Machine Learning 1 - Homework 4

Selene Baez Santamaria

1 Lagrange Multipliers

In this exercise, we will do optimization problems using Lagrange Multipliers. Suppose we would like to maximize the function

$$f(\mathbf{x}) = 1 - x_1^2 - 2x_2^2 \tag{1}$$

Answer the following questions:

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

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

   Partial derivatives:

   $$\frac{\partial L}{\partial x_1} = -2x_1 + \lambda = 0 \tag{2}$$
   $$\frac{\partial L}{\partial x_2} = -4x_2 + \lambda = 0 \tag{3}$$
   $$\frac{\partial L}{\partial \lambda} = x_1 + x_2 - 1 = 0 \tag{4}$$

   From 2 and 3 we have:

   $$x_1 = 2x_2$$

   Substituting in 4:

   $$3x_2 = 1$$

   Therefore:

   $$x_1 = \frac{2}{3} \qquad x_2 = \frac{1}{3} \qquad \lambda = \frac{4}{3}$$

2. Find the maximum of $1 - x_1^2 - x_2^2$ subject to the constraint $x_1 + x_2 - 1 \geqslant 0$
Solution:
Lagrangian:

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

Partial derivatives:

$$\frac{\partial L}{\partial x_1} = -2x_1 + \lambda = 0 \tag{5}$$
$$\frac{\partial L}{\partial x_2} = -2x_2 + \lambda = 0 \tag{6}$$
$$\frac{\partial L}{\partial \lambda} = x_1 + x_2 - 1 \geqslant 0 \tag{7}$$

Additional constraints:

$$\lambda \geqslant 0 \tag{8}$$
$$\lambda(x_1 + x_2 - 1) = 0 \tag{9}$$

From 5 and 6 we have:

$$x_1 = x_2$$
$$x_1 = \frac{\lambda}{2}$$

Substituting in 9:

$$2x_2 = 1$$

Therefore:

$$x_1 = \frac{1}{2} \qquad x_2 = \frac{1}{2} \qquad \lambda = 1$$

Checking the constraints we note that 7 and 8 hold, hence the solution satisfies all the conditions.

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

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

Partial derivatives:

$$\frac{\partial L}{\partial x_1} = -2x_1 - \lambda = 0 \tag{10}$$
$$\frac{\partial L}{\partial x_2} = -2x_2 - \lambda = 0 \tag{11}$$
$$\frac{\partial L}{\partial \lambda} = -x_1 - x_2 + 1 \geqslant 0 \tag{12}$$

Additional constraints:

$$\lambda \geqslant 0 \tag{13}$$
$$\lambda(-x_1 - x_2 + 1) = 0 \tag{14}$$

From 10 and 11 we have:

$$x_1 = x_2$$
$$x_1 = -\frac{\lambda}{2}$$

Substituting in 12:

$$2x_2 = 1$$

Therefore:

$$x_1 = \frac{1}{2} \qquad x_2 = \frac{1}{2} \qquad \lambda = -1$$

Checking the constraints we note that 13 does not hold, hence we need to revisit the solution for 12. Choosing the first term this time:

$$\lambda = 0$$

Therefore:

$$x_1 = 0 \qquad x_2 = 0 \qquad \lambda = 0$$

One more check at the constraints shows that 12 and 13 hold, hence the solution satisfies all the conditions.

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

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

Partial derivatives:

$$\frac{\partial L}{\partial x_1} = 1 + 2\lambda x_1 = 0 \tag{15}$$
$$\frac{\partial L}{\partial x_2} = 2 + 2\lambda x_2 = 0 \tag{16}$$
$$\frac{\partial L}{\partial x_3} = -2 + 2\lambda x_3 = 0 \tag{17}$$
$$\frac{\partial L}{\partial \lambda} = x_1^2 + x_2^2 + x_3^2 - 1 = 0 \tag{18}$$

From 15, 16 and 17 we have:

$$\lambda = -\frac{1}{2x_1} = -\frac{1}{x_2} = \frac{1}{x_3}$$
$$\therefore$$
$$x_2 = 2x_1, x_3 = -2x_1$$

Substituting in 18:

$$x_1^2 + (2x_1)^2 + (-2x_1)^2 - 1 = 0$$
$$x_1^2 = \frac{1}{9}$$

Therefore:

$$x_1 = \frac{1}{3} \qquad x_2 = \frac{2}{3} \qquad x_3 = -\frac{2}{3} \qquad \lambda = -\frac{3}{2}$$

Or

$$x_1 = -\frac{1}{3} \qquad x_2 = -\frac{2}{3} \qquad x_3 = \frac{2}{3} \qquad \lambda = \frac{3}{2}$$

In order to select the appropriate set of values, we substitute in the original function and choose the maximum:

$$x_1 + 2x_2 - 2x_3 = \frac{1}{3} + 2(\frac{2}{3}) - 2(-\frac{2}{3}) = \frac{9}{3} = 3 \tag{19}$$
$$x_1 + 2x_2 - 2x_3 = -\frac{1}{3} + 2(-\frac{2}{3}) - 2(\frac{2}{3}) = -\frac{9}{3} = -3 \tag{20}$$

Since 19 has the maximum, we select the first set of values.

5. A company manufactures a chemical product out of two ingredients, known as ingredient X and ingredient Y. The number of doses produced, D, is given

by $6x^{2/3}y^{1/2}$, where x and y are the number of grams of ingredients X and Y respectively. Suppose ingredient X consts 4 euro per gram, and ingredient Y costs 3 euro per gram. Find out the maximum number of doses that can be made if no more than 7000 euro can be spent on the ingredients.
Solution:
Constraint:

$$4x + 3y \leqslant 7000$$

Lagrangian:

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

Partial derivatives:

$$\frac{\partial L}{\partial x} = 4x^{-1/3}y^{1/2} - 4\lambda = 0 \tag{21}$$
$$\frac{\partial L}{\partial y} = 3x^{2/3}y^{-1/2} - 3\lambda = 0 \tag{22}$$
$$\frac{\partial L}{\partial \lambda} = 7000 - 4x - 3y \geqslant 0 \tag{23}$$

Additional constraints:

$$\lambda \geqslant 0 \tag{24}$$

From 21 and 22 we have:

$$\lambda = x^{-1/3}y^{1/2} = x^{2/3}y^{-1/2}$$
$$\therefore$$
$$x = y$$

Substituting in 23:

$$7000 = 7x$$
$$x = 1000$$

Therefore:

$$x = 1000 \qquad y = 1000 \qquad \lambda = 1000^{1/6}$$

Checking the constraints we note that 24 holds, hence the solution satisfies all the conditions.

2 Kernel Outlier Detection

Our task is to derive an algorithm that will detect the outliers (in this example there are 2 of them). To that end, we draw a circle rooted at location $\boldsymbol{a}$ and with radius R. All data-cases that fall outside the circle are detected as outliers.

We will now write down the primal program that will find such a circle:

$$\min_{\boldsymbol{a},R,\boldsymbol{\xi}} R^2 + C\sum_{i=1}^{N}\xi_i$$
$$s.t.\ \forall i : \|\boldsymbol{x_i} - \boldsymbol{a}\|^2 \leq R^2 + \xi_i,\ \xi_i \geq 0$$

In words: we want to minimize the radius of the circle subject to the constraint that most data-cases should lay inside it. Outliers are allowed to stay outside but they pay a price proportional their distance from the circle boundary and C.

Answer the following questions:

1. Introduce Lagrange multipliers for the constraints and write down the primal Lagrangian. Use the following notation: $\{\alpha_i\}$ are the Lagrange multipliers for the first constraint and $\{\mu_i\}$ for the second constraint.
   Solution:
   Given that α_i and μ_i are Lagrange multipliers:

$$L = R^2 + C\sum_i^N \xi_i + \sum_i^N \alpha_i \left(||\mathbf{x}_i - \mathbf{a}||^2 - R^2 - \xi_i\right) - \sum_i^N \mu_i\xi_i$$

2. Write down all KKT conditions. (Hint: take the derivative w.r.t. R^2 instead of R).
   Solution:
   Partial derivatives:

$$\frac{\partial L}{\partial R^2} = 1 - \sum_i^N \alpha_i = 0 \tag{1}$$
$$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \mu_i = 0 \tag{2}$$
$$\frac{\partial L}{\partial \mathbf{a}} = -2\sum_i^N \alpha_i(\mathbf{x}_i - \mathbf{a}) = 0 \tag{3}$$

Additional constraints:

$$||\mathbf{x}_i - \mathbf{a}||^2 - R^2 - \xi_i \leq 0 \tag{4}$$
$$\xi_i \geq 0 \tag{5}$$
$$\alpha_i \geq 0 \tag{6}$$
$$\mu_i \geq 0 \tag{7}$$
$$\mu_i \cdot \xi_i = 0 \tag{8}$$
$$\alpha_i \cdot \left(||\mathbf{x}_i - \mathbf{a}||^2 - R^2 - \xi_i\right) = 0 \tag{9}$$

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 $\mu_i > 0$.
Solution:
7 and 9 are the complementary slackness conditions. Considering α_i

   - Data cases inside the circle:
     $||\mathbf{x}_i - \mathbf{a}||^2 - R^2 - \xi_i < 0$
     $\therefore \alpha_i = 0$
   - Data cases on or outside the circle:
     $||\mathbf{x}_i - \mathbf{a}||^2 - R^2 - \xi_i = 0$
     $\therefore \alpha_i \geqslant 0$

   Considering for μ_i

   - Data cases on or inside the circle:
     $\xi_i = 0 \therefore \mu_i \geqslant 0$
   - Data cases outside the circle:
     $\xi_i > 0 \therefore \mu_i = 0$

4. Derive the dual Lagrangian and specify the dual optimization problem. Kernelize the problem, i.e. write the dual program only in terms of kernel entries and Lagrange multipliers.
Solution:
From 1, 2, and 3, we have:

$$\sum_i^N \alpha_i = 1 \tag{10}$$
$$\mu_i = C - \alpha_i \tag{11}$$
$$\mathbf{a} = \sum_i^N \alpha_i \mathbf{x}_i \tag{12}$$

Hereby we rewrite the Lagrangian:

Original Lagranian:

$$L = R^2 + C\sum_i^N \xi_i + \sum_i^N \alpha_i \left(||\mathbf{x}_i - \mathbf{a}||^2 - R^2 - \xi_i\right) - \sum_i^N \mu_i \xi_i$$

Expand:

$$L = R^2 + C\sum_i^N \xi_i + \sum_i^N \alpha_i \left(||\mathbf{x}_i - \mathbf{a}||^2\right) - \sum_i^N \alpha_i R^2 - \sum_i^N \alpha_i \xi_i - \sum_i^N \mu_i \xi_i$$

Using 10:

$$L = R^2 + C\sum_i^N \xi_i + \sum_i^N \alpha_i \left(||\mathbf{x}_i - \mathbf{a}||^2\right) - R^2 - \sum_i^N \alpha_i \xi_i - \sum_i^N \mu_i \xi_i$$

Cancel R^2

$$L = C\sum_i^N \xi_i + \sum_i^N \alpha_i \left(||\mathbf{x}_i - \mathbf{a}||^2\right) - \sum_i^N \alpha_i \xi_i - \sum_i^N \mu_i \xi_i$$

Using 11:

$$L = C\sum_i^N \xi_i + \sum_i^N \alpha_i \left(||\mathbf{x}_i - \mathbf{a}||^2\right) - \sum_i^N \alpha_i \xi_i - \sum_i^N (C - \alpha_i) \xi_i$$

Expand:

$$L = C\sum_i^N \xi_i + \sum_i^N \alpha_i \left(||\mathbf{x}_i - \mathbf{a}||^2\right) - \sum_i^N \alpha_i \xi_i - \sum_i^N C\xi_i + \sum_i^N \alpha_i \xi_i$$

Cancel terms:

$$L = \sum_i^N \alpha_i ||\mathbf{x}_i - \mathbf{a}||^2$$

$$L = \sum_{i}^{N} \alpha_i \left(\mathbf{x}_i^T \mathbf{x} + \mathbf{a}^T \mathbf{a} - 2\mathbf{x}_i^T \mathbf{a} \right)$$

$$L = \sum_{i}^{N} \alpha_i \mathbf{x}_i^T \mathbf{x}_i + \sum_{i}^{N} \alpha_i \mathbf{a}^T \mathbf{a} - 2 \sum_{i}^{N} \alpha_i \mathbf{x}_i^T \mathbf{a}$$

$$L = \sum_{i}^{N} \alpha_i \mathbf{x}_i^T \mathbf{x}_i + \mathbf{a}^T \mathbf{a} - 2\mathbf{a}^T \mathbf{a}$$

$$L = \sum_{i}^{N} \alpha_i \mathbf{x}_i^T \mathbf{x}_i - \mathbf{a}^T \mathbf{a}$$

$$L = \sum_{i}^{N} \alpha_i K(\mathbf{x}_i, \mathbf{x}_i) - \sum_{i,j}^{N} \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j)$$

Hence the dual program becomes:

$$\max_{\alpha} \sum_{i}^{N} \alpha_i K(\mathbf{x}_i, \mathbf{x}_i) - \sum_{i,j}^{N} \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j)$$
$$s.t. 0 \leq a_i \leq C, \forall i$$

5. The dual program will return optimal values for $\{\alpha_i\}$. In terms of these, compute the optimal values for the other dual variables $\{\mu_i\}$.

   Then, solve the primal variables $\{\boldsymbol{a}, R, \boldsymbol{\xi}\}$ (in that order) in terms of the dual variables $\{\mu_i, \alpha_i\}$. Note that you do not need to know the dual optimization program to solve this question. You only need the KKT conditions.
   Solution:

   - Optimal values of α_i
     $\alpha_i^\star$ Given from dual program
   - Optimal values of μ_i
     From 11: $\mu_i^\star = C - \alpha_i^\star$
   - Optimal values of $\mathbf{a}$

From 3:

$$
\begin{aligned}
-\sum_{i}^{N} \alpha_i^\star (\mathbf{x}_i - \mathbf{a}) &= 0 \\
-\sum_{i}^{N} \alpha_i^\star \mathbf{x}_i + \sum_{i}^{N} \alpha_i^\star \mathbf{a} &= 0 \\
\mathbf{a} &= \frac{\sum_i^N \alpha_i^\star \mathbf{x}_i}{\sum_i^N \alpha_i^\star} \\
\mathbf{a} &= \sum_{i}^{N} \alpha_i^\star \mathbf{x}_i
\end{aligned}
$$

- Optimal values of R
  The optimal value for R is determined by $\mathbf{x}_i$.

$$
R = \begin{cases} ||\mathbf{x}_i - \mathbf{a}||^2 & \text{if } \mathbf{x}_i \text{ is on the circle} \\ \frac{1}{N_{ball}} \sum_i^N ||\mathbf{x}_i - \mathbf{a}^\star||^2 & \text{if } \mathbf{x}_i \text{ otherwise} \end{cases}
$$

- Optimal values of $\boldsymbol{\xi}$
  Similarly, the optimal value for ξ_i is determined by $\mathbf{x}_i$.

$$
\xi_i = \begin{cases} 0 & \text{if } \mathbf{x}_i \text{ is on or inside the circle} \\ ||\mathbf{x}_i - \mathbf{a}^\star||^2 - R^2 & \text{if } \mathbf{x}_i \text{ is outside the circle} \end{cases}
$$

6. Assume we have solved the dual program. We now want to apply it to new test cases. Describe a test in the dual space (i.e. in terms if kernels and Lagrange multipliers) that could serve to detect outliers. (Students who got stuck along the way may describe the test in primal space).
   Solution:

$$||\mathbf{x}^* - \mathbf{a}||^2 > R^2$$

Expand:

$$||\mathbf{x}^* - \sum_i^N \alpha_i \mathbf{x}_i||^2 > R^2$$

Norm identity:

$$(\mathbf{x}^* - \sum_i^N \alpha_i \mathbf{x}_i)^T (\mathbf{x}^* - \sum_i^N \alpha_i \mathbf{x}_i) > R$$

$$\mathbf{x}^{*T}\mathbf{x}^* + \sum_{i,j}^N \alpha_i \alpha_j \mathbf{x}_i^T \mathbf{x}_j - 2\mathbf{x}^{*T} \sum_i^N \alpha_i \mathbf{x}_i > R$$

$$\text{If } K(\mathbf{x}^*, \mathbf{x}^*) + \sum_{i,j}^N \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) - 2\sum_i^N \alpha_i K(\mathbf{x}^*, \mathbf{x}_i) > R^2$$

then the point is an outlier

7. What kind of solution do you expect if we use $C = 0$. And what solution if we use $C = \infty$?
   Solution:
   If $C \to 0$ the cost of a variable being outside of the circle is nothing. Therefore the values of ξ_i can become as large as necessary without any penalty and the minimization plainly minimizes R^2. Thus, R will approach zero if $C \to 0$.

   If $C \to \infty$ the cost of a variable being outside the circle is infinitely large. To minimize the function, R is made large enough to accommodate every data point. In other words, every data point lies in the circle. Because the cost of a data point being outside the bounds is so large, R is made large enough so that it encircles all data points.

8. Describe geometrically what kind of solutions we may expect if we use a RBF kernel (Gaussian) with very small bandwidth (sigma = small), i.e. describe how these solutions can be different geometrically (in x-space) from the case with a linear kernel.
   Solution:
   The output of an RBF kernel is high if two points are alike and low if two points are apart. An RBF kernel is therefore able to make clusters of data,

not limited to a single circle. It actually models data points on how close they are.

9. Now assume that you are given labels (e.g. y=1 for outlier and y=-1 for "inlier"). Change the primal problem to include these labels and turn it into a classification problem similar to the SVM. (You do not have to derive the dual program).
Solution:
Restating the original optimization problem.

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

Now it is assumed that there is some boundary at distance R that separates inliers from outliers. With $y = -1$ for inliers and $y = 1$ for outliers. The optimization problem is now defined as depicted below. Data points are assumed to be distanced 1 away from the separation boundary. The problem is now described as a two-sided SVM.

$$\text{Our objective is } \min_{\mathbf{a},R,\boldsymbol{\xi}} R^2 + C\sum_{i}^{N} \xi_i$$
$$y(||\mathbf{x}_i - \mathbf{a}||^2 - R^2) > 1 - \xi_i$$
$$s.t.\ \forall i : y(||\mathbf{x}_i - \mathbf{a}||^2 - R^2) - 1 + \xi_i \geq 0$$
$$for\ \ \xi_i \geq 0$$