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

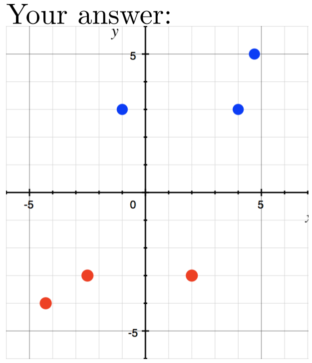
i	$\mathbf{x}_1^{(i)}$	$\mathbf{x}_2^{(i)}$	$y^{(i)}$
1	-1	3	1
2	-2.5	-3	-1
3	2	-3	-1
4	4.7	5	1
5	4	3	1
6	-4.3	-4	-1

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:



$\mathbf{w} = [0; \frac{1}{3}]$, $b = 0$

From the graph its quite clear the $\mathbf{x}_1$ has no effect on the classification. This leaves $\mathbf{w}_2$ and b unknown. We can use the four support vectors to determine the optimal values for these two parameters. We want the decision hyper-plane to be equidistant from each of these vectors which leaves us with $\mathbf{w}_2 = 1$ and $b = 0$. $\mathbf{w}_2$ can be further minimized down to $\mathbf{w}_2 = \frac{1}{3}$ to reduce the total cost while conforming the hard SVM constraint. Its important to note that increasing b off of 0 in either direction requires $\mathbf{w}_2$ to also increase, thereby increasing the total cost. Ex. if $b = \frac{1}{2}$, $\mathbf{w}_2$ would have to be increased to $\frac{1}{2}$, increasing the total cost as mentioned.

(b) Which of the examples are support vectors?

Your answer: The support vector are examples: 1, 2, 3, and 5

(c) A standard quadratic program is as follows,

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

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:

$$\mathbf{z} = \begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \\ b \end{bmatrix} \; P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \; \mathbf{q} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \; G = \begin{bmatrix} 1 & -3 & -1 \\ -2.5 & -3 & 1 \\ 2 & -3 & 1 \\ -4.7 & -5 & -1 \\ -4 & -3 & -1 \\ -4.3 & -4 & 1 \end{bmatrix} \; \mathbf{h} = \begin{bmatrix} -1 \\ -1 \\ -1 \\ -1 \\ -1 \\ -1 \end{bmatrix}$$

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

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

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

Your answer: When $C = 0$ the soft-SVM becomes a hard-SVM and has a margin of 1. If $C = \infty$ the margin is reduced to 0 in order to lower the cost as much as possible.

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:

$$\begin{aligned} \alpha K_1(\mathbf{x}, \mathbf{z}) + \beta K_2(\mathbf{x}, \mathbf{z}) &= \alpha \Phi(\mathbf{x})^T \Phi(\mathbf{z}) + \beta \Phi(\mathbf{x})^T \Phi(\mathbf{z}) \\ &= \begin{bmatrix} \alpha \Phi(\mathbf{x})^T & \beta \Phi(\mathbf{x})^T \end{bmatrix} \begin{bmatrix} \alpha \Phi(\mathbf{z}) \\ \beta \Phi(\mathbf{z}) \end{bmatrix} \\ &= (\alpha K_1 + \beta K_2)(\mathbf{x}, \mathbf{z}) \\ &= K_3(\mathbf{x}, \mathbf{z}) \end{aligned}$$

Which forms a valid kernel function, as the key restriction is that the kernel function must form a proper inner product.

(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:

$$
\begin{aligned}
K(\mathbf{x}, \mathbf{z}) &= (\mathbf{x}^T \mathbf{z})^2 \\
&= (\mathbf{x}_1 \mathbf{z}_1 + \mathbf{x}_2 \mathbf{z}_2)^2 \\
&= (\mathbf{x}_1 \mathbf{z}_1)^2 + 2(\mathbf{x}_1 \mathbf{z}_1)(\mathbf{x}_2 \mathbf{z}_2) + (\mathbf{x}_2 \mathbf{z}_2)^2 \\
&= \mathbf{x}_1^2 \mathbf{z}_1^2 + 2\mathbf{x}_1 \mathbf{z}_1 \mathbf{x}_2 \mathbf{z}_2 + \mathbf{x}_2^2 \mathbf{z}_2^2 \\
&= \begin{bmatrix} \mathbf{x}_1^2 \\ \sqrt{2}\mathbf{x}_1 \mathbf{x}_1 \\ \mathbf{x}_2^2 \end{bmatrix}^T \begin{bmatrix} \mathbf{z}_1^2 \\ \sqrt{2}\mathbf{z}_1 \mathbf{z}_1 \\ \mathbf{z}_2^2 \end{bmatrix}
\end{aligned}
$$

$$
\Phi(\mathbf{x}) = \begin{bmatrix} \mathbf{x}_1^2 \\ \sqrt{2}\mathbf{x}_1 \mathbf{x}_1 \\ \mathbf{x}_2^2 \end{bmatrix}
$$