Andrii Skliar
11636785
andrii.skliar@student.uva.nl

Homework Assignment 1
Machine Learning 1, 17/18

2019-01-08

1 Lagrange Multipliers: Warm-up

1.

$$f(x_1, x_2) = 1 - x_1^2 - 2x_2^2$$
$$g(x_1, x_2) = x_1 + x_2 - 1 = 0$$

Therefore, Lagrangian looks as following :

$$L(x_1, x_2, \lambda) = 1 - x_1^2 - 2x_2^2 - \lambda(x_1 + x_2 - 1)$$

Taking derivatives with respect to x_1, x_2 and λ, we get the following system of equations:

$$\begin{cases} -2x_1 - \lambda = 0 \\ -4x_2 - \lambda = 0 \\ x_1 + x_2 - 1 = 0 \end{cases} \quad \begin{cases} -2x_1 = \lambda \\ -4x_2 = \lambda \\ x_1 + x_2 = 1 \end{cases} \quad \begin{cases} x_1 = 2x_2 \\ \lambda = -2x_1 \\ 2x_2 + x_2 = 1 \end{cases} \quad \begin{cases} x_2 = \frac{1}{3} \\ x_1 = \frac{2}{3} \\ \lambda = -\frac{4}{3} \end{cases} \quad \Rightarrow \begin{cases} x_1* = \frac{2}{3} \\ x_2* = \frac{1}{3} \\ \lambda* = -\frac{4}{3} \end{cases}$$

$$\begin{cases} x_1* = \frac{2}{3} \\ x_2* = \frac{1}{3} \\ f(x_1*, x_2*) = \frac{1}{3} \end{cases}$$

2.

$$f(x_1, x_2) = 1 - x_1^2 - x_2^2$$
$$g(x_1, x_2) = -x_1 - x_2 + 1 \leq 0$$

Therefore, Lagrangian and KKT conditions look as following:

$$L(x_1, x_2, \lambda) = 1 - x_1^2 - x_2^2 - \lambda(-x_1 - x_2 + 1)$$

$$\begin{cases} \nabla f = \lambda \nabla g \\ g(x_1, x_2) \leq 0 \\ \lambda g(x_1, x_2) = 0 \\ \lambda \geq 0 \end{cases}$$

Taking derivatives with respect to x_1, x_2 and λ, we get the following system of equations:

$$\begin{cases} -2x_1 + \lambda = 0 \\ -2x_2 + \lambda = 0 \\ \lambda \geq 0 \\ \lambda(-x_1 - x_2 + 1) = 0 \\ g(x_1, x_2) \leq 0 \end{cases}$$

$$1) \begin{cases} -2x_1 + \lambda = 0 \\ -2x_2 + \lambda = 0 \\ \lambda = 0 \\ g(x_1, x_2) \leq 0 \end{cases} \quad \begin{cases} x_1 = x_2 \\ \lambda = 0 \\ -2x_1 = \lambda \\ g(x_1, x_2) \leq 0 \end{cases} \quad \begin{cases} x_1 = x_2 = 0 \\ \lambda = 0 \\ g(x_1, x_2) = 1 \geq 0 \end{cases} \Rightarrow \text{Not valid solution.}$$

$$2) \begin{cases} -2x_1 + \lambda = 0 \\ -2x_2 + \lambda = 0 \\ g(x_1, x_2) = 0 \\ \lambda > 0 \end{cases} \quad \begin{cases} x_1 = x_2 = \frac{\lambda}{2} \\ x_1 + x_2 = 1 \\ \lambda > 0 \end{cases} \quad \begin{cases} x_1* = x_2* = \frac{1}{2} \\ \lambda* = 1 \\ g(x_1*, x_2*) = 0 \leq 0 \\ f(x_1*, x_2*) = \frac{1}{2} \end{cases} \Rightarrow \text{Valid solution.}$$

$$\begin{cases} x_1* = \frac{1}{2} \\ x_2* = \frac{1}{2} \\ f(x_1*, x_2*) = \frac{1}{2} \end{cases}$$

3.

$$f(x_1, x_2, x_3) = x_1 + 2x_2 - 2x_3$$
$$g(x_1, x_2, x_3) = x_1^2 + x_2^2 + x_3^2 - 1 = 0$$

Therefore, Lagrangian looks as following :

$$L(x_1, x_2, \lambda) = x_1 + 2x_2 - 2x_3 - \lambda(x_1^2 + x_2^2 + x_3^2 - 1)$$

Taking derivatives with respect to x_1, x_2 and λ, we get the following system of equations:

$$\begin{cases} 1 - 2\lambda x_1 = 0 \\ 2 - 2\lambda x_2 = 0 \\ -2 - 2\lambda x_3 = 0 \\ x_1^2 + x_2^2 + x_3^2 - 1 = 0 \end{cases} \quad \begin{cases} x_1 = \frac{1}{2\lambda} \\ x_2 = \frac{1}{\lambda} \\ x_3 = -\frac{1}{\lambda} \\ x_1^2 + x_2^2 + x_3^2 = 1 \end{cases} \quad \begin{cases} x_1 = \frac{1}{2\lambda} \\ x_2 = \frac{1}{\lambda} \\ x_3 = -\frac{1}{\lambda} \\ \frac{9}{4\lambda^2} = 1 \end{cases}$$

$$1) \begin{cases} \lambda = \frac{3}{2} \\ x_1 = \frac{1}{3} \\ x_2 = \frac{2}{3} \\ x_3 = -\frac{2}{3} \\ f(x_1, x_2, x_3) = 3 \end{cases}$$

or

$$2) \begin{cases} \lambda = -\frac{3}{2} \\ x_1 = -\frac{1}{3} \\ x_2 = -\frac{2}{3} \\ x_3 = \frac{2}{3} \\ f(x_1, x_2, x_3) = -3 \end{cases}$$

Therefore, $x_1* = \frac{1}{3}, x_2* = \frac{2}{3}, x_3* = -\frac{2}{3}$ maximize f.

$$\begin{cases} x_1* = \frac{1}{3} \\ x_2* = \frac{2}{3} \\ x_3* = -\frac{2}{3} \\ f(x_1*, x_2*, x_3*) = 3 \end{cases}$$

4.

$$f(x_1, x_2) = 1 - x_1^2 - x_2^2$$
$$g(x_1, x_2) = x_1 + x_2 - 1 \leq 0$$

Therefore, Lagrangian and KKT conditions look as following:

$$L(x_1, x_2, \lambda) = 1 - x_1^2 - x_2^2 - \lambda(x_1 + x_2 - 1)$$

$$\begin{cases} \nabla f = \lambda \nabla g \\ g(x_1, x_2) \geq 0 \\ \lambda g(x_1, x_2) = 0 \\ \lambda \geq 0 \end{cases}$$

Taking derivatives with respect to x_1, x_2 and λ, we get the following system of equations:

$$\begin{cases} -2x_1 - \lambda = 0 \\ -2x_2 - \lambda = 0 \\ \lambda \geq 0 \\ \lambda(x_1 + x_2 - 1) = 0 \\ g(x_1, x_2) \leq 0 \end{cases}$$

$$1) \begin{cases} -2x_1 - \lambda = 0 \\ -2x_2 - \lambda = 0 \\ \lambda = 0 \\ g(x_1, x_2) \leq 0 \end{cases} \quad \begin{cases} x_1 = x_2 = -\lambda = 0 \\ g(x_1, x_2) = -1 \leq 0 \\ f(x_1, x_2) = 1 \end{cases} \quad \Rightarrow \text{Valid solution.}$$

$$2) \begin{cases} -2x_1 - \lambda = 0 \\ -2x_2 - \lambda = 0 \\ g(x_1, x_2) = 0 \\ \lambda > 0 \end{cases} \quad \begin{cases} x_1 = x_2 = -\frac{\lambda}{2} \\ x_1 + x_2 = 1 \\ g(x_1, x_2) = 0 \\ \lambda > 0 \end{cases} \quad \begin{cases} x_1 = x_2 = \frac{1}{2} \\ \lambda = -1 < 0 \\ g(x_1, x_2) = 0 \\ \lambda > 0 \end{cases} \quad \Rightarrow \text{Not valid solution.}$$

$$\begin{cases} x_1* = 0 \\ x_2* = 0 \\ f(x_1*, x_2*) = 1 \end{cases}$$

5.

$$f(x_1, x_2) = 6x^{\frac{2}{3}} y^{\frac{1}{2}}$$
$$g(x_1, x_2) = 4x + 3y - 7000 \leq 0$$

Therefore, Lagrangian and KKT conditions look as following:

$$L(x_1, x_2, \lambda) = 6x^{\frac{2}{3}} y^{\frac{1}{2}} - \lambda(4x + 3y - 7000)$$

$$\begin{cases} \nabla f = \lambda \nabla g \\ g(x_1, x_2) \leq 0 \\ \lambda g(x_1, x_2) = 0 \\ \lambda \geq 0 \end{cases}$$

Taking derivatives with respect to x_1, x_2 and λ, we get the following system of equations:

$$\begin{cases} 4x^{-\frac{1}{3}} y^{\frac{1}{2}} + 4\lambda = 0 \\ 3x^{\frac{2}{3}} y^{-\frac{1}{2}} + 3\lambda = 0 \\ \lambda \geq 0 \\ \lambda(-x_1 - x_2 + 1) = 0 \\ g(x_1, x_2) \geq 0 \end{cases}$$

$$1) \begin{cases} 4x^{-\frac{1}{3}} y^{\frac{1}{2}} + 4\lambda = 0 \\ 3x^{\frac{2}{3}} y^{-\frac{1}{2}} + 3\lambda = 0 \\ \lambda = 0 \\ g(x_1, x_2) \geq 0 \end{cases} \quad \begin{cases} 4x^{-\frac{1}{3}} y^{\frac{1}{2}} = 0 \\ 3x^{\frac{2}{3}} y^{-\frac{1}{2}} = 0 \\ \lambda = 0 \\ g(x_1, x_2) \geq 0 \end{cases} \quad \Rightarrow$$

$$1.1) \begin{cases} x = 0 \\ y \in \mathcal{R} \\ \lambda = 0 \\ g(x_1, x_2) = -7000 \leq 0 \\ f(x_1, x_2) = 0 \end{cases}$$

$$1.2) \begin{cases} y = 0 \\ x \in \mathcal{R} \\ \lambda = 0 \\ g(x_1, x_2) = -7000 \leq 0 \\ f(x_1, x_2) = 0 \end{cases}$$

$$2) \begin{cases} 4x^{-\frac{1}{3}}y^{\frac{1}{2}} + 4\lambda = 0 \\ 3x^{\frac{2}{3}}y^{-\frac{1}{2}} + 3\lambda = 0 \\ \lambda > 0 \\ g(x_1, x_2) = 0 \end{cases} \quad \begin{cases} 4x^{-\frac{1}{3}}y^{\frac{1}{2}} = -4\lambda \\ 3x^{\frac{2}{3}}y^{-\frac{1}{2}} = -3\lambda \\ \lambda > 0 \\ g(x_1, x_2) = 0 \end{cases} \quad \begin{cases} 4x^{-\frac{1}{3}}y^{\frac{1}{2}} = 3x^{\frac{2}{3}}y^{-\frac{1}{2}} \\ \lambda > 0 \\ 4x + 3y - 7000 = 0 \end{cases}$$

$$\begin{cases} x = y \\ \lambda > 0 \\ 7x = 7000 \end{cases} \quad \begin{cases} x = y = 1000 \\ f(x, y) = 6 \cdot 1000^{(\frac{2}{3} + \frac{1}{2})} = 6 \cdot 1000^{\frac{7}{6}} = 6 \cdot 10^{\frac{21}{6}} \end{cases}$$

Therefore, $x_1* = 1000, x_2* = 1000$ maximize f.

$$\begin{cases} x_1* = 1000 \\ x_2* = 1000 \\ f(x_1*, x_2*) = 6 * 10^{\frac{21}{6}} \end{cases}$$

2 Lagrange Multipliers: Warm-up

1.

$$L(\mathbf{a}, R, \xi, \alpha, \mu) = R^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i\left(-\|\mathbf{x}_i - \mathbf{a}\|^2 + R^2 + \xi_i\right) - \sum_{i=1}^{N}\mu_i\xi_i$$

2.

$$\frac{\partial L}{\partial R^2} = 1 - \sum_{i=1}^{N}\alpha_i = 0 \Rightarrow \sum_{i=1}^{N}\alpha_i = 1$$

$$\frac{\partial L}{\partial a} = -\sum_{i=1}^{N}2\alpha_i(\mathbf{x_i} - \mathbf{a}) = 2\sum_{i=1}^{N}\alpha_i(\mathbf{a} - \mathbf{x_i}) = 2(\sum_{i=1}^{N}\alpha_i\mathbf{a} - \sum_{i=1}^{N}\alpha_i\mathbf{x_i}) = 0$$

$$\sum_{i=1}^{N}\alpha_i\mathbf{a} = \sum_{i=1}^{N}\alpha_i\mathbf{x}_i \Rightarrow \mathbf{a} = \frac{\sum_{i=1}^{N}\alpha_i\mathbf{x}_i}{\sum_{i=1}^{N}} = \{because\sum_{i=1}^{N}\alpha_i = 1\} = \sum_{i=1}^{N}\alpha_i\mathbf{x}_i$$

$$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \mu_i \Rightarrow C = \alpha_i + \mu_i$$

Therefore KKT conditions are following:

$$\begin{cases} \sum_{i=1}^{N}\alpha_i = 1 \\ \mathbf{a} = \sum_{i=1}^{N}\alpha_i\mathbf{x}_i \\ C = \alpha_i + \mu_i \\ \forall i \in 1,..N : \xi_i \geq 0 \\ \forall i \in 1,..N : -\|\mathbf{x}_i - \mathbf{a}\|^2 + R^2 + \xi_i \geq 0 \\ \forall i \in 1,..N : \alpha_i \geq 0 \\ \forall i \in 1,..N : \alpha_i\left(-\|\mathbf{x}_i - \mathbf{a}\|^2 + R^2 + \xi_i\right) = 0 \\ \forall i \in 1,..N : \mu_i \geq 0 \\ \forall i \in 1,..N : \mu_i\xi_i = 0 \end{cases}$$

3.

1) If $\alpha_i > 0$, then $-\|\mathbf{x}_i - \mathbf{a}\|^2 + R^2 + \xi_i = 0 \Rightarrow \|\mathbf{x}_i - \mathbf{a}\|^2 = R^2 + \xi_i$, so points either lie on a circle or outside of it, depending on value of ξ_i.
2) If $\mu_i > 0$, then $\xi_i = 0 \Rightarrow \|\mathbf{x}_i - \mathbf{a}\|^2 \leq R^2 \Rightarrow \|\mathbf{x}_i - \mathbf{a}\| \leq R$ - points either lie on a circle or inside of it, depending on the value of $\mathbf{a}$.

4.

$$
\begin{aligned}
\tilde{L} &= R^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i\left(-\|\mathbf{x}_i - \mathbf{a}\|^2 + R^2 + \xi_i\right) - \sum_{i=1}^{N}\mu_i\xi_i \\
&= \{because\ \sum_{i=1}^{N}\alpha_i\} \\
&= \cancel{R^2} + C\sum_{i=1}^{N}\xi_i + \sum_{i=1}^{N}\alpha_i\|\mathbf{x}_i - \mathbf{a}\|^2 - \cancel{R^2\sum_{i=1}^{N}\alpha_i} - \sum_{i=1}^{N}\alpha_i\xi_i - \sum_{i=1}^{N}\mu_i\xi_i \\
&= \{because\ C = \alpha_i + \mu_i\} \\
&= \cancel{\sum_{i=1}^{N}\xi_i(\alpha_i + \mu_i)} + \sum_{i=1}^{N}\alpha_i\|\mathbf{x}_i - \mathbf{a}\|^2 - \cancel{\sum_{i=1}^{N}\alpha_i\xi_i} - \cancel{\sum_{i=1}^{N}\mu_i\xi_i} \\
&= \sum_{i=1}^{N}\alpha_i\|\mathbf{x}_i - \mathbf{a}\|^2 \\
&= \sum_{i=1}^{N}\alpha_i(\mathbf{x}_i - \sum_{j=1}^{N}\alpha_j\mathbf{x}_j)^T(\mathbf{x}_i - \sum_{j=1}^{N}\alpha_j\mathbf{x}_j) \\
&= \sum_{i=1}^{N}\alpha_i(\mathbf{x}_i^T - \sum_{j=1}^{N}\alpha_j\mathbf{x}_j^T)(\mathbf{x}_i - \sum_{j=1}^{N}\alpha_j\mathbf{x}_j) \\
&= \sum_{i=1}^{N}\alpha_i(\mathbf{x}_i^T\mathbf{x}_i) - (\mathbf{x}_i^T(\sum_{j=1}^{N}\alpha_j\mathbf{x}_j)) - ((\sum_{j=1}^{N}\alpha_j\mathbf{x}_j^T)\mathbf{x}_i) + (\sum_{j=1}^{N}\alpha_j\mathbf{x}_j^T)(\sum_{j=1}^{N}\alpha_j\mathbf{x}_j) \\
&= \{because\ (\sum_{j=1}^{N}\alpha_j\mathbf{x}_j^T)\mathbf{x}_i\ is\ a\ number,\ we\ can\ transpose\ it\} \\
&= \sum_{i=1}^{N}\alpha_i\left((\mathbf{x}_i^T\mathbf{x}_i) - 2(\mathbf{x}_i^T(\sum_{j=1}^{N}\alpha_j\mathbf{x}_j)) + (\sum_{j=1}^{N}\alpha_j\mathbf{x}_j^T)(\sum_{j=1}^{N}\alpha_j\mathbf{x}_j)\right) \\
&= \sum_{i=1}^{N}\alpha_i\mathbf{x}_i^T\mathbf{x}_i - 2\sum_{i=1}^{N}\alpha_i\sum_{j=1}^{N}\alpha_i\alpha_j\mathbf{x}_i^T\mathbf{x}_j + \sum_{i=1}^{N}\alpha_i\sum_{j=1}^{N}\alpha_j\sum_{m=1}^{N}\alpha_m\mathbf{x}_j^T\mathbf{x}_m \\
&= \{K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T\mathbf{x}_j\ -\ kernel\} \\
&= \sum_{i=1}^{N}\alpha_iK(\mathbf{x}_i, \mathbf{x}_i) - 2\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_jK(\mathbf{x}_i, \mathbf{x}_j) + \sum_{i=1}^{N}\alpha_i\sum_{j=1}^{N}\alpha_j\sum_{m=1}^{N}\alpha_mK(\mathbf{x}_j, \mathbf{x}_m) \\
&= \{because\ kernel\ in\ last\ term\ doesn't\ depend\ on\ \alpha_i\ and\ \sum_{i=1}^{N}\alpha_i = 1\} \\
&= \sum_{i=1}^{N}\alpha_iK(\mathbf{x}_i, \mathbf{x}_i) - 2\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_jK(\mathbf{x}_i, \mathbf{x}_j) + 1\cdot\sum_{j=1}^{N}\alpha_j\sum_{m=1}^{N}\alpha_mK(\mathbf{x}_j, \mathbf{x}_m) \\
&= \{because\ \sum_{j=1}^{N}\alpha_j\sum_{m=1}^{N}\alpha_mK(\mathbf{x}_j, \mathbf{x}_m) = \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_jK(\mathbf{x}_i, \mathbf{x}_j)\} \\
&= \sum_{i=1}^{N}\alpha_iK(\mathbf{x}_i, \mathbf{x}_i) - \sum_{j=1}^{N}\sum_{m=1}^{N}\alpha_j\alpha_mK(\mathbf{x}_j, \mathbf{x}_m)
\end{aligned}
$$

So, the dual kernelized problem is following:

$$\max_{\boldsymbol{\alpha},\boldsymbol{\mu}} \tilde{L}(\boldsymbol{\alpha},\boldsymbol{\mu}) = \sum_{i=1}^{N} \alpha_i K(\mathbf{x}_i,\mathbf{x}_i) - \sum_{j=1}^{N}\sum_{m=1}^{N} \alpha_j \alpha_m K(\mathbf{x}_j,\mathbf{x}_m),$$
$$s.t. \forall i : \alpha_i \geq 0, \mu_i \geq 0$$
$$where\ kernel\ is\ defined\ as\ following: \ K(\mathbf{x}_i,\mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$$

5.

$\mu_i = C - \alpha_i$

$a = \sum_{i=1}^{N} \alpha_i \mathbf{x}_i$

1)Considering optimal α_i for which $0 < \alpha_i < C$:

$$\begin{cases} \xi_i = 0 \\ R^2 = \|\mathbf{x}_i - \mathbf{a}\|^2 \end{cases}$$

$$\begin{cases} \xi_i = 0 \\ R = \|\mathbf{x}_i - \mathbf{a}\| = \|\mathbf{x}_i - \sum_{j=1}^{N} \alpha_j \mathbf{x}_j\| \end{cases}$$

so R is a distance from point on circle ($\xi_i = 0$) to the center $\mathbf{a}$.

2)For general case :

$$\begin{cases} \xi_i > 0 \\ \xi_i = \|\mathbf{x}_i - \mathbf{a}\|^2 - R^2 = \|\mathbf{x}_i - \sum_{j=1}^{N} \alpha_j \mathbf{x}_j\|^2 - \|\mathbf{x}_k - \sum_{m=1}^{N} \alpha_m \mathbf{x}_m\|^2 \end{cases}$$

so ξ_i would be distance to circle border (x_k lies on the circle border) for x_i outside circle

6.

Outliers test: $\|\mathbf{z} - \mathbf{a}\|^2 - R^2 = \|\mathbf{z} - \sum_{j=1}^{N} \alpha_j \mathbf{x}_j\|^2 - \|\mathbf{x}_k - \sum_{m=1}^{N} \alpha_m \mathbf{x}_m\|^2 > 0$, where $\mathbf{x}_k$ lies on the border of circle.

$$\begin{aligned}
\|\mathbf{z} - \sum_{j=1}^{N} \alpha_j \mathbf{x}_j\|^2 - \|\mathbf{x}_k - \sum_{m=1}^{N} \alpha_m \mathbf{x}_m\|^2 &= (\mathbf{z} - \sum_{j=1}^{N} \alpha_j \mathbf{x}_j)^T (\mathbf{z} - \sum_{j=1}^{N} \alpha_j \mathbf{x}_j) - (\mathbf{x}_k - \sum_{m=1}^{N} \alpha_m \mathbf{x}_m)^T (\mathbf{x}_k - \sum_{m=1}^{N} \alpha_m \mathbf{x}_m) \\
&= (\mathbf{z} - \sum_{j=1}^{N} \alpha_j \mathbf{x}_j)^T (\mathbf{z} - \sum_{j=1}^{N} \alpha_j \mathbf{x}_j) - (\mathbf{x}_m - \sum_{k=1}^{N} \alpha_k \mathbf{x}_k)^T (\mathbf{x}_m - \sum_{k=1}^{N} \alpha_j \mathbf{x}_k) \\
&= \left(\mathbf{z}^T\mathbf{z} - \mathbf{z}^T \sum_{j=1}^{N} \alpha_j \mathbf{x}_j - (\sum_{j=1}^{N} \alpha_j \mathbf{x}_j^T)\mathbf{z} + (\sum_{j=1}^{N} \alpha_j \mathbf{x}_j^T)(\sum_{l=1}^{N} \alpha_l \mathbf{x}_l) \right) \\
&- \left(\mathbf{x}_k^T\mathbf{x}_k - \mathbf{x}_m^T \sum_{m=1}^{N} \alpha_m \mathbf{x}_m - (\sum_{m=1}^{N} \alpha_m \mathbf{x}_m^T)\mathbf{x}_k + (\sum_{m=1}^{N} \alpha_m \mathbf{x}_m^T)(\sum_{n=1}^{N} \alpha_n \mathbf{x}_n) \right) \\
&= \{because\ (\sum_{j=1}^{N} \alpha_j \mathbf{x}_j^T)\mathbf{z}\ is\ a\ number,\ we\ can\ transpose\ it\} \\
&= \left(K(\mathbf{z},\mathbf{z}) - 2\sum_{j=1}^{N} \alpha_j K(\mathbf{z},\mathbf{x}_j) - \cancel{\sum_{j=1}^{N} \alpha_j \sum_{l=1}^{N} \alpha_l K(\mathbf{x}_j,\mathbf{x}_l)} \right) \\
&- \left(K(\mathbf{x}_k,\mathbf{x}_k) - 2\sum_{m=1}^{N} \alpha_m K(\mathbf{x}_k,\mathbf{x}_m) - \cancel{\sum_{m=1}^{N} \alpha_m \sum_{n=1}^{N} \alpha_n K(\mathbf{x}_m,\mathbf{x}_n)} \right) \\
&= K(\mathbf{z},\mathbf{z}) - 2\sum_{j=1}^{N} \alpha_j K(\mathbf{z},\mathbf{x}_j) - K(\mathbf{x}_k,\mathbf{x}_k) + 2\sum_{m=1}^{N} \alpha_m K(\mathbf{x}_k,\mathbf{x}_m) \\
&= K(\mathbf{z},\mathbf{z}) - K(\mathbf{x}_k,\mathbf{x}_k) - 2\sum_{j=1}^{N} \alpha_j (K(\mathbf{z},\mathbf{x}_j) - K(\mathbf{x}_k,\mathbf{x}_j)) > 0
\end{aligned}$$

7.

1) $C = 0$ - if $C = 0$, the only parameters to optimize are R^2 and $\mathbf{a}$ and because $C = 0$, there is no penalty for the outliers, so we can make ξ_i as big as possible, which would allow to make radius as small as possible and it will lead to R becoming zero, which would mean, that $\mathbf{a} = \sum_{i=1}^{N} \alpha_i x_i = \mathbf{0}$, because $\|\mathbf{x}_i - \mathbf{a}\|^2 - R^2 + \xi_i > 0$, so all the data points would then be outliers.
2) $C = \infty$ - if $C = \infty$, 'price' for being an outlier is so high, that there will be no outliers and radius will be covering nearly all the points.

8.

RBF kernel looks as following: $K(x_i, x_j) = \exp\left(\frac{\|\mathbf{x}_i - \mathbf{x}_j\|}{2\sigma^2}\right)$, so if $\sigma \to 0$, similarity regions between two points will be extremely small, so in such case only the circle itself or points, which lie in the same place as this exact point will be classified as inliers and all the other points will be classified as outliers. So, only inliers from the training set will be classified as inliers, while all the other datapoints from the training set and the new datapoints, which are different from the training ones, will be classified as outliers, which would lead to overfitting, because all the training data would be classified correct, while the new data will mostly be classified wrong (assuming there are more inliers in the new data than outliers).

Geometrically, it means, that for linear kernel, its value is changing linearly with the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$, while for RBF kernel, as we can see from the equation of kernel function above, value of σ behaves as an amplifier for the distance between x and x'. Therefore if σ is much smaller than the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$, the kernel function will become 0 and the 'radius' of the circle around point $\mathbf{x}_i$ will therefore be 0.

9.

Primal program will look as following:

$$\min_{\mathbf{a}, R, \xi} R^2 + C \sum_{i=1}^{N} \xi_i$$
$$s.t.\ \forall i : -y_i(\|\mathbf{x}_i - \mathbf{a}\| - R^2) \leq \xi_i, \xi_i \geq 0$$

To explain that, we first should note that we have introduced variable y_i, which is 1 for outliers and -1 for inliers. Thus, if point is an outlier, the value of $\|\mathbf{x}_i - \mathbf{a}\| - R^2$ will be positive and the value of $-y_i(\|\mathbf{x}_i - \mathbf{a}\| - R^2)$ will then be negative for the correctly classified data points. In contrast to that, for an inlier, the value of $\|\mathbf{x}_i - \mathbf{a}\| - R^2$ will be negative and in that case, the value of $-y_i(\|\mathbf{x}_i - \mathbf{a}\| - R^2)$ will be negative for the correctly classified data points.
To get more insight into the meaning of ξ_i, we need to understand, that it is only important when datapoint is wrongly classified and in that case, that would be a cost for wrong classification, while for correctly classified datapoints, the value of the right part is negative, so ξ_i could be set to zero without making constraint not valid.

3 Neural Network

1.

Note: here we are using binary cross entropy as a loss function.

$$\begin{aligned}
z_1 &= w_1 x_1 + w_2 x_2 = 0.4 \cdot 0.1 + 0.65 \cdot 0.35 = 0.2675 \\
z_2 &= w_3 x_1 + w_4 x_2 = 0.2 \cdot 0.1 + 0.1 \cdot 0.35 = 0.055 \\
a_1 &= \sigma(z_1) = \sigma(0.2675) = 0.566479056 \\
a_2 &= \sigma(z_2) = \sigma(0.055) = 0.5137465349 \\
y &= a_1 w_5 + a_2 w_6 = 0.8 \cdot 0.566479056 + 0.15 \cdot 0.5137465349 = 0.530245225 \\
a_y &= \sigma(y) = \sigma(0.530245225) = 0.629540305
\end{aligned}$$

$$\begin{aligned}
E(\mathbf{w}) &= -t \log a_y - (1 - t) \log (1 - a_y) \\
&= (-1 \cdot \log 0.629540305 - 0 \cdot \log (1 - 0.629540305)) \\
&= -\log 0.629540305 \\
&= 0.4627654006
\end{aligned}$$

2.

Note: here a means input into the layer, which we apply activation function to.

$$\delta^y = \frac{\partial E(\mathbf{w})}{\partial y} = \frac{\partial E(\mathbf{w})}{\partial a_y}\frac{\partial a_y}{\partial y} = -\frac{a_y - t}{a_y(1-a_y)}a_y(1-a_y) = a_y - t$$
$$= 0.4627654006 - 1 = -0.370459695$$

$$\begin{aligned}\frac{\partial E(\mathbf{w})}{\partial w_5} &= \frac{\partial E(\mathbf{w})}{\partial y}\frac{\partial y}{\partial w_5} = \delta^y a_1 \\ &= -0.370459695 \cdot 0.566479056 \\ &= -0.209857658 \\ \frac{\partial E(\mathbf{w})}{\partial w_6} &= \frac{\partial E(\mathbf{w})}{\partial y}\frac{\partial y}{\partial w_6} = \delta^y a_2 \\ &= -0.370459695 \cdot 0.5137465349 \\ &= -0.1903223846\end{aligned}$$

$$w_5* = w_5 - \eta\frac{\partial E}{\partial w_5} = 0.8 - 0.05 \cdot -0.209857658 = 0.8104928829$$
$$w_6* = w_6 - \eta\frac{\partial E}{\partial w_6} = 0.15 - 0.05 \cdot -0.1903223846 = 0.1595161192$$

3.

$$\begin{aligned}\delta^{z_1} &= \frac{\partial E(\mathbf{w})}{\partial z_1} = \frac{\partial E(\mathbf{w})}{\partial y}\frac{\partial y}{\partial a_1}\frac{\partial a_1}{\partial z_1} = \delta^y w_5 a_1(1-a_1) \\ &= -0.370459695 \cdot 0.8 \cdot 0.566479056 \cdot (1 - 0.566479056) = -0.0727821521 \\ \delta^{z_2} &= \frac{\partial E(\mathbf{w})}{\partial z_2} = \frac{\partial E(\mathbf{w})}{\partial y}\frac{\partial y}{\partial a_2}\frac{\partial a_2}{\partial z_2} = \delta^y w_6 a_2(1-a_2) \\ &= -0.370459695 \cdot 0.15 \cdot 0.5137465349 \cdot (1 - 0.5137465349) = -0.0138817379\end{aligned}$$

$$\begin{aligned}\frac{\partial E(\mathbf{w})}{\partial w_1} &= \frac{\partial E(\mathbf{w})}{\partial z_1}\frac{\partial z_1}{\partial w_1} = \delta^{z_1} x_1 \\ &= -0.0727821521 \cdot 0.1 = -0.0072782152 \\ \frac{\partial E(\mathbf{w})}{\partial w_2} &= \frac{\partial E(\mathbf{w})}{\partial z_1}\frac{\partial z_1}{\partial w_2} = \delta^{z_1} x_2 \\ &= -0.0727821521 \cdot 0.35 = -0.0254737532 \\ \frac{\partial E(\mathbf{w})}{\partial w_3} &= \frac{\partial E(\mathbf{w})}{\partial z_2}\frac{\partial z_2}{\partial w_3} = \delta^{z_2} x_1 \\ &= -0.0138817379 \cdot 0.1 = -0.00138817379 \\ \frac{\partial E(\mathbf{w})}{\partial w_4} &= \frac{\partial E(\mathbf{w})}{\partial z_1}\frac{\partial z_2}{\partial w_4} = \delta^{z_2} x_2 \\ &= -0.0138817379 \cdot 0.35 = -0.00485860826\end{aligned}$$

$$w_1* = w_1 - \eta\frac{\partial E}{\partial w_1} = 0.4 - 0.05 \cdot -0.0072782152 = 0.400363911$$
$$w_2* = w_2 - \eta\frac{\partial E}{\partial w_2} = 0.65 - 0.05 \cdot -0.0254737532 = 0.651273688$$
$$w_3* = w_3 - \eta\frac{\partial E}{\partial w_3} = 0.2 - 0.05 \cdot -0.00138817379 = 0.200069409$$
$$w_4* = w_4 - \eta\frac{\partial E}{\partial w_4} = 0.1 - 0.05 \cdot -0.00485860826 = 0.10024293$$

$$z_1 = w_1 * x_1 + w_2 * x_2 = 0.400363911 \cdot 0.1 + 0.651273688 \cdot 0.35 = 0.267982182$$
$$z_2 = w_3 * x_1 + w_4 * x_2 = 0.200069409 \cdot 0.1 + 0.10024293 \cdot 0.35 = 0.0550919664$$

$$
\begin{aligned}
a_1 &= \sigma(z_1) = \sigma(0.267982182) = 0.5665974667 \\
a_2 &= \sigma(z_2) = \sigma(0.0550919664) = 0.51376950909 \\
y &= a_1 w_5 * + a_2 w_6 * = 0.8104928829 \cdot 0.5665974667 + 0.1595161192 \cdot 0.51376950909 = 0.541177732 \\
a_y &= \sigma(0.541177732) = 0.6320863457
\end{aligned}
$$

$$
\begin{aligned}
E(\mathbf{w}) &= -t \log a_y - (1-t) \log (1 - a_y) \\
&= (-1 \cdot \log 0.6320863457 - 0 \cdot \log (1 - 0.6320863457)) \\
&= -\log 0.629540305 \\
&= 0.458729271
\end{aligned}
$$