Name: ______________________

University of Illinois Spring 2020

CS 446/ECE 449 Machine Learning
Homework 3: Support Vector Machine (SVM)

Due on Thursday February 20 2020, noon Central Time

1. **[30 points]** Max-Margin Support Vector Machine

We are given a dataset $\mathcal{D} = \left\{ \left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}, 1 \right), \left(\begin{bmatrix} 0 \\ 1 \end{bmatrix}, 1 \right), \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, -1 \right), \left(\begin{bmatrix} -1 \\ -1 \end{bmatrix}, -1 \right) \right\}$ containing four pairs (x, y), where each $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \in \mathbb{R}^2$ denotes a 2-dimensional point and $y \in \{-1, +1\}$.

We want to train the parameters w and the bias b of a max-margin support vector machine (SVM) using (with hyperparameter $C > 0$)

$$\min_{w,b} \frac{C}{2}\|w\|_2^2 \quad \text{s.t.} \quad \forall (x^{(i)}, y^{(i)}) \in \mathcal{D} \quad y^{(i)}(w^\top x^{(i)} + b) \geq 1. \tag{1}$$

(a) (5 points) For the given data $\mathcal{D}$, how many constraints are part of the program in Eq. (1)? Specify all of them **explicitly**.

Your answer:

c1) $w^T \begin{bmatrix} 1 \\ 0 \end{bmatrix} + b \geq 1$

c2) $w^T \begin{bmatrix} 0 \\ 1 \end{bmatrix} + b \geq 1$

c3) $w^T \begin{bmatrix} 0 \\ 0 \end{bmatrix} + b \leq -1$

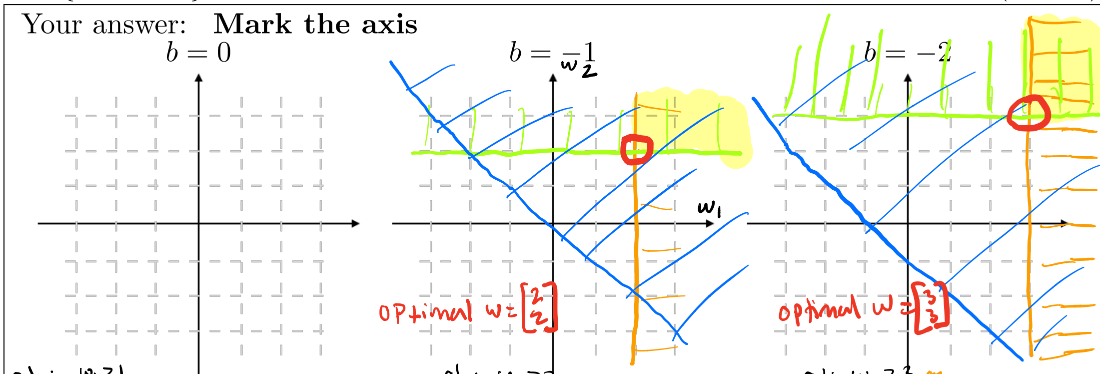
c4) $w^T \begin{bmatrix} -1 \\ -1 \end{bmatrix} + b \leq -1$

4 CONSTRAINTS IN TOTAL.

Name:

(b) (8 points) Highlight the feasible set in w_1-w_2-space for $b = 0$, $b = -1$ and $b = -2$. For each of the three choices for b also highlight the optimal w. Given only the three options $b \in \{0, -1, -2\}$ what is the optimal solution? Does a better solution exist (reason)?

Your answer: **Mark the axis**



c1 : $w_1 \geq 1$
c2: $w_2 \geq 1$
c3: $0 \leq -1$ ✗
c4: $w_1 + w_2 \geq 1$
No FEASIBLE SET FOR $b=0$ b/c pt 3 violates constraints.

c1 : $w_1 \geq 2$
c2: $w_2 \geq 2$
c3: $-1 \leq -1$ ✓
c4: $w_2 \geq -w_1$
= feasible set

c1: $w_1 \geq 3$
c2: $w_2 \geq 3$
c3: $-2 \leq -1$ ✓
c4: $w_2 \geq -w_1 - 1$

Given these 3 options, $b = -1$ gives optimal $w = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$, b/c this gives the smallest $\|w\|_2^2$.
No better solution exists b/c if we try to make $\|w\|_2^2$ smaller, ie make $b=0$, the constraints fail, specifically constraint 3.

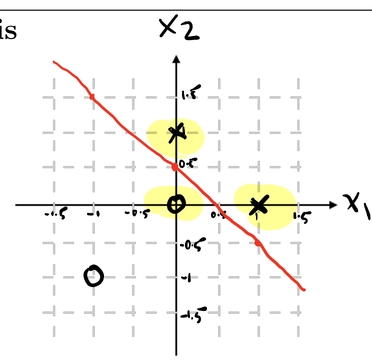
(c) (6 points) Draw the dataset in x_1-x_2-space using crosses for the points belonging to class 1 and circles for the points belonging to class -1. Find by inspection and highlight the support vectors, *i.e.*, those points for which the constraints hold with equality at the optimal solution. Solve the resulting linear system w.r.t. w and b and draw the solution into x_1-x_2-space.

Your answer: **Mark the axis**

$w_1 + b = 1$

$w_2 + b = 1$

$$\boxed{\begin{aligned} b &= -1 \\ w_1 &= 2 \\ w_2 &= 2 \end{aligned}}$$



$\begin{bmatrix} 2 \\ 2 \end{bmatrix} \begin{bmatrix} x_1 & x_2 \end{bmatrix} - 1 = 0$

$2x_1 + 2x_2 - 1 = 0$

$x_1 + x_2 = \frac{1}{2}$

$x_2 = -x_1 + 1/2$

Name: ____________________

(d) (1 point) What conditions do the datapoints have to fulfill such that the program in Eq. (1) has a feasible solution?

Your answer:

all points need to be linearly seperable.

(e) (6 points) In practice, for large datasets, it is hard to find the support vectors by inspection. A gradient based method is applicable. Use **general** notation, introduce slack variables into the program given in Eq. (1) and state the corresponding program (including all constraints). Subsequently, reformulate this program into an unconstrained program. Finally compute the gradient of this unconstrained program w.r.t. w (use $\frac{\partial}{\partial x}\max\{0, x\} = 1$ for $x > 0$, 0 otherwise). Evaluate the gradient at $w_1 = 2$, $w_2 = 2$ and $b = -1$. What can we conclude?

Your answer:

$$\min_{w,\xi\geq 0} \underbrace{\frac{c}{2}\|w\|_2^2}_{\text{MARGIN}} + \underbrace{\sum_{i\in D} \xi^{(i)}}_{\text{Correct classification}} \quad \text{ST} \quad y^{(i)} w^T x^{(i)} + b \geq 1 - \xi^{(i)} \;\; \forall i \in D$$

$$\xi^{(i)} \geq 1 - y^{(i)} w^T x^{(i)} - b$$

$$\xi^{(i)} \geq 0 \;\; \forall i \in D.$$

$$\max(0, 1 - y^{(i)} w^T x^{(i)} - b)$$

$$\min_{w,\xi\geq 0} \frac{c}{2}\|w\|_2^2 + \sum_{i\in D} \max(0, 1 - y^{(i)} w^T x^{(i)}).$$

$$\nabla = cw + \begin{cases} 0 & \text{when } y^{(i)} w^T x^{(i)} + b \geq 1 \text{ (correct classification)} \\ -y^{(i)} w^T x^{(i)} & \text{when } y^{(i)} w^T x + b < 1 \text{ (incorrect classification)} \end{cases}$$

$$\nabla = c\begin{bmatrix} 2 \\ 2 \end{bmatrix} + \underbrace{0}_{\text{b/c data is perfectly linearally seperable}}$$

$$\nabla = c\begin{bmatrix} 2 \\ 2 \end{bmatrix}.$$

reason

b/c the gradient is not ∅ at this point, we did not fully optimize this objective even tho this is proven to be optimal in previous questio

If $c \to \infty$: It behaves like soft svm b/c we care more about maximizing margin than we do about correct classification

If $c \to 0$: behaves like hard svm b/c we care more about correct classification and less about max margin

(f) (4 points) Complete A3_SVM.py and verify your reply for the previous answer. What is the optimal solution (w, b) that your program found and what's the corresponding loss? Explain the solution and what you observe when running the program, as well as how to fix this issue.

Your answer:

$$w = \begin{bmatrix} 0.678 \\ 0.678 \end{bmatrix} \quad b = 0.3329$$

loss = 1.781

Our loss does not converge to ∅. This is b/c $c=1$ and we have soft SVM as explained in e. To fix this we make $c \to 0$. Ie $c = 0.1$ gives expected optimal values and a loss = 0.04 after 1000 steps.