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**.

Solution: 4 constraints

$$\begin{aligned} &w_1 + b \geq 1 \\ &w_2 + b \geq 1 \\ &-b \geq 1 \\ &w_1 + w_2 - b \geq 1 \end{aligned}$$

(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)?

Solution: $b = 0$: no feasible set
$b = -1$: $w_1 \geq 2$, $w_2 \geq 2$; $w^* = [2, 2]^\top$
$b = -2$: $w_1 \geq 3$, $w_2 \geq 3$; $w^* = [3, 3]^\top$
Optimal solution is $w_1 = 2$, $w_2 = 2$, $b = -1$. No better solution exists because the feasible set is empty for $b > -1$ and the cost function increases for $b < -1$.

(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.

Solution: Support vectors: $x^{(1)}$, $x^{(2)}$ and $x^{(3)}$
Linear system:

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

Solution: $b = -1$, $w_1 = 2$, $w_2 = 2$

Name:

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

Solution:
linearly separable

(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?

Solution:

$$\min_{w,b,\xi_i \geq 0} \frac{C}{2}\|w\|_2^2 + \sum_i \xi_i \qquad \text{s.t.} \quad y^{(i)}(w^\top x^{(i)} + b) \geq 1 - \xi_i$$

$$\min_{w,b} \frac{C}{2}\|w\|_2^2 + \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \max\{0, 1 - y^{(i)}(w^\top x^{(i)} + b)\}$$

$$Cw + \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \delta(y^{(i)}(w^\top x^{(i)} + b) < 1)(-y^{(i)}x^{(i)})$$

Gradient:

$$\begin{bmatrix} 2C \\ 2C \end{bmatrix}, \quad (\text{if } \frac{\partial}{\partial x}\max\{0, x\} = 1 \text{ for } x \geq 0 : \begin{bmatrix} 1 \\ 1 \end{bmatrix})$$

Not optimal. This is due to the fact that constraints are now soft and norm plays a role. Note that strictly speaking we'd need to consider the sub-gradients and also the gradient w.r.t. b which we ignore here for simplicity.

(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.

Solution:

$$(w, b) = \left(\begin{bmatrix} 0.6674 \\ 0.6674 \end{bmatrix}, 0.3330\right)$$

loss is around 1.778; wrong result; $C = 1$ is too large and needs to be reduced to get the correct answer