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

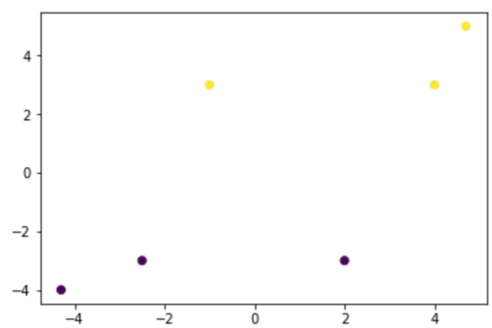
i	$\mathbf{x}_1^{(i)}$	$\mathbf{x}_2^{(i)}$	$y^{(i)}$
1	-1	3	1
2	-2.5	-3	-1
3	2	-3	-1
4	4.7	5	1
5	4	3	1
6	-4.3	-4	-1

Recall a hard SVM is as follows:

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

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

Your answer:
Below is the corrdinates of those those data point. Same color means data points with same label.



So the hyperplane which seperates those data points the best is the x-axis, which gives the highest distance of margin which equals 6. Recall that the distance of margin equals $\frac{2}{||\mathbf{w}||}$.
So we can figure out that $\mathbf{w} = [0 \;\; \frac{1}{3}]$ and $b = 0$.

(b) Which of the examples are support vectors?

Your answer: Examples with indices 1, 2, 3 and 5 are support vectors.

(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:
Define variables as follows:

$\mathbf{z}$ has a dimension of (d+1,1), which d denotes the features' dimension.

$$\mathbf{z} = \begin{bmatrix} b \\ \mathbf{w} \end{bmatrix}$$

$\mathbf{q}$ has a dimension of (d+1,1), which d denotes the features' dimension.

$$\mathbf{q} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

P has a dimension of (d+1, d+1), which d denotes the features' dimension.

$$P = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \dots & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \dots & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \dots & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \dots & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \dots & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 \dots & 1 \end{bmatrix}$$

I use m to demote the number of observations (number of labels), $\mathbf{h}$ has dimension (m, 1).

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

In matrix G, I use m to demote the number of observations (number of labels) and d to demote the number of features.

$$G = -\begin{bmatrix} y_1 & x_{11}y_1 & x_{12}y_1 \dots & x_{1d}y_1 \\ y_2 & x_{21}y_2 & x_{22}y_2 \dots & x_{2d}y_2 \\ y_3 & x_{31}y_3 & x_{32}y_3 \dots & x_{3d}y_3 \\ \vdots & \vdots & \ddots & \vdots \\ y_m & x_{m1}y_m & x_{m2}y_m \dots & x_{md}y_m \end{bmatrix}$$

(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:
I understand that C determines the influence of the misclassification on the objective function. The objective function is the sum of a regularization term and the misclassification rate.

If $C = 0$, then the weight of regularization term will be infinity. Then SVM gives a large margin, even though there are misclassifications there, because of the 'maximum of the margin'.

IF $C = \infty$, then the weight of regularization term will be zero. So SVM gives a much smaller margin will calssify all samples 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:

$$K_1(\mathbf{x}, \mathbf{z}) = \Phi^{(1)}(\mathbf{x})^\intercal \Phi^{(1)}(\mathbf{z})$$
$$K_2(\mathbf{x}, \mathbf{z}) = \Phi^{(2)}(\mathbf{x})^\intercal \Phi^{(2)}(\mathbf{z})$$

Let us construct:

$$\Phi(\mathbf{x}) = [\sqrt{\alpha}\Phi^{(1)} \sqrt{\beta}\Phi^{(2)}(\mathbf{x})]$$

Clearly then:

$$\begin{aligned} K(\mathbf{x}, \mathbf{z}) &= \Phi(\mathbf{x})^\intercal \Phi(\mathbf{z}) \\ &= \alpha[\Phi^{(1)}\mathbf{x}, \;\; \Phi^{(1)}\mathbf{z}] + \beta[\Phi^{(2)}\mathbf{x}, \;\; \Phi^{(2)}\mathbf{z}] \\ &= \alpha K_1(\mathbf{x}, \; \mathbf{z}) + \beta K_2(\mathbf{x}, \mathbf{z}) \end{aligned}$$

So $K(\mathbf{x}, \mathbf{z}) = \alpha K_1(\mathbf{x}, \mathbf{z}) + \beta K_2(\mathbf{x}, \mathbf{z})$ is also a valid kernel function.

(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:

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

So $K(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^\intercal \mathbf{z})^2$ is a valid kernel.