UNIVERSITY OF AMSTERDAM

University of Amsterdam

Homework Assignment 4

Machine Learning I

2024

Pedro M. P. Curvo
15713725

Contents

1 Maximum Margin Classifier

a)

By applying the transformation $\phi(x)$ to the dataset, the points are mapped onto a circumference in the transformed space, with the boundary alternating between the *red* and *blue* points. This transformation can be decomposed into $\phi(x) = \phi_1(x) \circ \phi_2(x)$, where $\phi_1(x)$ maps the points as $\phi_1(x) = (x^{(1)} - x^{(2)}, x^{(1)} - x^{(2)})$, projecting the data along the principal component $[1, -1]$ and forming a diagonal where the points alternate between the red and blue regions in a periodic way. The second transformation, $\phi_2(x)$, applies trigonometric functions, $\phi_2(x) = (\cos(x^{(1)}), \sin(x^{(2)}))$, which maps the points onto a circular path, resulting in the points alternating between the red and blue regions along the circumference.

b)

The decision boundary is defined by the equation:

$$y(x) = \boldsymbol{w}^T \phi(x) + b = w_0 \cos(x^{(1)} - x^{(2)}) + w_1 \sin(x^{(1)} - x^{(2)}) - b$$

If $y(x) = 0$, then the decision boundary is defined by the equation:

$$\begin{aligned} y(x) &= 0 \\ w_0 cos(x^{(1)} - x^{(2)}) + w_1 \sin(x^{(1)} - x^{(2)}) - b &= 0 \\ \sin(x^{(1)} - x^{(2)}) &= -\frac{w_0}{w_1} \cos(x^{(1)} - x^{(2)}) + \frac{b}{w_1} \end{aligned}$$

In the new space, $\phi(x)^{(1)} = \cos(x^{(1)} - x^{(2)})$ and $\phi(x)^{(2)} = \sin(x^{(1)} - x^{(2)})$, so this becomes:

$$\begin{aligned} y(x) &= 0 \\ \phi(x)^{(2)} &= -\frac{w_0}{w_1} \phi(x)^{(1)} + \frac{b}{w_1} \end{aligned}$$

Which is the equation of a straight line in the new space, hence the w is the vector that controls the slope of the decision boundary, and b is the bias term that controls the offset of the decision boundary shifting it up or down.

c)

By introducing the Lagrange multipliers λ_n for the first constraint and μ_n for the second constraint, we can write the primal Lagrangian as:

$$\mathcal{L}(w, b, \lambda, \mu) = \frac{1}{2}||w||^2 + C\sum_{n=1}^{N}\xi_n - \sum_{n=1}^{N}\lambda_n\left(t_n(\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x_n}) + b) - 1 + \xi_n\right) - \sum_{n=1}^{N}\mu_n\xi_n$$

With:

$$\xi_n \geq 0, \quad \lambda_n \geq 0, \quad \mu_n \geq 0$$

d)

The KKT conditions for the primal problem are:

Primal Feasibility:

$$\begin{aligned} t_n(\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x_n}) + b) &\geq 1 - \xi_n, \quad \forall n \\ \xi_n &\geq 0 \quad \forall n \end{aligned}$$

Dual Feasibility:

$$\begin{aligned} \lambda_n &\geq 0 \\ \mu_n &\geq 0 \end{aligned}$$

Complementary Slackness:

$$\begin{aligned} \lambda_n \left(t_n(\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x_n}) + b) - 1 + \xi_n\right) &= 0, \quad \forall n \\ \mu_n\xi_n &= 0, \quad \forall n \end{aligned}$$

With this we have 6 KKT conditions per point, 3 for each constraint condition in the primal problem. So, in the overall problem we have $6N$ KKT conditions.

e)

To derive the dual Lagrangian $\tilde{\mathcal{L}}(\lambda)$, we need to find the stationary conditions for the primal Lagrangian $\mathcal{L}(w, b, \lambda, \mu)$ by taking the derivatives with respect to w, b and ξ_n and setting them to zero, to find the optima of the primal problem.

$$\frac{\partial \mathcal{L}}{\partial w} = 0$$

$$0 = \frac{\partial}{\partial w}\frac{1}{2}||w||^2 + C\sum_{n=1}^{N}\xi_n - \sum_{n=1}^{N}\lambda_n\left(t_n(\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x_n}) + b) - 1 + \xi_n\right) - \sum_{n=1}^{N}\mu_n\xi_n$$

$$0 = w - \sum_{n=1}^{N}\lambda_n t_n \boldsymbol{\phi}(\boldsymbol{x_n})$$

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

$$\frac{\partial \mathcal{L}}{\partial b} = 0$$

$$0 = \frac{\partial}{\partial b}\frac{1}{2}||w||^2 + C\sum_{n=1}^{N}\xi_n - \sum_{n=1}^{N}\lambda_n\left(t_n(\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x_n}) + b) - 1 + \xi_n\right) - \sum_{n=1}^{N}\mu_n\xi_n$$

$$0 = -\sum_{n=1}^{N}\lambda_n t_n$$

$$0 = \sum_{n=1}^{N}\lambda_n t_n$$

$$\frac{\partial \mathcal{L}}{\partial \xi_n} = 0$$

$$0 = \frac{\partial}{\partial \xi_n}\frac{1}{2}||w||^2 + C\sum_{n=1}^{N}\xi_n - \sum_{n=1}^{N}\lambda_n\left(t_n(\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x_n}) + b) - 1 + \xi_n\right) - \sum_{n=1}^{N}\mu_n\xi_n$$

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

$$\lambda_n = C - \mu_n$$

Now by substituting the values into the primal Lagrangian, we can obtain the dual Lagrangian $\tilde{\mathcal{L}}(\lambda)$:

$$
\begin{aligned}
\tilde{\mathcal{L}}(\lambda) &= \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{n=1}^{N}\xi_n - \sum_{n=1}^{N}\lambda_n\left(t_n(\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x_n}) + b) - 1 + \xi_n\right) - \sum_{n=1}^{N}\mu_n\xi_n \\
&= \boldsymbol{w}^T\left(\frac{1}{2}\boldsymbol{w} - \sum_{n=1}^{N}\lambda_n t_n\boldsymbol{\phi}(\boldsymbol{x_n})\right) + C\sum_{n=1}^{N}\xi_n - \sum_{n=1}^{N}\lambda_n t_n b + \sum_{n=1}^{N}\lambda_n - \sum_{n=1}^{N}\lambda_n\xi_n - \sum_{n=1}^{N}\mu_n\xi_n \\
&= \boldsymbol{w}^T\left(\frac{1}{2}\boldsymbol{w} - \boldsymbol{w}\right) - b\sum_{n=1}^{N}\lambda_n t_n + \sum_{n=1}^{N}\lambda_n + C\sum_{n=1}^{N}\xi_n - \sum_{n=1}^{N}\lambda_n\xi_n - \sum_{n=1}^{N}\mu_n\xi_n \\
&= -\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \sum_{n=1}^{N}\lambda_n + C\sum_{n=1}^{N}\xi_n - \sum_{n=1}^{N}(\lambda_n + \mu_n)\xi_n \quad \text{, since } \sum_{n=1}^{N}\lambda_n t_n = 0 \\
&= -\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \sum_{n=1}^{N}\lambda_n \quad \text{, since } \lambda_n = C - \mu_n \\
&= -\frac{1}{2}\sum_{n=1}^{N}\lambda_n t_n\boldsymbol{\phi}(\boldsymbol{x_n})^T\sum_{m=1}^{N}\lambda_m t_m\boldsymbol{\phi}(\boldsymbol{x_m}) + \sum_{n=1}^{N}\lambda_n \quad \text{, since } w = \sum_{n=1}^{N}\lambda_n t_n\boldsymbol{\phi}(\boldsymbol{x_n}) \\
&= -\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\lambda_n\lambda_m t_n t_m\boldsymbol{\phi}(\boldsymbol{x_n})^T\boldsymbol{\phi}(\boldsymbol{x_m}) + \sum_{n=1}^{N}\lambda_n
\end{aligned}
$$

With the remaining constraints of the dual problem:

$$
\begin{aligned}
0 \leq \lambda_n \leq C, \quad \forall n \\
\sum_{n=1}^{N}\lambda_n t_n = 0
\end{aligned}
$$

The dual optimization problem is then:

$$
\lambda^* = arg\min_{\lambda}\tilde{\mathcal{L}}(\lambda) = arg\min_{\lambda}\left(-\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\lambda_n\lambda_m t_n t_m\boldsymbol{\phi}(\boldsymbol{x_n})^T\boldsymbol{\phi}(\boldsymbol{x_m}) + \sum_{n=1}^{N}\lambda_n\right)
$$

Subject to the constraints presented above.

f)

The kernel function $\kappa(x, x')$ in the dual Lagrangian is given by:

$$\begin{aligned}
\kappa(x_n, x_m) &= \boldsymbol{\phi(x_n)}^T \boldsymbol{\phi(x_m)} \\
&= \cos(x_n^{(1)} - x_n^{(2)}) \cos(x_m^{(1)} - x_m^{(2)}) + \sin(x_n^{(1)} - x_n^{(2)}) \sin(x_m^{(1)} - x_m^{(2)}) \\
&= \cos((x_n^{(1)} - x_n^{(2)}) - (x_m^{(1)} - x_m^{(2)})) \quad \text{, by the trigonometric identity}
\end{aligned}$$

g)

To classify a new point x^*, we need to evaluate $y(x^*)$, but since we have the dual solution λ_n, we can write $y(x^*)$ as:

$$\begin{aligned} y(x^*) &= \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x^*}) + b \\ &= \sum_{n=1}^{N} \lambda_n t_n \kappa(x_n, x_m^*) + b \quad \text{, by the definition of } w \text{ in the dual solution} \end{aligned}$$

Where $\kappa(x_n, x_m)$ is the kernel function $\boldsymbol{\phi}(\boldsymbol{x_n})^T \boldsymbol{\phi}(\boldsymbol{x_m})$.

To evaluate the point we can just evaluate the sign of $y(x^*)$, if $y(x^*) > 0$, then the point is classified as $t = 1$, if $y(x^*) < 0$, then the point is classified as $t = -1$. We can do this because we already know the λ_n and b values, and we can evaluate the kernel function $k(x_n, x_m^*)$ for all the support vectors.

h)

For a new prediticion we have that:

$$y(x) = \sum_{n=1}^{N} \lambda_n t_n \kappa(x_n, x_m) + b$$

with

$$0 \leq \lambda_n \leq C, \quad \forall n \quad \text{and} \quad \sum_{n=1}^{N} \lambda_n t_n = 0$$

And from the KKT conditions we have that:

$$\begin{aligned} \lambda_n \geq 0, \quad \mu_n \geq 0 \\ \lambda_n \left(t_n(\boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x_n}) + b) - 1 + \xi_n \right) = 0, \quad \mu_n \xi_n = 0 \\ t_n(\boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x_n}) + b) \geq 1 - \xi_n, \quad \xi_n \geq 0 \end{aligned}$$

Now, with this constraints and by the way we define the problem, than if $\lambda_n > 0$ we have a support vector that is on the region that is defined by the margin and the decision boundary.

Now, if $\lambda_n = 0$, then $t_n y(x_n) = 1 - \xi_n$. With this we can conclude several things:

- If $\lambda_n < C$, then $\mu_n > 0$ because $\lambda_n = C - \mu_n$. Meaning that $\xi_n = 0$, since $\mu_n \xi_n = 0$. Since ξ_n defines distance from the margin, if $\xi_n = 0$, then the point is on the margin.
- If $\lambda_n = C$, then $\mu_n = 0$ because $\lambda_n = C - \mu_n$. Meaning that $\xi_n > 0$, since $\mu_n \xi_n = 0$. Now, by the definition of ξ_n, if $\xi_n > 1$, then the point is on the wrong side of the margin, and is misclassified. If $\xi_n \leq 1$, then the point is within the margin and is correctly classified.

With this we can resume in the following:

- Outside the margin, correctly classified: $\lambda_n = 0$
- On the margin (support vector), correctly classified: $0 < \lambda_n < C$, $\mu_n > 0$, $\xi_n = 0$
- Inside the margin, correctly classified: $\lambda_n = C$, $\mu_n = 0$, $0 \leq \xi_n \leq 1$
- Inside the margin, misclassified: $\lambda_n = C$, $\mu_n = 0$, $\xi_n > 1$

i)

To find the optimal values for the dual variables μ_n given the optimal values for the dual variables λ_n, we can use the the previous shown conditions that:

$$\lambda_n = C - \mu_n$$

With this we can find the optimal values for μ_n.

Now, by using the KKT conditions, we can solve for the primal variables w, b and ξ_n:

$$\begin{aligned} w &= \sum_{n=1}^{N} \lambda_n t_n \boldsymbol{\phi}(\boldsymbol{x_n}) \\ b &= \frac{1}{N} \sum_{n=1}^{N} \left(t_n - \sum_{m=1}^{N} \lambda_m t_m \kappa(x_m, x_n) \right) \quad \text{, by inverting the problem and isolating } b \\ \xi_n &= \max(0, 1 - t_n(\boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x_n}) + b)) \end{aligned}$$

Where in b, we are taking the average of the support vectors N just to ensure that we are getting a good estimate of the bias, since the program solving for the dual problem might not be able to find the exact value of b.

In ξ_n we are taking the maximum of 0 and $1 - t_n(\boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x_n}) + b)$, since this shows if the margin is being violated or not, i.e., if the points is correctly classified and lies on/or outside the margin then $\xi_n = 0$. For points misclassified or inside the margin, then $\xi_n > 0$. Now, according to the restritions, $\xi_n \leq 1 - t_n(\boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x_n}) + b)$, so taking the maximum of 0 and $1 - t_n(\boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x_n}) + b)$ ensures that the ξ_n is within the bounds.

j)

Dataset 1:

This dataset looks like the XOR function but rotated, so first we can rotate the dataset by 45 degrees counter clockwise:

$$\begin{bmatrix}\phi_1(x)^{(1)}\\ \phi_1(x)^{(2)}\end{bmatrix} = \begin{bmatrix}\cos(45) & -\sin(45)\\ \sin(45) & \cos(45)\end{bmatrix}\begin{bmatrix}x^{(1)}\\ x^{(2)}\end{bmatrix} = \begin{bmatrix}\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}}\\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\end{bmatrix}\begin{bmatrix}x^{(1)}\\ x^{(2)}\end{bmatrix} = \begin{bmatrix}\frac{x^{(1)}-x^{(2)}}{\sqrt{2}}\\ \frac{x^{(1)}+x^{(2)}}{\sqrt{2}}\end{bmatrix}$$

Then, we can see that the *blue* points, in the new space, have the characteristic that $x^{(1)} \cdot x^{(2)} < 0$, and the *red* points have the characteristic that $x^{(1)} \cdot x^{(2)} > 0$, because they are going to be in different quadrants. So we can just apply the transformation $\phi_2(x) = ((x^{(1)} \cdot x^{(2)}), (x^{(1)} \cdot x^{(2)}))$ to the dataset and then they will be separated by a line passing through quadrants 2 and 4.

The total trasnformation is then:

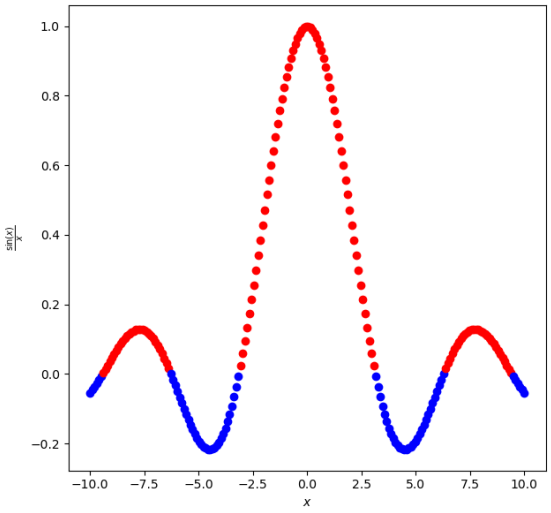
$$\phi(x) = (\frac{(x^{(1)} - x^{(2)})(x^{(1)} + x^{(2)})}{2}, \frac{(x^{(1)} - x^{(2)})(x^{(1)} + x^{(2)})}{2})$$

Dataset 2:

Looking into the dataset, we can see some periodicity in the direction $[1, 1]$ but with varying amplitude, so we can apply first a trasnformation that projects the data onto that component $[1, 1]$ such as $\phi_1(x) = (x^{(1)} + x^{(2)}, x^{(1)} + x^{(2)})$. Then we can apply a transformation that maps the data onto the function $\frac{\sin x}{x}$, since the periodicity looks like the function and the width can be estimated. With that the total transformation is:

$$\phi(x) = \left((x^{(1)} + x^{(2)}), \frac{\sin(x^{(1)} + x^{(2)})}{x^{(1)} + x^{(2)}}\right)$$

Looking something like:



To be in mind that constant values of the function, such as the constants for dilation and translation are not defined, but can be estimated by the data. Moreover, each periodic signal can be expressed as

a fourier series, so we could also use a sum of sins and cosines to map the periodicity of the data as the second coordinate of the transformation.

Dataset 3:

For this dataset there are several approaches we could try. First, we can rotate the dataset by 45 degrees counter clockwise and apply the transformation from the previous dataset $sin(x)/x$ to map the data to a 2D space where the data is linearly separable. Besides that, after the rotation we can also apply a sum of sins and cosines since its periodicity can be mapped by a fourier series. Hence, one possible solution could be the following transformation:

$$\phi(x) = \left(\frac{(x^{(1)} - x^{(2)})(x^{(1)} + x^{(2)})}{2}, \frac{\sin\left(\frac{(x^{(1)} - x^{(2)})(x^{(1)} + x^{(2)})}{2}\right)}{\frac{(x^{(1)} - x^{(2)})(x^{(1)} + x^{(2)})}{2}} \right)$$

Another solution, since the data looks like two spirals intertwined in a 3D space, but seen from above, we could try to first estimate the radius of the spiral, the angle of each point relative to the center of the spiral, and width of the spiral by the distance between "wavelengths". With this we can create the typical 3D space of a spiral, where the axis are given by $cos(\theta)$, $sin(\theta)$ and θ. Now, we could simply apply the function arccos to one of the trigonometric axis to map the data to a any 2D space where this is one axis. Why this function? So the spirals are intertwined, but not overlapping, meaning that the angle of one is shifted by a constant value relative to the other, so the arccos function could map the data to a 2D space where the spirals are linearly separable, because it would look something like θ for one spiral and $\theta + \delta$ for the other.

Due to the complexity of what was presented above, an alternative would be to use a RBF kernel, since it can map the data into a space where it is linearly separable, due to the periodicity of the data.