Homework 4

JD Gallego Posada, *University of Amsterdam* 12/10/2016

1 Lagrange Multipliers

1. Find the maximum of $f = 1 - x_1^2 - 2x_2^2$, subject to the constraint that $x_1 + x_2 = 1$.

$$\mathscr{L}(\mathbf{x}, \lambda) = 1 - x_1^2 - 2x_2^2 + \lambda (x_1 + x_2 - 1)$$

$$\begin{aligned} \frac{\partial \mathscr{L}}{\partial x_1} &= -2x_1 + \lambda = 0 \\ \frac{\partial \mathscr{L}}{\partial x_2} &= -4x_2 + \lambda = 0 \\ \frac{\partial \mathscr{L}}{\partial \lambda} &= x_1 + x_2 - 1 = 0 \end{aligned} \quad \rightarrow \begin{bmatrix} -2 & 0 & 1 \\ 0 & -4 & 1 \\ 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \lambda \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \quad \rightarrow \begin{bmatrix} x_1^* \\ x_2^* \\ \lambda^* \end{bmatrix} = \begin{bmatrix} \frac{2}{3} \\ \frac{1}{3} \\ \frac{4}{3} \end{bmatrix} \quad \rightarrow f(\mathbf{x}^*) = \frac{1}{3}$$

Note that f is clearly concave. Thus $\mathbf{x}^*$ is indeed a maximum.

2. Find the maximum of $f = 1 - x_1^2 - x_2^2$ subject to the constraint $x_1 + x_2 - 1 \geqslant 0$

$$\mathscr{L}(\mathbf{x}, \lambda) = 1 - x_1^2 - x_2^2 + \lambda (x_1 + x_2 - 1)$$

$$\begin{aligned} \frac{\partial \mathscr{L}}{\partial x_1} &= -2x_1 + \lambda = 0 \\ \frac{\partial \mathscr{L}}{\partial x_2} &= -2x_2 + \lambda = 0 \\ \frac{\partial \mathscr{L}}{\partial \lambda} &= x_1 + x_2 - 1 = 0 \end{aligned} \quad \rightarrow \begin{bmatrix} -2 & 0 & 1 \\ 0 & -2 & 1 \\ 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \lambda \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \quad \rightarrow \begin{bmatrix} x_1^* \\ x_2^* \\ \lambda^* \end{bmatrix} = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \\ 1 \end{bmatrix} \quad \rightarrow f(\mathbf{x}^*) = \frac{1}{2}$$

Note that f is clearly concave and $\lambda^* > 0$. Thus $\mathbf{x}^*$ is indeed a maximum.

3. Find the maximum of $f = 1 - x_1^2 - x_2^2$ subject to the constraint $-x_1 - x_2 + 1 \geqslant 0$

$$\mathscr{L}(\mathbf{x}, \lambda) = 1 - x_1^2 - x_2^2 + \lambda (-x_1 - x_2 + 1)$$

$$\begin{aligned} \frac{\partial \mathscr{L}}{\partial x_1} &= -2x_1 - \lambda = 0 \\ \frac{\partial \mathscr{L}}{\partial x_2} &= -2x_2 - \lambda = 0 \\ \frac{\partial \mathscr{L}}{\partial \lambda} &= -x_1 - x_2 + 1 = 0 \end{aligned} \quad \rightarrow \begin{bmatrix} -2 & 0 & -1 \\ 0 & -2 & -1 \\ 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \lambda \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \quad \rightarrow \begin{bmatrix} x_1 \\ x_2 \\ \lambda \end{bmatrix} = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \\ -1 \end{bmatrix}$$

Since $\lambda < 0$, set $\lambda^* = 0$ and solve. Note that f is concave. Clearly, $x_1^* = x_2^* = 0$ and $f(\mathbf{x}^*) = 1$.

4. Find the maximum of $x_1 + 2x_2 - 2x_3$, subject to the constraint that $x_1^2 + x_2^2 + x_3^2 = 1$.

$$\mathscr{L}(\mathbf{x}, \lambda) = x_1 + 2x_2 - 2x_3 + \lambda \left(x_1^2 + x_2^2 + x_3^3 - 1\right)$$

$$\begin{aligned} \frac{\partial \mathscr{L}}{\partial x_1} &= 1 + 2x_1\lambda = 0 \\ \frac{\partial \mathscr{L}}{\partial x_2} &= 2 + 2x_2\lambda = 0 \\ \frac{\partial \mathscr{L}}{\partial x_3} &= -2 + 2x_3\lambda = 0 \\ \frac{\partial \mathscr{L}}{\partial \lambda} &= x_1^2 + x_2^2 + x_3^3 - 1 = 0 \end{aligned} \quad \rightarrow \quad \begin{aligned} x_1 &= \frac{-1}{2\lambda} \\ x_2 &= \frac{-1}{\lambda} \\ x_3 &= \frac{1}{\lambda} \end{aligned}$$

$$x_1^2 + x_2^2 + x_3^3 = 1 \rightarrow \frac{1}{4\lambda^2} + \frac{1}{\lambda^2} + \frac{1}{\lambda^2} = 1 \rightarrow \frac{1}{\lambda^2}\left(\frac{1}{4} + 2\right) = 1 \rightarrow \lambda^2 = \frac{9}{4} \rightarrow \lambda = \pm\frac{3}{2}$$

For $\lambda^* = -3/2$ we obtain $f(\mathbf{x}^*) = 3$. For $\lambda^* = 3/2$ we obtain $f(\mathbf{x}^*) = -3$. Thus the maximum is achieved at $\lambda^* = -3/2 \rightarrow x_1^* = \frac{1}{3}$, $x_2^* = \frac{2}{3}$, $x_3^* = -\frac{2}{3}$.

5. Find out the maximum number of doses that can be made if no more than 7000 euro can be spent on the ingredients.

We want to find the maximum of $6x^{2/3}y^{1/2}$ subject to $0 \leq 7000 - 4x - 3y$. Besides, we have physical constraints $x \geq 0$ and $y \geq 0$.

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

$$\begin{aligned} \frac{\partial \mathscr{L}}{\partial x} &= 6\frac{2}{3}x^{-1/3}y^{1/2} - 4\lambda = 0 & \lambda &= x^{-1/3}y^{1/2} \\ \frac{\partial \mathscr{L}}{\partial y} &= 6\frac{1}{2}x^{2/3}y^{-1/2} - 3\lambda = 0 \quad \rightarrow & \lambda &= x^{2/3}y^{-1/2} \quad \rightarrow \begin{bmatrix} x^* \\ y^* \\ \lambda^* \end{bmatrix} = \begin{bmatrix} 1000 \\ 1000 \\ \sqrt{10} \end{bmatrix} \\ \frac{\partial \mathscr{L}}{\partial \lambda} &= 7000 - 4x - 3y = 0 & 7000 &= 4x + 3y \end{aligned}$$

The Hessian matrix is given by:

$$H = \begin{bmatrix} -\frac{4}{3}x^{-4/3}y^{1/2} & 2x^{-1/3}y^{-1/2} \\ 2x^{-1/3}y^{-1/2} & -\frac{3}{2}x^{2/3}y^{-3/2} \end{bmatrix} \rightarrow |H| = -2x^{-2/3}y^{-1}$$

Note that H is negative definite at (x^*, y^*, λ^*). Therefore this point is a maximum and the maximum number of doses that can be produced with at most €7.000 is $f(\mathbf{x}^*) = 6000\sqrt{10}$.

2 Kernel Outlier Detection

1. Introduce Lagrange multipliers for the constraints and write down the primal Lagrangian.

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

subject to the original constraints and $\alpha_i \geq 0$, $\mu_i \geq 0 \quad \forall i = 1, \ldots, N$.

2. Write down all KKT conditions.

KKT Conditions

$$\begin{aligned} \alpha_i &\geq 0 & \mu_i &\geq 0 \\ R^2 + \xi_i - \|\mathbf{x}_i - \mathbf{a}\|^2 &\geq 0 & \xi_i &\geq 0 \\ \alpha_i \left(R^2 + \xi_i - \|\mathbf{x}_i - \mathbf{a}\|^2\right) &= 0 & \mu_i \xi_i &= 0 \end{aligned}$$

for all $i = 1, \ldots, N$.

FOC Conditions

$$\begin{aligned} \nabla_{R^2}\mathscr{L} &= 1 - \sum_i \alpha_i = 0 \rightarrow \sum_i \alpha_i = 1 \\ \nabla_{\xi_i}\mathscr{L} &= C - \alpha_i - \mu_i = 0 \rightarrow C = \alpha_i + \mu_i \quad \forall i = 1, \ldots, N \\ \nabla_{\boldsymbol{\alpha}}\mathscr{L} &= \sum_i \alpha_i(\mathbf{a} - \mathbf{x}_i) = 0 \rightarrow \mathbf{a} = \sum_i \alpha_i \mathbf{x}_i \end{aligned}$$

3. Identify the complementary slackness conditions. Use these conditions to derive what data-cases will have $\alpha_i > 0$ (support vectors) and which ones will have $\mu_i > 0$.
The complementary slackness conditions are given by:

$$\alpha_i \left(R^2 + \xi_i - \|\mathbf{x}_i - \mathbf{a}\|^2\right) = 0 \qquad \mu_i \xi_i = 0 \quad \forall i = 1, \ldots, N$$

α ***conditions***	μ ***conditions***	***Consequence***
$\alpha_i > 0$	$\mu_i = 0$	$\begin{cases} \xi_i = 0 \rightarrow \text{Lies on the boundary} \\ \xi_i > 0 \rightarrow \text{Outside the region} \end{cases}$
$\alpha_i > 0$	$\mu_i > 0$	$\xi_i = 0 \rightarrow$ Lies on the boundary
$\alpha_i = 0$	$\mu_i = 0$	Impossible assuming $C > 0$
$\alpha_i = 0$	$\mu_i > 0$	$\xi_i = 0 \rightarrow$ Inside the region

4. Derive the dual Lagrangian and specify the dual optimization problem. Kernelize the problem.

$$\begin{aligned}
\mathscr{L}(\mathbf{a}, R^2, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\mu}) &= R^2 + c\sum_i \xi_i - \sum_i \alpha_i \left(R^2 + \xi_i - \|\mathbf{x}_i - \mathbf{a}\|^2\right) - \sum_i \mu_i \xi_i \\
&= \cancel{R^2} + c\cancel{\sum_i \xi_i} - \cancel{\sum_i \alpha_i R^2} - \cancel{\sum_i \alpha_i \xi_i} + \sum_i \alpha_i \|\mathbf{x}_i - \mathbf{a}\|^2 - \cancel{\sum_i \mu_i \xi_i} \\
&= \sum_i \alpha_i (\mathbf{x}_i - \mathbf{a})^T (\mathbf{x}_i - \mathbf{a}) \\
&= \sum_i \alpha_i (\mathbf{x}_i - \sum_j \alpha_j \mathbf{x}_j)^T (\mathbf{x}_i - \sum_j \alpha_j \mathbf{x}_j) \\
&= \sum_i \alpha_i (\mathbf{x}_i^T - \sum_j \alpha_j \mathbf{x}_j^T)(\mathbf{x}_i - \sum_j \alpha_j \mathbf{x}_j) \\
&= \sum_i \alpha_i \left[\mathbf{x}_i^T \mathbf{x}_i - \sum_j \alpha_j \mathbf{x}_j^T \mathbf{x}_i - \mathbf{x}_i^T \sum_j \alpha_j \mathbf{x}_j + \left(\sum_j \alpha_j \mathbf{x}_j^T\right)\left(\sum_j \alpha_j \mathbf{x}_j\right)\right] \\
&= \sum_i \alpha_i \mathbf{x}_i^T \mathbf{x}_i - \sum_i \alpha_i \sum_j \alpha_j \mathbf{x}_j^T \mathbf{x}_i - \cancel{\sum_i \alpha_i \mathbf{x}_i^T \sum_j \alpha_j \mathbf{x}_j} + \cancel{\left(\sum_j \alpha_j \mathbf{x}_j^T\right)\left(\sum_j \alpha_j \mathbf{x}_j\right)} \\
&= \sum_i \alpha_i \mathbf{x}_i^T \mathbf{x}_i - \sum_i \sum_j \alpha_i \alpha_j \mathbf{x}_j^T \mathbf{x}_i \\
\tilde{\mathscr{L}}(\boldsymbol{\alpha}) &= \sum_i \alpha_i k(\mathbf{x}_i, \mathbf{x}_i) - \sum_i \sum_j \alpha_i \alpha_j k(\mathbf{x}_j, \mathbf{x}_i)
\end{aligned}$$

where $k(\cdot, \cdot)$ is the linear kernel.
The dual optimization problem is given by:

$$\begin{aligned}
&\max_{\boldsymbol{\alpha}} \quad \tilde{\mathscr{L}}(\boldsymbol{\alpha}) = \sum_i \alpha_i k(\mathbf{x}_i, \mathbf{x}_i) - \sum_i \sum_j \alpha_i \alpha_j k(\mathbf{x}_j, \mathbf{x}_i) \\
&\text{s.t} \qquad \sum_i \alpha_i = 1, \;\; 0 \leq \alpha_i \leq C \;\; \forall i = 1, \ldots, N.
\end{aligned}$$

5. The dual program will return optimal values for $\{\alpha_i\}$. Compute the optimal values for the other dual variables $\{\mu_i\}$. Then, solve the primal variables $\{\boldsymbol{a}, R, \boldsymbol{\xi}\}$ in terms of the dual variables $\{\mu_i, \alpha_i\}$.

Assuming we have $\boldsymbol{\alpha}^*$ we can compute $\boldsymbol{\mu}^*$ as $\mu_i^* = C - \alpha_i^*$. Furthermore, $\mathbf{a}^* = \sum_i \alpha_i^* \mathbf{x}_i$. To calculate R^{*2}, choose a support vector $\mathbf{x}_i$ for which $\alpha_i^* > 0$ and $\mu_i^* > 0$. Therefore, $\xi_i^* = 0$ and $R^{*2} = \|\mathbf{x}_i - \mathbf{a}^*\|^2$. Finally, $\xi_i^* = \max\{0, \|\mathbf{x}_i - \mathbf{a}^*\|^2 - R^{*2}\}$.

6. Describe a test in the dual space that could serve to detect outliers.
We want to build a function T which, given a new test case $\mathbf{z}$, allows us to classify the new point as an outlier if T is true and as not an outlier if T is false. Consider the test function T, given by:

$$T(\mathbf{z}) = \|\mathbf{z} - \mathbf{a}^*\|^2 - R^{*2} > 0$$

Note that T can be rewritten in terms of variables of the dual space as:

$$\begin{aligned} T(\mathbf{z}) &= k(\mathbf{z},\mathbf{z}) - 2k(\mathbf{z},\mathbf{a}^*) + k(\mathbf{a}^*,\mathbf{a}^*) - R^{*2} > 0 \\ &= k(\mathbf{z},\mathbf{z}) - 2k(\mathbf{z},\mathbf{a}^*) + \cancel{k(\mathbf{a}^*,\mathbf{a}^*)} - k(\mathbf{x}_0,\mathbf{x}_0) + 2k(\mathbf{x}_0,\mathbf{a}^*) - \cancel{k(\mathbf{a}^*,\mathbf{a}^*)} > 0 \\ &= k(\mathbf{z},\mathbf{z}) - k(\mathbf{x}_0,\mathbf{x}_0) - 2\sum_i \alpha_i \left(k(\mathbf{z},\mathbf{x}_i) - k(\mathbf{x}_0,\mathbf{x}_i)\right) > 0 \end{aligned}$$

where $\mathbf{x}_0$ is a support vector on the boundary.

7. What kind of solution do you expect if we use $C = 0$. And what solution if we use $C = \infty$?

Case $C \to 0$ Because of the constraint $\sum_i \alpha_i = 1$, for $C < 1/N$ no solution can be found, but clearly for $C > 1$ a solution always exits. Therefore the only choices for which C has an influence on the solution of the dual problem is when $1/N \leq C \leq 1$.

When C is restricted to small values, there is a low penalization to points outside the region and a larger fraction of points is allowed outside the region.

Case $C \to \infty$ This case gives an infinite penalization to points outside the region. Thus, we try to find the smallest region which is large enough to include all the data points. Which implies that all points in the data sets will be treated as non-outliers.

8. Describe geometrically what kind of solutions we may expect if we use a RBF kernel with very small bandwidth.

RBF Kernel

If we use a RBF kernel with very small bandwidth, then we have $k(\mathbf{w},\mathbf{z}) \approx 1$ if $\|\mathbf{w} - \mathbf{z}\| \approx 0$ and $k(\mathbf{w},\mathbf{z}) \approx 0$ if $\|\mathbf{w} - \mathbf{z}\| > 0$.

Recall the test function defined in question 6:

$$T(\mathbf{z}) \approx 1 - 2\sum_i \alpha_i^* k(\mathbf{z},\mathbf{x}_i) + \sum_i \alpha_i^{*2} - R^{*2} > 0$$

The only term depending on $\mathbf{z}$ which contributes negatively is $-2\sum_i \alpha_i^* k(\mathbf{z},\mathbf{x}_i)$. In order to accept point $\mathbf{z}$, the left hand side of the previous inequality must be negative. Therefore we have to make $-2\sum_i \alpha_i^* k(\mathbf{z},\mathbf{x}_i)$ as negative as possible. That occurs when $\mathbf{z}$ is very close to (at least) one of the data points. This will generate highly non-spherical and complex decision boundary, very sensitive to the location of the individual data points.

Linear Kernel

If we use a linear kernel, all the support vectors will contribute to the classification of a new test point (with a inversely proportional relevance determined by its dot product with every support vector). Since this kernel induces an Euclidean norm, even if the points are non-spherically distributed, the decision boundary will be an sphere centered in $\mathbf{a}$ with radius R.

N.B. Note that in either case there is always a dependency on the hyperparameter C, which controls the *size* of the region, while the kernel controls its *shape*.

9. Change the primal problem to include these labels and turn it into a classification problem similar to the SVM.

Label	***Situation***	***Condition***
$y_i = 1$	Outlier	$\|\mathbf{x}_i - \mathbf{a}\|^2 - R^2 > 0$
$y_i = -1$	Normal	$\|\mathbf{x}_i - \mathbf{a}\|^2 - R^2 \leq 0$

If we are given labels $y_i \in \{-1, 1\}$ we can modify the optimization problem to be:

$$\min_{\mathbf{a}, R, \boldsymbol{\xi}} R^2 + C \sum_i \xi_i$$

$$\text{s.t } y_i \left(\|\mathbf{x}_i - \mathbf{a}\|^2 - R^2 \right) + \xi_i \geq 0, \;\; \xi_i \geq 0 \;\; \forall i = 1, \ldots, N.$$