi(\mathbf{x})^\intercal \Phi(\mathbf{z})$)

Your answer: For $\mathbf{x} = (x_1, x_2), \mathbf{z} = (z_1, z_2)$:

$$K(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^\intercal \mathbf{z})^2$$

$$= (x_1^2 z_1^2 + 2x_1 z_1 x_2 z_2 + x_2^2 z_2^2)$$

$$= \Phi(\mathbf{x})^\intercal \Phi(\mathbf{z})$$

where, $\Phi(\mathbf{x}) = (x_1^2, \sqrt{2} x_1 x_2, x_2^2)$

2.2 b) **2 / 2**

✓ - **0 pts** Correct

- **0.5 pts** Typo

- **1 pts** Partial credit

- **0 pts** Incorrect