Machine Learning 1 - Homework 4 - Solutions

1 Lagrange Multipliers: Warm-up

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

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

which has two input dimensions x_1 and x_2 (they can be considered as the parameters that we would like to learn). This function is plotted in Figure 1(a). We can see the function is concave and there is no local minimum. The optimization of the function is subject to a constraint function. Therefore the optimal solution that is found has to satisfies the constraints.

For example, if we set the constraints $x_1 + x_2 = 1$, then the optimization problem is to find the maximal value of $f(\mathbf{x})$ where $\mathbf{x}$ is also on the constraint plane. Figure 1(b) shows the constraint function (black) separates $f(\mathbf{x})$ into two parts. The 3D view of $f(\mathbf{x})$ is slighted changed in Figure 1(b) for better visualization.

Answer the following questions:

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

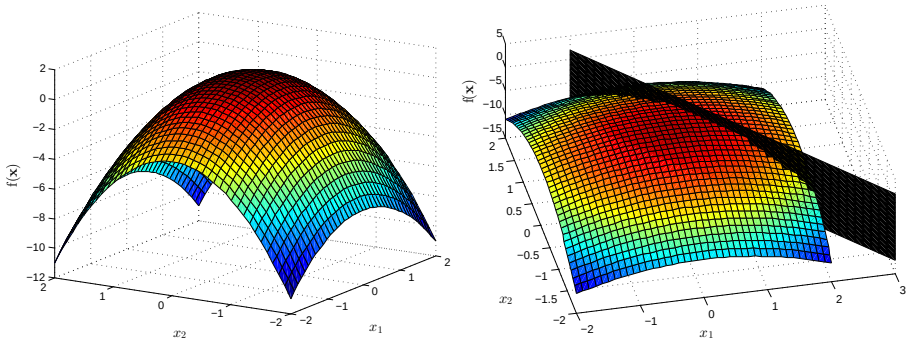
   Solution:
   The lagrangian is given by

   $$L = 1 - x_1^2 - 2x_2^2 + \lambda(x_1 + x_2 - 1). \tag{2}$$

   Taking the derivatives with respect to x and λ are then

   $$\begin{cases} -2x_1 + \lambda = 0 & \text{(A)} \\ -4x_2 + \lambda = 0 & \text{(B)} \\ x_1 + x_2 = 1 & \text{(C)} \end{cases} \tag{3}$$

   Subtracting (A) from (B) tells us that $-2x_1 + 4x_2 = 0$ so that $x_1 = 2x_2$. In (C): $3x_2 = 1$, so that $x_2 = \frac{1}{3}$ and $x_1 = \frac{2}{3}$.



(a) (b)

Figure 1: (a) $f(\mathbf{x}) = 1 - x_1^2 - 2x_2^2$ Plot of the example function. Plot (b) also illustrates the constraint surface.

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

   Solution:
   The lagrangian is given by $L = 1 - x_1^2 - x_2^2 + \lambda(x_1 + x_2 - 1)$. Since this is an inequality constraints, the resulting set of constraints to satisfy are:

$$\begin{cases} -2x_1 + \lambda = 0 & \text{(A)} \\ -2x_2 + \lambda = 0 & \text{(B)} \\ x_1 + x_2 - 1 \geqslant 0 & \text{(C)} \\ \lambda \geqslant 0 & \text{(D)} \\ \lambda(x_1 + x_2 - 1) = 0 & \text{(E)} \end{cases} \tag{4}$$

   From (A) and (B) we get that $x_1 = x_2$ which, in (E) results in $x_1 = x_2 = \frac{1}{2}$. Knowing this we get from (A) and (B) that $\lambda = 1$, so that (D) holds and the constraint is active.

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

Solution:
The lagrangian is given by Ł $= 1 - x_1^2 - x_2^2 + \lambda(-x_1 - x_2 + 1)$. The resulting set of constraints to satisfy are:

$$\begin{cases} -2x_1 - \lambda = 0 & \text{(A)} \\ -2x_2 - \lambda = 0 & \text{(B)} \\ -x_1 - x_2 + 1 \geqslant 0 & \text{(C)} \\ \lambda \geqslant 0 & \text{(D)} \\ \lambda(-x_1 - x_2 + 1) = 0 & \text{(E)} \end{cases} \tag{5}$$

Again from (A) and (B) we get that $x_1 = x_2$ which, in (E) results in $x_1 = x_2 = 1/2$ if $\lambda \neq 0$. Knowing this we get from (A) and (B) that $\lambda = -1$, so that (D) does not hold. This is therefore not the correct solution, so that λ must be zero and the constraint is inactive. Filling this value into (A) and (B) results in $x_1 = x_2 = 0$, which, when filled into (C) results in $1 \geqslant 0$, which is correct.

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: The Lagrangian is $L = x_1 + 2x_2 - 2x_3 + \lambda(x_1^2 + x_2^2 + x_3^2 - 1)$. The constraints are therefore

$$\begin{cases} 1 + 2\lambda x_1 = 0 & \text{(A)} \\ 2 + 2\lambda x_2 = 0 & \text{(B)} \\ -2 + 2\lambda x_3 = 0 & \text{(C)} \\ x_1^2 + x_2^2 + x_3^2 = 1 & \text{(D)} \end{cases} \tag{6}$$

   From (A) we get that $\lambda = -\frac{1}{2x_1}$. Using this in (B), we get that $x_2 = 2x_1$ and in (C) we get that $x_3 = -2x_1$. Filling those in (D) results in $9x_1^2 = 1$, so that the function is optimal (subject to the constraint) for either $x = (x_1, x_2, x_3) = (\frac{1}{3}, \frac{2}{3}, -\frac{2}{3})$ or $x = (-\frac{1}{3}, -\frac{2}{3}, \frac{2}{3})$. If we compute the corresponding values for $f(x)$, the result is 3 and -3 respectively, so that the maximum is obtained for $x = (\frac{1}{3}, \frac{2}{3}, -\frac{2}{3})^t p$. The other solution is the minimum.

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:
Our constraint is that $4x + 3y \leqslant 7000$, so that the Lagrangian is $L = 6x^{2/3}y^{1/2} + \lambda(7000 - 4x - 3y)$. From this, we get the set of equations:

$$\begin{cases} 4x^{-1/3}y^{1/2} - 4\lambda = 0 & \text{(A)} \\ 3x^{2/3}y^{-1/2} - 3\lambda = 0 & \text{(B)} \\ \lambda \geqslant 0 & \text{(C)} \\ 7000 - 4x - 3y \geqslant 0 & \text{(D)} \end{cases} \tag{7}$$

From (A), we get

$$\lambda = \frac{y^{1/2}}{x^{1/3}} \tag{8}$$

which, in (B) gives that $x = y$. In (D), this tells us that the larges possible positive value of $x = y = 1000$. To double-check that lambda is positive, we have

$$\lambda = \frac{\sqrt{1000}}{10} > 0 \tag{9}$$

2 Kernel Outlier Detection

Consider the picture in Figure 2. The dots represent data-items. 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.

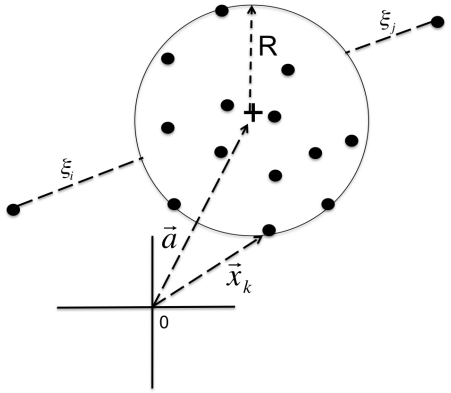


Figure 2: Kernel Outlier Detection

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

3. Identify the complementary slackness conditions. Use these conditions to derive what data-cases (e.g. in Figure 2) will have $\alpha_i > 0$ (support vectors) and which ones will have $\mu_i > 0$.

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.

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.

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).

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

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.

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).

$$\min_{\bar{a}, R, \bar{\xi}} \quad R^2 + C \sum_i \xi_i$$

$$\text{s.t.} \quad \|x_i - \bar{a}\|^2 \leq R^2 + \xi_i \quad \forall i$$

$$\text{s.t} \quad \xi_i \geq 0 \quad \forall i$$

1 $\mathcal{L}(a, R, \xi, \alpha, \mu) = R^2 + C \sum_i \xi_i$

$$+ \sum_i \alpha_i \left(\|x_i - \bar{a}\|^2 - R^2 - \xi_i \right) + - \sum_i \mu_i \xi_i$$

2 K.K.T :

$$\frac{\partial \mathcal{L}}{\partial R^2} = 0 \Rightarrow \boxed{1 = \sum_i \alpha_i} \quad (A)$$

$$\frac{\partial \mathcal{L}}{\partial a} = 0 \Rightarrow -2 \sum_i \alpha_i (x_i - a) = 0$$

$$\sum_i \alpha_i x_i = a \sum_i \alpha_i \Rightarrow (B) \boxed{a = \sum_i \alpha_i x_i}$$

$$\frac{\partial \mathcal{L}}{\partial \xi_i} = 0 \Rightarrow \boxed{C - \alpha_i - \mu_i = 0} \quad (C)$$

(D) $\|x_i - \bar{a}\|^2 - R^2 - \xi_i \leq 0 \quad \forall i$

(E) $\xi_i \geq 0 \quad \forall i$

(F) $\alpha_i \geq 0 \quad \forall i$

(G) $\mu_i \geq 0 \quad \forall i$

(H) $\alpha_i \left(\|x_i - \bar{a}\|^2 - R^2 - \xi_i \right) = 0 \quad \forall i$

(I) $\mu_i \xi_i = 0 \quad \forall i$

3. (G) & (I) are C.S. conditions

=> Data - cases on inside ball:

$\|x_i - \bar{a}\|^2 - R^2 - \xi_i \leq 0 \Rightarrow \alpha_i = 0$

Data - cases on or outside ball

$\|x_i - \bar{a}\|^2 - R^2 - \xi_i = 0 \Rightarrow \alpha_i \geq 0$

Data - cases: Inside or on inside ball

$\xi_i = 0 \Rightarrow \mu_i \geq 0$

Data - cases outside ball $\xi_i > 0 \Rightarrow \mu_i = 0$

Note that by (C) & (F) & (G)

$\alpha_i = 0 \Rightarrow \mu_i = C$

$\mu_i = 0 \Rightarrow \alpha_i = C$

4. $L = R^2 + C\sum_i \xi_i + \sum_i \alpha_i x_i^T x_i - 2a^T(\sum_i \alpha_i x_i) + a^T a \sum_i \alpha_i - R^2 \sum_i \alpha_i - \sum_i \alpha_i \xi_i + \sum_i (\alpha_i - C)\xi_i$

$+ \sum_i \alpha_i \xi_i - C\sum_i \xi_i$

$L(\alpha) = \sum_i \alpha_i x_i^T x_i - a^T a = \sum_i \alpha_i \|x_i\|^2 - \sum_{ij} \alpha_i \alpha_j x_i^T x_j$

$$\mathcal{L}(\alpha) = \sum_i \alpha_i K(x_i, x_i) - \sum_{ij} \alpha_i K(x_i, x_j) \alpha_j$$

Dual program

$$\max_{\{\alpha\}} \sum_i \alpha_i K_{ii} - \sum_{ij} \alpha_i \alpha_j K_{ij}$$

$$\text{s.t.} \quad \alpha_i \in [0, C] \quad \forall i$$

6 α_i^* from dual program

$$\mu_i^* = C - \alpha_i^*$$

$$a^* = \sum_i \alpha_i^* x_i$$

$$\xi_i^* = \|x_i - a^*\|^2 - R^2 \quad \text{if } x_i \text{ outside ball}$$

$$= 0 \quad \text{if } x_i \text{ inside ball or on ball}$$

$$R^{*2} = \|x_i - a^*\|^2 \quad \text{for } x_i \text{ on ball}$$

$$R^{*2} = \frac{1}{N_{\text{ball}}} \sum_{i \in \text{s.v.}} \|x_i - a^*\|^2$$

7 $\|x_* - a\|^2 > R^2$?

$$x_*^T x - 2 x_*^T a^* + a^T a > R^{*2}$$

$$K(x_*, x_*) - 2 \sum_i \alpha_i K(x_i, x_i) + \sum_{ij} \alpha_i \alpha_j K(x_i, x_j)$$

$$> R^* \quad \text{or} \quad K(x_*, x_*) - 2 \sum_i \alpha_i K(x_i, x_i) > R_*^2 - \|a^*\|^2$$

⑨ $C = 0$ $R \Rightarrow 0$ ⑨

$C = \infty$ $R \Rightarrow$ outside all data-cases

⑨ Overfitted solution of the form

⑩ Same objective $R^2 + C \sum_i \xi_i$

$y_i(\|x_i - a\|^2 - R^2) \geq 1 - \xi_i$

↑ margin.

$\xi_i \geq 0$