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 $y^{(i)} \in \{1, -1\}$

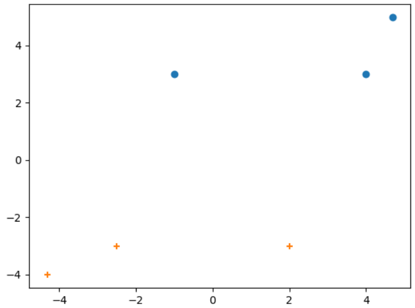
i	$\mathbf{x}_1^{(i)}$	$\mathbf{x}_2^{(i)}$	$y^{(i)}$
1	-1	3	1
2	-2.5	-3	-1
3	2	-3	-1
4	4.7	5	1
5	4	3	1
6	-4.3	-4	-1

   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: Below there is the plot showing the different examples in our dataset.



.

Now it seems evident that the support vectors are (-1, 3), (4, 3), (-2.5, -3) and (2, -3). So the margin is defined by $x_2 = 3$ and $x_2 = -3$ and the width of the margin is 6. Since $\mathbf{w}$ has to be perpendicular to the margin we have that $w_1 = 0$. Now to find w_2 we can use the relation

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

Since $w_1 = 0$ we have that $w_2 = \frac{1}{3}$ and

$$\mathbf{w} = \begin{bmatrix} 0 \\ \frac{1}{3} \end{bmatrix}$$

To find b we can use one of the support vectors, lets take (-1, 3):

$$(1)\left(0 \cdot -1 + \frac{1}{3} \cdot 3\right) + b = 1$$

Then $b = 0$.

(b) Which of the examples are support vectors?

Your answer: The support vectors are instances 1, 2, 3, 5.

(c) A standard quadratic program is as follows,

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

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: Let D be the number of dimensions of $\mathbf{x}$ and $N = |\mathcal{D}|$ the number of elements in our data set.
Lets first multiply the constraint by -1 so that we can match the components with the QP.

$$-y^{(i)}(\mathbf{w}^\intercal \mathbf{x}^{(i)} + b) \leq -1$$

Since this is true for all i we can write it in matrix form as follow

$$-\begin{bmatrix} y^{(1)} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & y^{(N)} \end{bmatrix} \begin{bmatrix} x^{(1)} & 1 \\ \vdots & \vdots \\ x^{(N)} & 1 \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ b \end{bmatrix} = \begin{bmatrix} -1 \\ \vdots \\ -1 \end{bmatrix}$$

where the first matrix is $N \times N$ and it is created by putting the $y^{(i)}$ in the i-th diagonal position; the second matrix is $N \times (D+1)$.
Now we can take $\mathbf{z} = \begin{bmatrix} \mathbf{w} \\ b \end{bmatrix}$.
Then

$$G = -\begin{bmatrix} y^{(1)} & 0 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & y^{(N)} \end{bmatrix} \begin{bmatrix} x^{(1)} & 1 \\ \vdots & \vdots \\ x^{(N)} & 1 \end{bmatrix}$$

$$\mathbf{h} = \begin{bmatrix} -1 \\ \vdots \\ -1 \end{bmatrix}$$

Lets now remember that $||\mathbf{w}||^{\mathbf{2}}$ can be written as $\mathbf{w}^\intercal \mathbf{w}$. We can take $\mathbf{q}^\intercal = \begin{bmatrix} 0 & \cdots & 0 \end{bmatrix}$.
Finally we want $\mathbf{z}^\intercal P \mathbf{z}$ to be $||\mathbf{w}||^2$.

$$\begin{aligned} \mathbf{z}^\intercal P \mathbf{z} &= \begin{bmatrix} \mathbf{w} & b \end{bmatrix} P \begin{bmatrix} \mathbf{w} \\ b \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{w} & b \end{bmatrix} \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & 1 & 0 \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ b \end{bmatrix} \end{aligned}$$

So P is a $(D+1) \times (D+1)$ matrix where the $D \times D$ upper-left matrix is an identity matrix and the last column and row are filled with zeros to get rid of the b.

(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: When $C = \infty$ we are making the second term (the cost of the slack variables) to prevail over $\mathbf{w}$ so the minimization process will need to minimize the $\xi^{(i)}$. This means that it will try to find a separation that perfectly classifies all the data.

When $C = 0$ we are saying that we don't care about the slack variables at all. They can be anything so there may be many mis-classifications. Note that this is equivalent to hard-SVM because minimizing $||\mathbf{w}||$ is the same as maximizing the margin $\left(\frac{2}{||\mathbf{w}||}\right)$.

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:

$$\begin{aligned} K_3(\mathbf{x}, \mathbf{z}) &= \alpha K_1(\mathbf{x}, \mathbf{z}) + \beta K_2(\mathbf{x}, \mathbf{z}) \\ &= \alpha \phi_1(\mathbf{x})^\intercal \phi_1(\mathbf{z}) + \beta \phi_2(\mathbf{x})^\intercal \phi_2(\mathbf{z}) \\ &= \begin{bmatrix} \sqrt{\alpha}\phi_1(\mathbf{z}) & \sqrt{\beta}\phi_2(\mathbf{z}) \end{bmatrix} \begin{bmatrix} \sqrt{\alpha}\phi_1(\mathbf{x})^\intercal \\ \sqrt{\beta}\phi_2(\mathbf{x})^\intercal \end{bmatrix} \end{aligned}$$

Therefore, we can define our new ϕ function in terms of ϕ_1 and ϕ_2.
Suppose $\phi_1(\cdot) \in \mathbb{R}^m$ and $\phi_2(\cdot) \in \mathbb{R}^n$

$$\phi(\cdot) = \begin{bmatrix} \sqrt{\alpha}\phi_1(\cdot) & \sqrt{\beta}\phi_2(\cdot) \end{bmatrix}^\intercal \in \mathbb{R}^{m+n}$$

(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:

$$\begin{aligned} K(\mathbf{x}, \mathbf{z}) &= (\mathbf{x}^\intercal \mathbf{z})^2 \\ &= (x_1 z_1 + x_2 z_2)^2 \\ &= (x_1 z_1)^2 + 2x_1 x_2 z_1 z_2 + (x_2 z_2)^2 \\ &= x_1^2 z_1^2 + 2x_1 x_2 z_1 z_2 + x_2^2 z_2^2 \\ &= \begin{bmatrix} z_1^2 & \sqrt{2} z_1 z_2 & z_2^2 \end{bmatrix} \begin{bmatrix} x_1^2 \\ \sqrt{2} x_1 x_2 \\ x_2^2 \end{bmatrix} \\ &= \phi(\mathbf{x})^\intercal \phi(\mathbf{z}) \end{aligned}$$

$$\therefore \phi(\mathbf{a}) = \begin{bmatrix} a_1^2 & \sqrt{2} a_1 a_2 & a_2^2 \end{bmatrix}^\intercal$$

where

$$\mathbf{a} = \begin{bmatrix} a_1 & a_2 \end{bmatrix}$$