Machine Learning 1 - Practice exercises 5

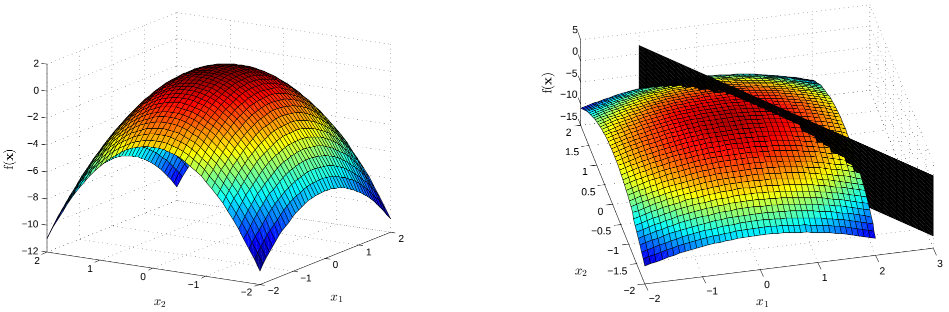
1 Lagrange Multipliers: Warm-up

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

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

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 satisfy 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 that the constraint function (black) separates $f(\mathbf{x})$ into two parts. The 3D view of $f(\mathbf{x})$ is slightly changed in Figure 1(b) for better visualization.



(a)

(b)

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

Answer the following questions below. Note that the first two questions are extra practice questions that will not be graded. They are just meant to provide you with more practice material.

1. **To practice, will not be graded!** 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

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

   Taking the derivatives with respect to x and λ are then

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

   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 $x_1 = \frac{2}{3}$.

2. **To practice, will not be graded!** 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:

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

   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 $x_1 + 2x_2 - 2x_3$, 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

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

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.

4. 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:

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

   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.

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 costs 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:

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

From (A), we get

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

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

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

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:

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

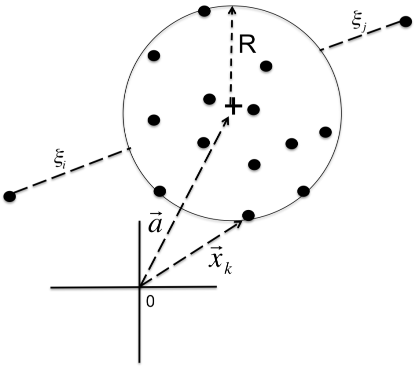


Figure 2: Kernel Outlier Detection

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:

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

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

Solution: The first three conditions are obtained by setting the derivative of the primal Lagrangian to zero with respect to R^2, a and ξ_i.

(a)

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

(b)

$$\frac{\partial \mathcal{L}}{\partial a} = -2 \sum_{i=1}^{N} \alpha_i (\boldsymbol{x}_i - \boldsymbol{a})$$
$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{a}} = 0 \implies \sum_{i=1}^{N} \alpha_i \boldsymbol{x}_i = \boldsymbol{a} \sum_{i=1}^{N} \alpha_i \implies \boldsymbol{a} = \sum_{i=1}^{N} \alpha_i \boldsymbol{x}_i.$$

(c)

$$\frac{\partial \mathcal{L}}{\partial \xi_i} = C - \alpha_i - \mu_i$$
$$\frac{\partial \mathcal{L}}{\partial \xi_i} = 0 \implies C - \alpha_i - \mu_i = 0, \forall i.$$

Note that the three conditions above are **not** KKT conditions. The KKT conditions for this model are given below:

(d) $\|\boldsymbol{x}_i - \boldsymbol{a}\|^2 - R^2 - \xi_i \leq 0 \ \forall i.$

(e) $\xi_i \geq 0 \ \forall i.$

(f) $\alpha_i \geq 0 \ \forall i.$

(g) $\mu_i \geq 0 \ \forall i.$

(h) $\alpha_i \left(\|\boldsymbol{x}_i - \boldsymbol{a}\|^2 - R^2 - \xi_i \right) = 0 \ \forall i.$

(i) $\mu_i \xi_i = 0 \ \forall i.$

3. Use these conditions to derive which data-cases $\boldsymbol{x_i}$ will have $\alpha_i > 0$ and

which ones will have $\mu_i > 0$.

Solution: The complementary slackness conditions are

- Inside the ball: $\|\boldsymbol{x}_i - \boldsymbol{a}\|^2 - R^2 - \xi_i < 0 \implies \alpha_i = 0 \implies \mu_i = C$ (conditions h and c).
- Outside the ball: $\xi_i > 0 \implies \mu_i = 0 \implies \alpha_i = C$ (conditions i and c).
- Inside or on ball: $\xi_i = 0 \implies \mu_i \geq 0$ (condition i).
- Outside or on ball: $\|\boldsymbol{x}_i - \boldsymbol{a}\|^2 - R^2 - \xi_i = 0 \implies \alpha_i \geq 0$ (condition h).

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: We use conditions a, b and c to eliminate R^2, a and ξ_i from the primal Lagrangian to obtain the dual representation.

$$\begin{aligned}
\mathcal{L}(a, R, \xi, \alpha, \mu) &= R^2 + C\sum_{i=1}^{N}\xi_i + \sum_{i=1}^{N}\alpha_i\left(\|\boldsymbol{x}_i - \boldsymbol{a}\|^2 - R^2 - \xi_i\right) - \sum_{i=1}^{N}\mu_i\xi_i \\
&= R^2 + C\sum_{i=1}^{N}\xi_i + \sum_{i=1}^{N}\alpha_i\boldsymbol{x}_i^T\boldsymbol{x}_i - 2\sum_{i=1}^{N}\alpha_i\boldsymbol{x}_i^T\boldsymbol{a} + \sum_{i=1}^{N}\alpha_i\boldsymbol{a}^T\boldsymbol{a} \\
&- \sum_{i=1}^{N}\alpha_i R^2 - \sum_{i=1}^{N}\alpha_i\xi_i - \sum_{i=1}^{N}\mu_i\xi_i \\
&= \left(R^2 - \sum_{i=1}^{N}\alpha_i R^2\right) + \sum_{i=1}^{N}(C - \alpha_i - \mu_i)\xi_i \\
&- 2\left(\sum_{i=1}^{N}\alpha_i\boldsymbol{x}_i^T\right)\boldsymbol{a} + \left(\sum_{i=1}^{N}\alpha_i\right)\boldsymbol{a}^T\boldsymbol{a} + \sum_{i=1}^{N}\alpha_i\boldsymbol{x}_i^T\boldsymbol{x}_i \\
&= \sum_{i=1}^{N}\alpha_i\boldsymbol{x}_i^T\boldsymbol{x}_i - \boldsymbol{a}^T\boldsymbol{a} \\
&= \sum_{i=1}\alpha_i\|\boldsymbol{x}_i\|^2 - \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j\boldsymbol{x}_i^T\boldsymbol{x}_j.
\end{aligned}$$

In the first step we expand the expression for the primal Lagrangian, in the second step we rearrange the terms, finally we apply the conditions. Kernelize the problem:

$$\begin{aligned}
\sum_{i=1}^{N}\alpha_i\boldsymbol{x}_i^T\boldsymbol{x}_i - \boldsymbol{a}^T\boldsymbol{a} &= \sum_{i=1}\alpha_i\|\boldsymbol{x}_i\|^2 - \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j\boldsymbol{x}_i^T\boldsymbol{x}_j \\
&= \sum_{i=1}^{N}\alpha_i K(\boldsymbol{x}_i, \boldsymbol{x}_i) - \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i K(\boldsymbol{x}_i, \boldsymbol{x}_j)\alpha_j
\end{aligned}$$

The dual program is

$$\arg\max_{\alpha}\sum_{i=1}^{N}\alpha_i K_{ii} - \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j K_{ij}$$

$$\text{with } \alpha_i \in [0, C]\ \forall i.$$

5. The dual program will return optimal values for $\{\alpha_i\}$. Assume that at least one of these is such that $0 < \alpha_i < C$. In terms of the optimal values for α_i, 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: Note that when $0 < \alpha_i < C$, then $\boldsymbol{x}_i$ must be on the ball. When $\alpha_i = 0$, then $\boldsymbol{x}_i$ is on or inside the ball, and when $\alpha_i = C$ then $\boldsymbol{x}_i$ is on or outside the ball.

$$\begin{aligned} \mu_i^* &= C - \alpha_i^* \\ \boldsymbol{a}^* &= \sum_{i=1}^{N} \alpha_i^* \boldsymbol{x}_i \\ R^{*2} &= \|\boldsymbol{x}_i - \boldsymbol{a}^*\|^2 \text{ (when } 0 < \alpha_i^* < C) \\ \xi_i^* &= \begin{cases} \|\boldsymbol{x}_i - \boldsymbol{a}^*\|^2 - R^{*2} & \text{(when } \alpha_i^* = C) \\ 0 & \text{(when } \alpha_i^* = 0) \end{cases} \end{aligned}$$

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: A new test case $\boldsymbol{x}_t$ is an outlier when

$$\begin{aligned} \|\boldsymbol{x}_t - \boldsymbol{a}^*\|^2 &> R^{*2} \\ \boldsymbol{x}_t^T \boldsymbol{x}_t - 2\boldsymbol{x}_t^T \boldsymbol{a}^* + \boldsymbol{a}^{*T} \boldsymbol{a}^* &> R^{*2} \\ K(\boldsymbol{x}_t, \boldsymbol{x}_t) - 2\sum_{i=1}^{N} \alpha_i^* \boldsymbol{x}_i^T \boldsymbol{x}_t + \sum_{i,j} \alpha_i \alpha_j \boldsymbol{x}_i^T \boldsymbol{x}_j &> R^{*2} \\ K(\boldsymbol{x}_t, \boldsymbol{x}_t) - 2\sum_{i=1}^{N} \alpha_i^* K(\boldsymbol{x}_i, \boldsymbol{x}_t) + \sum_{i,j} \alpha_i \alpha_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) &> R^{*2} \end{aligned}$$

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

Solution: When $C \to 0$ we expect $R \to 0$, and when $C \to \infty$ we expect R to be such that all data-cases are inside the ball.

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: We expect an over-fitted solution where none of the data-points are outliers. An RBF kernel will result in flexible and possibly disjunct decision boundaries, whereas a linear kernel can only result in a decision boundary shaped as a circle.

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: We replace the hard margin constraint, with a soft margin that allows some of the data-points to be misclassified. (See Bishop 7.20 and additional explanatory document).

$$\begin{aligned}&\min_{\boldsymbol{a},R,\xi} R^2 + C\sum_{i=1}^{N}\xi_i\\ &y_i\left(\|\boldsymbol{x}_i - \boldsymbol{a}\|^2 - R^2\right) \geq -\xi_i\\ &\xi_i \geq 0.\end{aligned}$$