Homework 3

CSCI 567: Machine Learning

Seyed Mohammad Asghari Pari

Kernels

Solution to Question 1.1:

For any N and any pairs $\{\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_N}\}$ such that $\boldsymbol{x_i} \neq \boldsymbol{x_j}$ if $i \neq j$, the matrix

$$K = \begin{bmatrix} k(\boldsymbol{x_1}, \boldsymbol{x_1}) & k(\boldsymbol{x_1}, \boldsymbol{x_2}) & \ldots & k(\boldsymbol{x_1}, \boldsymbol{x_N}) \\ k(\boldsymbol{x_2}, \boldsymbol{x_1}) & k(\boldsymbol{x_2}, \boldsymbol{x_2}) & \ldots & k(\boldsymbol{x_2}, \boldsymbol{x_N}) \\ \vdots & \vdots & \ddots & \vdots \\ k(\boldsymbol{x_N}, \boldsymbol{x_1}) & k(\boldsymbol{x_N}, \boldsymbol{x_2}) & \ldots & k(\boldsymbol{x_N}, \boldsymbol{x_N}) \end{bmatrix} = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{bmatrix} = I, \tag{1}$$

is positive semi-definite (PSD) where $I \in R^{N \times N}$ is the identity matrix. This is because we know that the $N \times N$ identity matrix is PSD (all the eigen-values are 1). This can be also proved as follows. A matrix M is PSD if for any vector $\boldsymbol{z} = [z_1 \quad z_2 \quad \ldots \quad z_N]^\intercal$, we have $\boldsymbol{z}^\intercal M \boldsymbol{z} \geq 0$. If we calculate $\boldsymbol{z}^\intercal I \boldsymbol{z}$, we get

$$z_1^2 + z_2^2 + \ldots + z_N^2 \geq 0 \tag{2}$$

where the inequality is correct because $z_n^2 \geq 0$ for any z_n, $n = 1, \ldots, N$.

Solution to Question 1.2:

If $\lambda = 0$, we have

$$J(\boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{\alpha}^\intercal K^\intercal K \boldsymbol{\alpha} - \boldsymbol{y}^\intercal K \boldsymbol{\alpha} + \frac{1}{2}\boldsymbol{y}^\intercal \boldsymbol{y}. \tag{3}$$

Now, from part 1, we know that the kernel matrix K is the identity matrix I. Then, (3) can be written as

$$J(\boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{\alpha}^\intercal \boldsymbol{\alpha} - \boldsymbol{y}^\intercal \boldsymbol{\alpha} + \frac{1}{2}\boldsymbol{y}^\intercal \boldsymbol{y}. \tag{4}$$

By taking derivatives w.r.t $\boldsymbol{\alpha}$ and set it to zero, we get,

$$\frac{\partial J(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} = 0, \Longrightarrow \boldsymbol{\alpha}^* - \boldsymbol{y} = 0, \Longrightarrow \boldsymbol{\alpha}^* = \boldsymbol{y}. \tag{5}$$

Substituting $\boldsymbol{\alpha}^*$ back into (4), we get

$$J(\boldsymbol{\alpha}^*) = \frac{1}{2}\boldsymbol{y}^\intercal \boldsymbol{y} - \boldsymbol{y}^\intercal \boldsymbol{y} + \frac{1}{2}\boldsymbol{y}^\intercal \boldsymbol{y} = 0. \tag{6}$$

Furthermore, the learned regressor is

$$f(\boldsymbol{x}) = [k(\boldsymbol{x}, \boldsymbol{x_1}) \quad k(\boldsymbol{x}, \boldsymbol{x_2}) \quad \ldots \quad k(\boldsymbol{x}, \boldsymbol{x_N})]\, \boldsymbol{y}. \tag{7}$$

Solution to Question 1.3:

If $\boldsymbol{x} \neq \boldsymbol{x_n}$, $\forall n = 1, \ldots, N$, then by definition of kernel function given in the question we $k(\boldsymbol{x}, \boldsymbol{x_n}) = 0$, $\forall n = 1, \ldots, N$. Then, from (7), we get

$$f(\boldsymbol{x}) = [0 \quad 0 \quad \ldots \quad 0]\, \boldsymbol{y} = \boldsymbol{0}. \tag{8}$$

Support Vector Machines

Solution to Question 2.1:

No. To show this, by contradiction assume there is a linear separator $f(x) = wx + b$. To separate the points correctly, we want

$$f(x_1) = -w + b < 0, \quad f(x_2) = w + b < 0, \quad f(x_3) = 0 + b > 0. \tag{9}$$

No by summing the first and the second inequality, we get $2b < 0$ which means we should have $b < 0$. But, this contradicts the third inequality above, that is, $b > 0$. Hence, there is no linear separator.

Solution to Question 2.2:

Yes, there is a linear separator. Let $f(\boldsymbol{z}) = \boldsymbol{w}^\intercal \boldsymbol{z} + b$ be the linear separator where $\boldsymbol{z} = \phi(x) = \begin{bmatrix} x & x^2 \end{bmatrix}^\intercal$. To separate the points correctly, we want

$$f(x_1) = \boldsymbol{w} \begin{bmatrix} -1 \\ 1 \end{bmatrix} + b < 0, \quad f(x_2) = \boldsymbol{w} \begin{bmatrix} 1 \\ 1 \end{bmatrix} + b < 0, \quad f(x_3) = \boldsymbol{w} \begin{bmatrix} 0 \\ 0 \end{bmatrix} + b > 0 \tag{10}$$

This means that we should have,

$$f(x_1) = -w_1 + w_2 + b < 0, \quad f(x_2) = w_1 + w_2 + b < 0, f(x_3) = b > 0. \tag{11}$$

Now, if choose for example, $w_1 = 0, w_2 = -2, b = 1$, all the above inequalities are satisfied. Hence, indeed there is a linear separator.

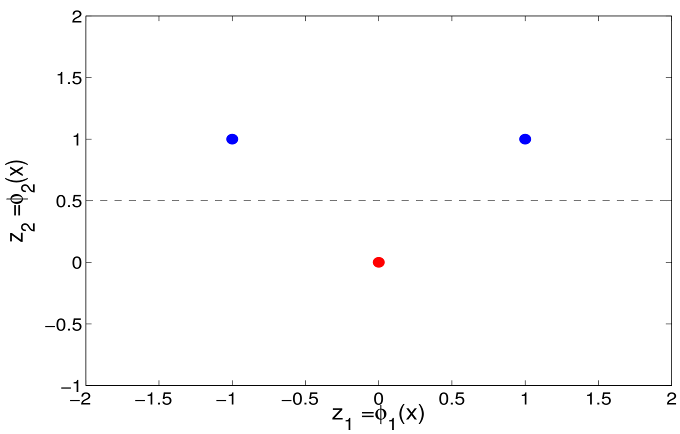

Figure 1: Q 2.2

Solution to Question 2.3:

Since $\phi(x) = \begin{bmatrix} x & x^2 \end{bmatrix}^\intercal$, we have

$$k(x, x') = \phi(x)^\intercal \phi(x') = \begin{bmatrix} x & x^2 \end{bmatrix} \begin{bmatrix} x \\ x^2 \end{bmatrix} = xx' + (xx')^2. \tag{12}$$

Then using (12) and the fact that $(x_1, y_1) = (-1, -1), (x_3, y_3) = (0, 1), (x_2, y_2) = (1, -1)$, we get

$$K = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{13}$$

To show that K is PSD, we show that for any vector $\boldsymbol{z} = [z_1 \quad z_2 \quad z_3]^\intercal$, we have $\boldsymbol{z}^\intercal K \boldsymbol{z} \geq 0$. If we calculate $\boldsymbol{z}^\intercal K \boldsymbol{z}$, we get

$$\boldsymbol{z}^\intercal K \boldsymbol{z} = [z_1 \quad z_2 \quad z_3] \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = [z_1 \quad z_2 \quad z_3] \begin{bmatrix} 2z_1 \\ 2z_2 \\ 0 \end{bmatrix} = 2z_1^2 + 2z_2^2 \geq 0, \tag{14}$$

where the last inequality is correct because $z_1^2, z_2^2 \geq 0$ for any z_1, z_2.

Solution to Question 2.4:

Primal problem:

$$\begin{aligned}
\min_{\boldsymbol{w},b} \quad & \frac{1}{2}||\boldsymbol{w}||_2^2, \\
s.t. \quad & y_n[\boldsymbol{w}^\intercal \phi(x_n) + b] \geq 1, \quad n = 1, 2, 3.
\end{aligned} \tag{15}$$

Dual problem:

$$\begin{aligned}
\max_{\boldsymbol{\alpha}} \quad & \sum_{n=1}^{3} \alpha_n - \frac{1}{2} \sum_{m=1}^{3} \sum_{n=1}^{3} y_m y_n \alpha_m \alpha_n K_{mn}, \\
s.t. \quad & 0 \leq \alpha_n, \quad n = 1, 2, 3, \\
& \sum_{n=1}^{3} \alpha_n y_n = 0.
\end{aligned} \tag{16}$$

Note that we have $(x_1, y_1) = (-1, -1), (x_3, y_3) = (0, 1), (x_2, y_2) = (1, -1)$. Further, in (13), we found,

$$K = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{17}$$

Hence, the dual problem can be written as,

$$\begin{aligned}
\min_{\boldsymbol{\alpha}} \quad & \alpha_1^1 + \alpha_2^1 - (\alpha_1 + \alpha_2 + \alpha_3), \\
s.t. \quad & 0 \leq \alpha_n, \quad n = 1, 2, 3, \\
& -\alpha_1 - \alpha_2 + \alpha_3 = 0.
\end{aligned} \tag{18}$$

Solution to Question 2.5:

To solve the dual problem, note that from $-\alpha_1 - \alpha_2 + \alpha_3 = 0$ in (18), we get $\alpha_3 = \alpha_1 + \alpha_2$. By substituting this in (18), we get

$$\begin{aligned}
\min_{\boldsymbol{\alpha}} \quad & \alpha_1^1 + \alpha_2^1 - 2(\alpha_1 + \alpha_2), \\
s.t. \quad & 0 \leq \alpha_1, \quad 0 \leq \alpha_2.
\end{aligned} \tag{19}$$

Let's call the objective function $g(\alpha_1, \alpha_2)$. Then g is a convex function of (α_1, α_2), because the Hessian H is PSD.

$$\nabla g = \begin{bmatrix} 2\alpha_1 - 2 \\ 2\alpha_2 - 2 \end{bmatrix}, \quad H = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}. \tag{20}$$

By setting the gradient to zero, we get $\alpha_1 = \alpha_2 = 1$ with $g(1, 1) = -2$. To make sure that $\alpha_1 = \alpha_2 = 1$ is the minimizer of g, we need to calculate g at the boundary points. Since we have $g(0, 0) = 0$, $\alpha_1 = \alpha_2 = 1$ is the minimizer. Hence, we have

$$\alpha_1^* = \alpha_2^* = 1, \quad \alpha_3^* = 2. \tag{21}$$

The weight vector $\boldsymbol{w}$ then can be calculated as,

$$\boldsymbol{w} = \sum_{n=1}^{3} \alpha_n y_n \phi(x_n) = -1 \begin{bmatrix} -1 \\ 1 \end{bmatrix} + -1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} + 1 \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ -2 \end{bmatrix}. \tag{22}$$

Hence, the decision boundary is $\hat{y} = \boldsymbol{w}^\intercal \phi(x) + b = \boldsymbol{w}^\intercal \boldsymbol{z} + b = 0$. We want the points to be classified correctly, hence,

$$[0 \quad 2] \begin{bmatrix} -1 \\ 1 \end{bmatrix} + b = -1, [0 \quad 2] \begin{bmatrix} 1 \\ 1 \end{bmatrix} + b = -1, \quad f(x_3) = [0 \quad 2] \begin{bmatrix} 0 \\ 0 \end{bmatrix} + b = 1, \tag{23}$$

where all three inequalities result in $b^* = 1$.

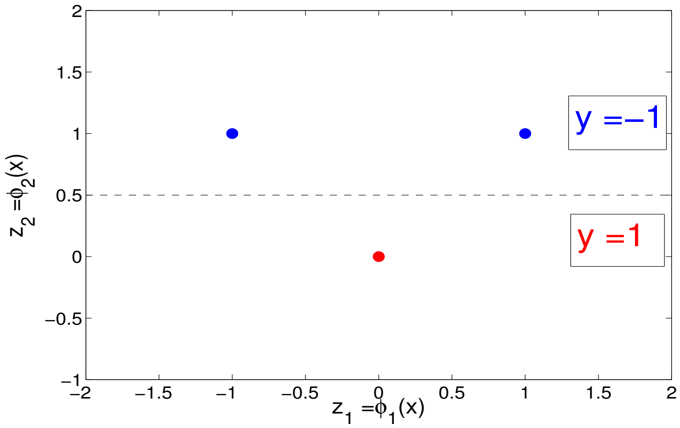

Figure 2: Q 2.6: Decision boundary in 2D

Solution to Question 2.6:

We have $\hat{y} = \boldsymbol{w}^\intercal \phi(x) + b = \boldsymbol{w}^\intercal \boldsymbol{z} + b = -2z_2 + 1$. Hence, the decision boundary is $z_2 = 0.5$. Figure 2 shows the decision boundary in 2D.

To find the decision boundary in 1D, note that $z_2 = \phi_2(x) = x^2$. Hence, from boundary $z_2 = 0.5$, we have $x^2 = 0.5$ which results in $x = -\frac{1}{\sqrt{2}}$ and $x = \frac{1}{\sqrt{2}}$. Since if $z_2 < 0.5$, the label is 1 and otherwise -1, we obtain that the label is 1 if $-\frac{1}{\sqrt{2}} < x < \frac{1}{\sqrt{2}}$, and -1 otherwise. Figure 3 shows the decision boundary in 1D.

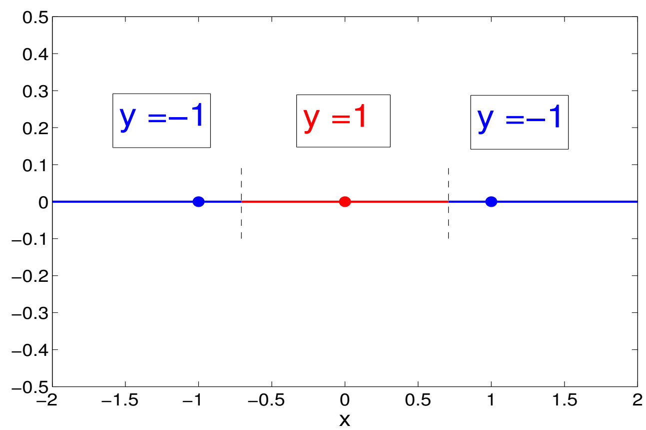

Figure 3: Q 2.6: Decision boundary in 1D

Adaboost for building up a nonlinear classifier

Solution to Question 3.1:

Since $w_1(1) = w_1(2) = w_1(3) = w_1(4) = 0.25$, from Table 1, we get $\sum_{n=1}^{4} w_1(n) 1[y_n = h_{(s,b,d)}(\boldsymbol{x_n})] = 0.5$ for all possible values of (s, b, d). We arbitrarily choose $f_1 = h_{(1,-2,1)}$. Then, $\epsilon_1 = \sum_{n=1}^{4} w_1(n) 1[y_n = h_{(1,-2,1)}(\boldsymbol{x_n})] = 0.5$ and $\beta_1 = \frac{1}{2} \log \frac{1-\epsilon_1}{\epsilon} = 0$.

Table 1: Q 3.1: For original training set

(s,b,d)	$1[y_1 = h_{(s,b,d)}(\boldsymbol{x_1})]$	$1[y_1 = h_{(s,b,d)}(\boldsymbol{x_2})]$	$1[y_1 = h_{(s,b,d)}(\boldsymbol{x_3})]$	$1[y_1 = h_{(s,b,d)}(\boldsymbol{x_4})]$
$(1,-2,1)$	0	1	0	1
$(-1,-2,1)$	1	0	1	0
$(1,-2,2)$	0	1	0	1
$(-1,-2,2)$	1	0	1	0
$(1,-0.5,1)$	0	0	1	1
$(-1,-0.5,1)$	1	1	0	0
$(1,-0.5,2)$	0	1	1	0
$(-1,-0.5,2)$	1	0	0	1
$(1,0.5,1)$	0	0	1	1
$(-1,0.5,1)$	1	1	0	0
$(1,0.5,2)$	0	1	1	0
$(-1,0.5,2)$	1	0	0	1
$(1,2,1)$	1	0	1	0
$(-1,2,1)$	0	1	0	1
$(1,2,2)$	1	0	1	0
$(-1,2,2)$	0	1	0	1

Table 2: Q 3.3: For transformed training set

(s,b,d)	$1[y_1 = h_{(s,b,d)}(\boldsymbol{x_1})]$	$1[y_1 = h_{(s,b,d)}(\boldsymbol{x_2})]$	$1[y_1 = h_{(s,b,d)}(\boldsymbol{x_3})]$	$1[y_1 = h_{(s,b,d)}(\boldsymbol{x_4})]$
$(1,-2,1)$	0	1	0	1
$(-1,-2,1)$	1	0	1	0
$(1,-2,2)$	0	1	0	1
$(-1,-2,2)$	1	0	1	0
$(1,-0.5,1)$	0	0	0	1
$(-1,-0.5,1)$	1	1	1	0
$(1,-0.5,2)$	0	1	1	1
$(-1,-0.5,2)$	1	0	0	0
$(1,0.5,1)$	1	0	1	1
$(-1,0.5,1)$	0	1	0	0
$(1,0.5,2)$	0	0	1	0
$(-1,0.5,2)$	1	1	0	1
$(1,2,1)$	1	0	1	0
$(-1,2,1)$	0	1	0	1
$(1,2,2)$	1	0	1	0
$(-1,2,2)$	0	1	0	1

Solution to Question 3.2:

Since $\beta_1 = 0$, we get $w_2(1) = w_1(1) = 0.25$, $w_2(2) = w_1(2) = 0.25$, $w_2(3) = w_1(3) = 0.25$, $w_2(4) = w_1(4) = 0.25$.

Solution to Question 3.3:

Since $w_1(1) = w_1(2) = w_1(3) = w_1(4) = 0.25$, from Table 2, the minimum for $\sum_{n=1}^{4} w_1(n) 1[y_n = h_{(s,b,d)}(\boldsymbol{x_n})]$ is 0.25. We arbitrarily choose $f_1 = h_{(1,-0.5,1)}$ as one of the minimizers. Then, $\epsilon_1 = \sum_{n=1}^{4} w_1(n) 1[y_n = h_{(1,-0.5,1)}(\boldsymbol{x_n})] = 0.25$ and $\beta_1 = \dfrac{1}{2} \log \dfrac{1-\epsilon_1}{\epsilon} = 0.55$.

Solution to Question 3.4:

Since $\beta_1 = 0.55$ and we chose $f_1 = h_{(1,-0.5,1)}$, we first get $w_2(1) = w_1(1)\exp(-0.55) = 0.14$, $w_2(2) = w_1(2)\exp(-0.55) = 0.14$, $w_2(3) = w_1(3)\exp(-0.55) = 0.14$, and $w_2(4) = w_1(4)\exp(0.55) = 0.43$. Then, after normalization, we get $w_2(1) = w_2(2) = w_2(3) = 0.16$, and $w_2(4) = 0.5$.

Since $w_2(1) = w_2(2) = w_2(3) = 0.16$, and $w_2(4) = 0.5$, from Table 2, the minimum for $\sum_{n=1}^{4} w_2(n)1[y_n = h_{(s,b,d)}(\boldsymbol{x_n})]$ is 0.16. We arbitrarily choose $f_2 = h_{(1,0.5,2)}$ as one of the minimizers. Then, $\epsilon_2 = \sum_{n=1}^{4} w_2(n)1[y_n = h_{(1,0.5,2)}(\boldsymbol{x_n})] = 0.16$ and $\beta_2 = \dfrac{1}{2}\log\dfrac{1-\epsilon_1}{\epsilon} = 0.83$.

Solution to Question 3.5:

Since $\beta_2 = 0.83$ and we chose $f_2 = h_{(1,0.5,2)}$, we first get $w_3(1) = w_2(1)\exp(-0.83) = 0.07$, $w_3(2) = w_2(2)\exp(-0.83) = 0.07$, $w_3(3) = w_2(3)\exp(0.83) = 0.37$, and $w_3(4) = w_2(4)\exp(-0.83) = 0.22$. Then, after normalization, we get $w_3(1) = w_3(2) = 0.10$, $w_3(3) = 0.51$ and $w_3(4) = 0.30$.

Since $w_3(1) = w_3(2) = 0.10$, $w_3(3) = 0.51$, and $w_3(4) = 0.30$, from Table 2, the minimum for $\sum_{n=1}^{4} w_3(n)1[y_n = h_{(s,b,d)}(\boldsymbol{x_n})]$ is 0.10. We arbitrarily choose $f_1 = h_{(-1,0.5,1)}$ as one of the minimizers. Then, $\epsilon_3 = \sum_{n=1}^{4} w_3(n)1[y_n = h_{(-1,0.5,1)}(\boldsymbol{x_n})] = 0.10$ and $\beta_3 = \dfrac{1}{2}\log\dfrac{1-\epsilon_1}{\epsilon} = 1.1$.

Solution to Question 3.6:

We have,

$$\begin{aligned} F(\boldsymbol{x}) &= \text{sign}[\beta_1 f_1(\boldsymbol{x}) + \beta_2 f_2(\boldsymbol{x}) + \beta_3 f_3(\boldsymbol{x})] \\ &= \text{sign}[0.55 h_{(1,-0.5,1)}(\boldsymbol{x}) + 0.83 h_{(1,0.5,2)}(\boldsymbol{x}) + 1.1 h_{(-1,0.5,1)}(\boldsymbol{x})]. \end{aligned} \tag{24}$$

Then, from Table 2,

$$\begin{aligned} F(\boldsymbol{x_1}) &= \text{sign}[0.55 \times 1 + 0.83 \times 1 + 1.1 \times 1] = 1, \\ F(\boldsymbol{x_2}) &= \text{sign}[0.55 \times (-1) + 0.83 \times (-1) + 1.1 \times 1] = -1, \\ F(\boldsymbol{x_3}) &= \text{sign}[0.55 \times 1 + 0.83 \times (-1) + 1.1 \times 1] = 1, \\ F(\boldsymbol{x_4}) &= \text{sign}[0.55 \times 1 + 0.83 \times (-1) + 1.1 \times (-1)] = -1. \end{aligned} \tag{25}$$

As can be seen, all the examples have been correctly labeled, that is, $\#\text{ of correctly labeled examples } = 4$.