Homework 4

Luis Vitor Zerkowski - 14895730

October 29, 2023

1 Maximum Marginal Classifier

1.a

Using the given hint, we can think of the final transformation ϕ as a composition of two transformations $\phi_2 \circ \phi_1$. Applying ϕ_1 first, we notice that data from each color stripe follows a pattern. For each stripe, $x^{(1)} - x^{(2)}$ approximately lies in a constant range. And for each class, this constant range varies constantly from one stripe to the next.

Just as an example, the right-most and bottom-most blue stripe has transformed values $\phi_1(\boldsymbol{x})$ around $x^{(1)}_{\phi_1} \in [8, 10], x^{(2)}_{\phi_1} \in [8, 10]$. And for consecutive stripes (from right to left and bottom to top), the transformed values vary -6, so the second right-most and bottom-most blue stripe has ranges around $x^{(1)}_{\phi_1} \in [2, 4], x^{(2)}_{\phi_1} \in [2, 4]$. Naturally, the analysis for the red class is analogous, only with different range values, which fill out the space between the blue stripes. Between the two blue stripes given as examples, we have one red stripe for which the transformation values $\phi_1(\boldsymbol{x})$ lie around $x^{(1)}_{\phi_1} \in [4, 8], x^{(2)}_{\phi_1} \in [4, 8]$.

The second transformation takes advantage of the patterns from the first transformation to make the data linearly separable. For the first coordinate of $\phi_2(\boldsymbol{x})$, we observe that points from different colors, will probably have different signs. This phenomenon can be explained by a few approximations. If we take the $x^{(1)}_{\phi_1} \in [2, 4], x^{(2)}_{\phi_1} \in [2, 4]$ as an example, we observe that $\frac{\pi}{2} < 2, 4 < \frac{3\pi}{2}$, which means that $\cos(x^{(1)}_{\phi_1}) < 0$. Since blue stripes ranges are separated by a constant value of $-6 \approx -2\pi$, the points from the next stripe will approximately lie on the same section of the trigonometric circle and thus have negative cosine.

For the red stripe example, $x^{(1)}_{\phi_1} \in [4, 8], x^{(2)}_{\phi_1} \in [4, 8]$, we know that 4 is only a bit smaller than $\frac{3\pi}{2}$ and $\frac{3\pi}{2} < 8 < \frac{5\pi}{2} \equiv \frac{\pi}{2}$, so points from this stripe will probably have $\cos(x^{(1)}_{\phi_1}) \geq 0$. Following the same pattern of stripe separation, $-6 \approx -2\pi$, the points from the next red stripe will approximately lie on the same section of the trigonometric circle and thus have positive cosine.

The final transformation $\phi = \phi_2 \circ \phi_1$, therefore, separates the classes on the $x^{(1)}_{\phi}$ axis by the sign, with blue points being negative and red points positive. And since the ranges we studied don't lead to a conclusion on the sine sign, datapoints for both classes are spread all over the $x^{(2)}_{\phi}$ axis.

1.b

In the transformed space, the decision boundary probably looks like a vertical line, splitting blue points and red points on a $x^{(1)}_{\phi}$ threshold given by b. Going a bit further into the details, since the $x^{(2)}_{\phi}$ is not an informative feature, we will probably have $w_0 \approx 1$ and $w_1 \approx 0$, leading to a decision boundary that depends only on the sign of $x^{(1)}_{\phi}$. But since there could be a few red points with negative $x^{(1)}_{\phi}$ coordinate too, we can horizontally adjust this vertical decision boundary with the parameter b.

Considering the example we went through on **1.a** for the red stripes, we notice that the inferior limit of the interval ($4 < \frac{3\pi}{2}$) still allows us to find some small negative cosine values for this class. With that in mind, b is probably negative, shifting the vertical boundary to the left and making it possible to identify all the points from the red class even if they have slightly negative $x^{(1)}_{\phi}$.

1.c

The primal Lagrangian is given by:

$$L(\boldsymbol{w}, b, \boldsymbol{\xi}, \{\lambda_n\}, \{\mu_n\}) = \frac{1}{2}||\boldsymbol{w}||_2^2 + C\sum_{n=1}^{N}\xi_n - \sum_{n=1}^{N}\lambda_n(t_n(\phi(\boldsymbol{x}_n)^T\boldsymbol{w} - b) - 1 + \xi_n) - \sum_{n=1}^{N}\mu_n\xi_n$$

1.d

There are exactly $6N$ KKT conditions. These are (for all $n \in 1, \ldots, N$):

$$t_n(\phi(\boldsymbol{x}_n)^T\boldsymbol{w} - b) - 1 + \xi_n \geq 0 \quad \text{(Primal feasibility for } (i)) \quad (1)$$

$$\lambda_n \geq 0 \quad \text{(Dual feasibility for } (i)) \quad (2)$$

$$\lambda_n(t_n(\phi(\boldsymbol{x}_n)^T\boldsymbol{w} - b) - 1 + \xi_n) = 0 \quad \text{(Complementary slackness for } (i)) \quad (3)$$

$$\xi_n \geq 0 \quad \text{(Primal feasibility for } (ii)) \quad (4)$$

$$\mu_n \geq 0 \quad \text{(Dual feasibility for } (ii)) \quad (5)$$

$$\mu_n\xi_n = 0 \quad \text{(Complementary slackness for } (ii)) \quad (6)$$

1.e

We start by computing $\frac{\partial}{\partial \boldsymbol{w}}L(\boldsymbol{w}, b, \boldsymbol{\xi}, \{\lambda_n\}, \{\mu_n\})$:

$$\frac{\partial}{\partial \boldsymbol{w}}L(\boldsymbol{w}, b, \boldsymbol{\xi}, \{\lambda_n\}, \{\mu_n\}) = \frac{\partial}{\partial \boldsymbol{w}}\left(\frac{1}{2}||\boldsymbol{w}||_2^2 + C\sum_{n=1}^{N}\xi_n - \sum_{n=1}^{N}\lambda_n(t_n(\phi(\boldsymbol{x}_n)^T\boldsymbol{w} - b) - 1 + \xi_n) - \sum_{n=1}^{N}\mu_n\xi_n\right) =$$

By the sum rule, removing the terms that don't depend on $\boldsymbol{w}$ and also rewriting $||\boldsymbol{w}||_2^2 = \boldsymbol{w}^T\boldsymbol{w}$ to use the product rule, we get:

$$= \frac{1}{2}\frac{\partial}{\partial \boldsymbol{w}}\boldsymbol{w}^T\boldsymbol{w} - \frac{\partial}{\partial \boldsymbol{w}}\sum_{n=1}^{N}\lambda_n(t_n(\phi(\boldsymbol{x}_n)^T\boldsymbol{w} - b) - 1 + \xi_n) = \boldsymbol{w}^T - \frac{\partial}{\partial \boldsymbol{w}}\sum_{n=1}^{N}\lambda_n(t_n(\phi(\boldsymbol{x}_n)^T\boldsymbol{w} - b) - 1 + \xi_n) =$$

Once again by the sum rule and removing terms that don't depend on $\boldsymbol{w}$, we have:

$$= \boldsymbol{w}^T - \sum_{n=1}^{N}\frac{\partial}{\partial \boldsymbol{w}}\lambda_n t_n \phi(\boldsymbol{x}_n)^T\boldsymbol{w} = \boldsymbol{w}^T - \sum_{n=1}^{N}\lambda_n t_n \frac{\partial}{\partial \boldsymbol{w}}\phi(\boldsymbol{x}_n)^T\boldsymbol{w} =$$

$$= \boldsymbol{w}^T - \sum_{n=1}^{N}\lambda_n t_n \phi(\boldsymbol{x}_n)^T$$

Now equaling it to zero, we get:

$$\frac{\partial}{\partial \boldsymbol{w}}L(\boldsymbol{w}, b, \boldsymbol{\xi}, \{\lambda_n\}, \{\mu_n\}) = 0 \implies \boldsymbol{w}^T - \sum_{n=1}^{N}\lambda_n t_n \phi(\boldsymbol{x}_n)^T = 0 \implies$$

$$\implies \boldsymbol{w}^T = \sum_{n=1}^{N}\lambda_n t_n \phi(\boldsymbol{x}_n)^T$$

Repeating the same process for $\frac{\partial}{\partial b}$, we have:

$$\frac{\partial}{\partial b}L(\boldsymbol{w}, b, \boldsymbol{\xi}, \{\lambda_n\}, \{\mu_n\}) = \frac{\partial}{\partial b}\left(\frac{1}{2}||\boldsymbol{w}||_2^2 + C\sum_{n=1}^{N}\xi_n - \sum_{n=1}^{N}\lambda_n(t_n(\phi(\boldsymbol{x}_n)^T\boldsymbol{w} - b) - 1 + \xi_n) - \sum_{n=1}^{N}\mu_n\xi_n\right) =$$

By the sum rule and removing the terms that don't depend on b, we get:

$$= -\frac{\partial}{\partial b}\sum_{n=1}^{N}\lambda_n(t_n(\phi(\boldsymbol{x}_n)^T\boldsymbol{w}-b)-1+\xi_n) = \sum_{n=1}^{N}\lambda_n t_n\frac{\partial}{\partial b}b =$$

$$= \sum_{n=1}^{N}\lambda_n t_n$$

Now equaling it to zero, we have:

$$\frac{\partial}{\partial b}L(\boldsymbol{w},b,\boldsymbol{\xi},\{\lambda_n\},\{\mu_n\}) = 0 \implies \sum_{n=1}^{N}\lambda_n t_n = 0$$

Finally, repeating the same process for $\frac{\partial}{\partial \xi_i}$, we get:

$$\frac{\partial}{\partial \xi_i}L(\boldsymbol{w},b,\boldsymbol{\xi},\{\lambda_n\},\{\mu_n\}) = \frac{\partial}{\partial \xi_i}\left(\frac{1}{2}||\boldsymbol{w}||_2^2 + C\sum_{n=1}^{N}\xi_n - \sum_{n=1}^{N}\lambda_n(t_n(\phi(\boldsymbol{x}_n)^T\boldsymbol{w}-b)-1+\xi_n) - \sum_{n=1}^{N}\mu_n\xi_n\right) =$$

By the sum rule and removing the terms that don't depend on ξ_i, we have:

$$= C\frac{\partial}{\partial \xi_i}\sum_{n=1}^{N}\xi_n - \frac{\partial}{\partial \xi_i}\sum_{n=1}^{N}\lambda_n(t_n(\phi(\boldsymbol{x}_n)^T\boldsymbol{w}-b)-1+\xi_n) - \frac{\partial}{\partial \xi_i}\sum_{n=1}^{N}\mu_n\xi_n =$$

$$= C\sum_{n=1}^{N}\delta_{in} - \sum_{n=1}^{N}\lambda_n\delta_{in} - \sum_{n=1}^{N}\mu_n\delta_{in} = C - \lambda_i - \mu_i$$

Now equaling it to zero, we have:

$$\frac{\partial}{\partial \xi_i}L(\boldsymbol{w},b,\boldsymbol{\xi},\{\lambda_n\},\{\mu_n\}) = 0 \implies C - \lambda_i - \mu_i = 0$$

We now proceed to derive the dual Lagrangian. We start by expanding the primal Lagrangian:

$$\frac{1}{2}||\boldsymbol{w}||_2^2 + C\sum_{n=1}^{N}\xi_n - \sum_{n=1}^{N}\lambda_n(t_n(\phi(\boldsymbol{x}_n)^T\boldsymbol{w}-b)-1+\xi_n) - \sum_{n=1}^{N}\mu_n\xi_n =$$

$$= \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{n=1}^{N}\xi_n - \sum_{n=1}^{N}\lambda_n t_n\phi(\boldsymbol{x}_n)^T\boldsymbol{w} + \sum_{n=1}^{N}\lambda_n t_n b + \sum_{n=1}^{N}\lambda_n - \sum_{n=1}^{N}\lambda_n\xi_n - \sum_{n=1}^{N}\mu_n\xi_n =$$

Using the derived $\boldsymbol{w}^T = \sum_{n=1}^{N}\lambda_n t_n\phi(\boldsymbol{x}_n)^T$ and $\sum_{n=1}^{N}\lambda_n t_n = 0$, we get:

$$= \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{n=1}^{N}\xi_n - \boldsymbol{w}^T\boldsymbol{w} + b*0 + \sum_{n=1}^{N}\lambda_n - \sum_{n=1}^{N}(\lambda_n+\mu_n)\xi_n =$$

$$= -\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{n=1}^{N}\xi_n + \sum_{n=1}^{N}\lambda_n - \sum_{n=1}^{N}(\lambda_n+\mu_n)\xi_n =$$

Using the derived $C - \lambda_i - \mu_i = 0 \implies C = \lambda_i + \mu_i$, we have:

$$= -\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{n=1}^{N}\xi_n + \sum_{n=1}^{N}\lambda_n - C\sum_{n=1}^{N}\xi_n = -\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \sum_{n=1}^{N}\lambda_n =$$

And now using $\boldsymbol{w}^T = \sum_{n=1}^{N}\lambda_n t_n\phi(\boldsymbol{x}_n)^T$ again to remove $\boldsymbol{w}$, we get:

$$= -\frac{1}{2}\left(\sum_{n=1}^{N}\lambda_n t_n \phi(\boldsymbol{x}_n)^T\right)\left(\sum_{n=1}^{N}\lambda_n t_n \phi(\boldsymbol{x}_n)\right) + \sum_{n=1}^{N}\lambda_n =$$

$$= -\frac{1}{2}\left(\sum_{n=1}^{N}\sum_{m=1}^{N}\lambda_n\lambda_m t_n t_m \phi(\boldsymbol{x}_n)^T\phi(\boldsymbol{x}_m)\right) + \sum_{n=1}^{N}\lambda_n$$

So we finally reach an expression for the dual Lagrangian given by:

$$\tilde{L}(\{\lambda_n\},\{\mu_n\}) = -\frac{1}{2}\left(\sum_{n=1}^{N}\sum_{m=1}^{N}\lambda_n\lambda_m t_n t_m \phi(\boldsymbol{x}_n)^T\phi(\boldsymbol{x}_m)\right) + \sum_{n=1}^{N}\lambda_n$$

1.f

The explicit form of $k(\boldsymbol{x}_n, \boldsymbol{x}_m)$ in the final solution to the dual Lagrangian is $k(\boldsymbol{x}_n, \boldsymbol{x}_m) = \phi(\boldsymbol{x}_n)^T\phi(\boldsymbol{x}_m)$. Using the given transformation $\phi(\boldsymbol{x}) = [\cos(x^{(1)} - x^{(2)}), \sin(x^{(1)} - x^{(2)})]^T$, we have:

$$k(\boldsymbol{x}_n, \boldsymbol{x}_m) = \phi(\boldsymbol{x}_n)^T\phi(\boldsymbol{x}_m) =$$
$$= \cos(x_n^{(1)} - x_n^{(2)})\cos(x_m^{(1)} - x_m^{(2)}) + \sin(x_n^{(1)} - x_n^{(2)})\sin(x_m^{(1)} - x_m^{(2)})$$

Leading to a dual Lagrangian of the form:

$$\tilde{L}(\{\lambda_n\},\{\mu_n\}) = -\frac{1}{2}\left(\sum_{n=1}^{N}\sum_{m=1}^{N}\lambda_n\lambda_m t_n t_m(\cos(x_n^{(1)} - x_n^{(2)})\cos(x_m^{(1)} - x_m^{(2)}) + \sin(x_n^{(1)} - x_n^{(2)})\sin(x_m^{(1)} - x_m^{(2)}))\right) +$$
$$+ \sum_{n=1}^{N}\lambda_n$$

1.g

The decision function is given by $\boldsymbol{w}^T\phi(\boldsymbol{x}) - b$. Using the primal variable we derived in question **1.e** $\boldsymbol{w}^T = \sum_{n=1}^{N}\lambda_n t_n \phi(\boldsymbol{x}_n)^T$, we can rewrite the decision function as $\sum_{n=1}^{N}\lambda_n t_n \phi(\boldsymbol{x}_n)^T\phi(\boldsymbol{x}) - b$. Using the kernel trick to rewrite it one last time, we have $\sum_{n=1}^{N}\lambda_n t_n k(\boldsymbol{x}_n, \boldsymbol{x}) - b$. Thus the classification of a new point $\boldsymbol{x}^*$ is given by:

$$t^* = \begin{cases} 1, & \text{if } \sum_{n=1}^{N}\lambda_n t_n k(\boldsymbol{x}_n, \boldsymbol{x}^*) - b \geq 0 \\ -1, & \text{otherwise} \end{cases}$$

1.h

For this question, we repeat the conditions defined in exercise **1.d** to make the KKT conditions indexes explicit:

$$t_n(\phi(\boldsymbol{x}_n)^T\boldsymbol{w} - b) - 1 + \xi_n \geq 0 \quad \text{(Primal feasibility for } (i)) \quad (1)$$
$$\lambda_n \geq 0 \quad \text{(Dual feasibility for } (i)) \quad (2)$$
$$\lambda_n(t_n(\phi(\boldsymbol{x}_n)^T\boldsymbol{w} - b) - 1 + \xi_n) = 0 \quad \text{(Complementary slackness for } (i)) \quad (3)$$
$$\xi_n \geq 0 \quad \text{(Primal feasibility for } (ii)) \quad (4)$$
$$\mu_n \geq 0 \quad \text{(Dual feasibility for } (ii)) \quad (5)$$
$$\mu_n\xi_n = 0 \quad \text{(Complementary slackness for } (ii)) \quad (6)$$

1.h.1

For a point $\boldsymbol{x}_n$ outside the margin that is correctly classified, we have $\xi_n = 0$. Using condition (1) and knowing that the point is not only correctly classified, but outside the margin, we have $t_n(\phi(\boldsymbol{x}_n)^T\boldsymbol{w} - b) - 1 > 0$. Now using condition (3), we conclude that $\lambda_n = 0$. Finally, because of conditions (5) and (6), but also using the stationary point for ξ_n, we conclude that $\mu_n = C > 0$.

1.h.2

For a point $\boldsymbol{x}_n$ on the margin, we still have $\xi_n = 0$, but this time we also have $t_n(\phi(\boldsymbol{x}_n)^T\boldsymbol{w} - b) - 1 = 0$, still respecting condition (1). This leaves us with a λ_n that can take any non-negative value, respecting conditions (2) and (3), and a μ_n that can also take any non-negative value, respecting conditions (5) and (6), but in a way that $\lambda_n + \mu_n = C > 0$, respecting the stationary point for ξ_n.

1.h.3

For a point $\boldsymbol{x}_n$ within the margin that is correctly classified, we have $\xi_n > 0$ and $t_n(\phi(\boldsymbol{x}_n)^T\boldsymbol{w} - b) - 1 = 0$, still respecting condition (1). Using condition (3), we get to $\lambda_n\xi_n = 0$, but since $\xi_n > 0$, we conclude that $\lambda_n = 0$. Finally, using condition (6) and the fact that $\xi_n > 0$, we conclude that $\mu_n = 0$.

1.h.4

For a point $\boldsymbol{x}_n$ on the wrong side of the decision boundary, we have $\xi_n > 1$.

1.i

Using the stationary point for ξ_n and assuming we have found the optimal λ_n, we have:

$$C - \lambda_n - \mu_n = 0 \implies \mu_n = C - \lambda_n$$

Now with the optimal values for both λ_n and μ_n, we can solve for the primal variables. For the stationary point for $\boldsymbol{w}$, we get:

$$\boldsymbol{w} = \sum_{n=1}^{N} \lambda_n t_n \phi(\boldsymbol{x}_n)$$

Now using the KKT conditions (3) with support vectors ($\lambda_n > 0$), we have:

$$t_n(\phi(\boldsymbol{x}_n)^T\boldsymbol{w} - b) - 1 + \xi_n = 0$$

From the above equation, we can solve for b because we know that for support vectors, $\xi_n = 0$. So we get:

$$t_n(\phi(\boldsymbol{x}_n)^T\boldsymbol{w} - b) - 1 = 0 \implies t_n\phi(\boldsymbol{x}_n)^T\boldsymbol{w} - t_n b - 1 = 0 \implies$$

$$\implies b = \phi(\boldsymbol{x}_n)^T\boldsymbol{w} - \frac{1}{t_n} \implies b = \sum_{m=1}^{N} \lambda_m t_m k(\boldsymbol{x}_n, \boldsymbol{x}_m) - \frac{1}{t_n}$$

So have $\boldsymbol{w} = \sum_{n=1}^{N} \lambda_n t_n \phi(\boldsymbol{x}_n)$ and $b = \sum_{m=1}^{N} \lambda_m t_m k(\boldsymbol{x}_n, \boldsymbol{x}_m) - \frac{1}{t_n}$.For ξ_n, we solved for $\boldsymbol{x}_n$ that are support vectors, because then $\xi_n = 0$.

1.j

1.j.1

Consider $\mathbb{I}[a]$ a function that returns 1 when condition a is fulfilled and 0 otherwise. Then we have:

$$\phi(\boldsymbol{x}) = [\mathbb{I}[|x^{(2)}| < |x^{(1)}|], \mathbb{I}[|x^{(2)}| > |x^{(1)}|]]^T$$

1.j.2

This one can be transformed by a slightly different version of the transformation provided in the beginning of the exercise:

$$\phi(\boldsymbol{x}) = [\cos(x^{(1)} + x^{(2)}), \sin(x^{(1)} + x^{(2)})]^T$$