Dr. Boqiang Huang

Mathematical Foundations of Data Analysis II — SS 2019

Problem Set 06

Published on 27.05.2019 To be collected on 04.06.2019

Problem 1: (4 points) $\underset{x}{\text{Max}}\,\underset{y}{\text{Min}} f(x,y) \le \underset{y}{\text{Min}}\,\underset{x}{\text{Max}} f(x,y)$
Consider a general optimization problem with inequality constraints

$$\begin{aligned} \underset{\boldsymbol{x}\in\mathbb{R}^p}{\text{Min}} \quad & f(\boldsymbol{x}) \\ \text{s.t.} \quad & g_i(\boldsymbol{x}) \le 0,\ i = 1,\dots,n, \end{aligned} \tag{1}$$

the corresponding Lagrangian function could be defined as

$$\mathcal{L}(\boldsymbol{x},\lambda) := f(\boldsymbol{x}) + \sum_{i=1}^{n} \lambda_i g_i(\boldsymbol{x}), \tag{2}$$

where $\lambda := [\lambda_1,\dots,\lambda_n]^T, \lambda_i \ge 0, i = 1,\dots,n$. If we define

$$\theta_{\text{p}}(\boldsymbol{x}) := \underset{\lambda,\,\lambda_i\ge 0}{\text{Max}}\ \mathcal{L}(\boldsymbol{x},\lambda), \tag{3}$$

one can notice that $\theta_{\text{p}}(\boldsymbol{x})$ will be the exact cost function if $\boldsymbol{x}$ satisfies all inequality constraints, which means that the optimization problem in (1) can be represented as the following version called the primal problem

$$\underset{\boldsymbol{x}}{\text{Min}}\ \theta_{\text{p}}(\boldsymbol{x}) = \underset{\boldsymbol{x}}{\text{Min}} \left(\underset{\lambda,\,\lambda_i\ge 0}{\text{Max}}\ \mathcal{L}(\boldsymbol{x},\lambda) \right). \tag{4}$$

Similarly, if we define

$$\theta_{\text{d}}(\lambda) := \underset{\boldsymbol{x}}{\text{Min}}\ \mathcal{L}(\boldsymbol{x},\lambda), \tag{5}$$

the corresponding dual problem of the original optimization in (1) can be represented by

$$\underset{\lambda,\,\lambda_i\ge 0}{\text{Max}}\ \theta_{\text{d}}(\lambda) = \underset{\lambda,\,\lambda_i\ge 0}{\text{Max}} \left(\underset{\boldsymbol{x}}{\text{Min}}\ \mathcal{L}(\boldsymbol{x},\lambda) \right). \tag{6}$$

Please show that

$$\underset{\lambda,\,\lambda_i\ge 0}{\text{Max}} \left(\underset{\boldsymbol{x}}{\text{Min}}\ \mathcal{L}(\boldsymbol{x},\lambda) \right) \le \mathcal{L}(\boldsymbol{x}^*,\lambda^*) \le \underset{\boldsymbol{x}}{\text{Min}} \left(\underset{\lambda,\,\lambda_i\ge 0}{\text{Max}}\ \mathcal{L}(\boldsymbol{x},\lambda) \right), \tag{7}$$

where $(\boldsymbol{x}^*,\lambda^*)$ is the solution of the optimization problem.
Remark: There is no limitation of the form of either primal or dual problem. If primal version is the maximization, the dual one will be the minimization. The point is that we may obtain the benefit if one optimization is easier to be solved to the other.

Problem 2: (8 points)
(a) Separating boundary without violation in the support vector machine (SVM)
Suppose a set of nonlinear basis functions is well designed such that the separating boundary/hyperplane could be represented as

$$\beta_0 + \beta^T \boldsymbol{x} = 0 \tag{8}$$

for a simple two-class linear/nonlinear SVM problem, where $\beta_0 \in \mathbb{R}$, $\beta := [\beta_1,\dots,\beta_p]^T \in \mathbb{R}^p$, and $\boldsymbol{x} := [x_1,\dots,x_p]^T \in \mathbb{R}^p$. If there is no violation between each observation pair $(y_i,\boldsymbol{x}_i), i = 1,\dots,n$ and the ideal separating boundary defined in (8), one can formulate the following quadratic minimization as the primal version in order to obtain the optimal parameter in (8)

$$\begin{aligned} \underset{\beta_0,\beta}{\text{Min}} \quad & \frac{1}{2}\|\beta\|_2^2 \\ \text{s.t.} \quad & y_i(\beta_0 + \beta^T \boldsymbol{x}_i) \ge 1,\ i = 1,\dots,n. \end{aligned} \tag{9}$$

Now, try to derive the dual version of the minimization problem in the following form

$$\begin{aligned} \underset{\lambda}{\text{Max}} \quad & \mathbf{1}^T\lambda - \frac{1}{2}\lambda^T K \lambda \\ \text{s.t.} \quad & \boldsymbol{y}^T\lambda = 0, \\ & \lambda \geq \mathbf{0}. \end{aligned} \tag{10}$$

where $\boldsymbol{y} := [y_1, \ldots, y_n]^T$, the (i, j)th element of the square matrix K represents the relationship between $(y_i, \boldsymbol{x}_i)$ and $(y_j, \boldsymbol{x}_j)$.

(b) Separating boundary with violations in the support vector machine (SVM)

Suppose that there are a few observations $(y_k, \boldsymbol{x}_k)$ violate the resulted separating boundary, the primal problem in (9) could be modified as

$$\begin{aligned} \underset{\beta_0, \beta, \varepsilon}{\text{Min}} \quad & \frac{1}{2}\|\beta\|_2^2 + C\|\varepsilon\|_1 \\ \text{s.t.} \quad & y_i(\beta_0 + \beta^T\boldsymbol{x}_i) \geq 1 - \varepsilon_i,\ i = 1, \ldots, n; \\ & \varepsilon_i \geq 0,\ i = 1, \ldots, n. \end{aligned} \tag{11}$$

where parameter C controls the size of margin. Now, try to derive the dual version of (11) which has a similar form shown in (10).

Problem 3: (4 points)

(a) Given four training data in $\mathbb{R}^2$ as ($y_1 = 1, \boldsymbol{x}_1 = [0, 0]^T$), ($y_2 = -1, \boldsymbol{x}_2 = [1, 0]^T$), ($y_3 = 1, \boldsymbol{x}_3 = [1, 1]^T$), ($y_4 = -1, \boldsymbol{x}_4 = [0, 1]^T$). Try to determine a separating hyperplane from the primal optimization problem;

(b) Try to determine a separating hyperplane from the dual version of the optimization problem shown in (a).

Remark: Theoretically speaking, to solve the dual version one may need the sequential minimal optimization (SMO) method. Since only the vector λ is unknown in the dual version, it will be easier to obtain the solution if the size of the unknown is too large in the primal problem. For example, let the following problem represent a dual version of the SVM optimization

$$\begin{aligned} \underset{\lambda}{\text{Max}} \quad & \mathcal{L}(\lambda) := \mathbf{1}^T\lambda - \frac{1}{2}\lambda^T K \lambda \\ \text{s.t.} \quad & \boldsymbol{y}^T\lambda = 0, \\ & 0 \leq \lambda_i \leq C, i = 1, \ldots, n. \end{aligned} \tag{12}$$

From the equality constraints, we can represent one λ_i as a linear combination of the others from the vector λ. Suppose that $\lambda^{(k)}$ is the candidate vector at the kth iteration, we can update two selected elements λ_i and λ_j in the $k+1$ iteration by

(1) fix the other $n-2$ elements, and set $\lambda_m^{(k+1)} = \lambda_m^{(k)}, m \neq i$ and $m \neq j$;

(2) obtain the optimal $\lambda_i^{(k+1)} = \underset{\lambda_i}{\text{argmax}}\ \mathcal{L}(\lambda_i)$, and update $\lambda_j^{(k+1)}$ by the rest $n-1$ elements;

Then, we implement the update iteration until the algorithm converges to the optimal solution.

In fact, the so-called Karush-Kuhn-Tucker (KKT) conditions (see Chapter 10 of the optimization reference book) may be considered for the convergence determination. For this problem, it is not necessary to check the KKT conditions.

Problem 4: (4 points) k-means clustering

Consider the following data set consisting of the scores of two variables on each of seven individuals:

Subject	**A**	**B**
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

Try to separate the data set into two clusters by using the k-means method.