Machine Learning and Computational Statistics

Theodoros Georgiopoulos

Homework 9

Exercises for Unit 9: Support Vector Machines

Solution for exercise 1

Consider a hyperplane L defined by the equation $\theta_0 + \boldsymbol{\theta}^T \boldsymbol{x} = 0$. For SVM problem, we want to maximize the distance of our data points from this decision boundary. The signed distance of any point $\boldsymbol{x}$ to L is given by $\frac{1}{\|\boldsymbol{\theta}\|}(\theta_0 + \boldsymbol{\theta}^T \boldsymbol{x})$. If we have binary classification problem with $y \in \{-1, 1\}$ and we have 2 decision boundaries, we have

$$\begin{aligned}\theta_0 + \boldsymbol{\theta}^T \boldsymbol{x}_i &\geq +1, \text{if } y_i = +1 \\ \theta_0 + \boldsymbol{\theta}^T \boldsymbol{x}_i &\leq -1, \text{if } y_i = -1.\end{aligned}$$

This problem is equivalent to

$$\begin{gathered}\min \frac{1}{2}\|\boldsymbol{\theta}\|^2 \\ \text{subject to } y_i(\boldsymbol{\theta}^T \boldsymbol{x}_i + \theta_0) \geq 1, i = 1, \ldots, N.\end{gathered}$$

(a) The Lagrangian function is given by

$$L(\boldsymbol{\theta}, \theta_0, \lambda) = \frac{1}{2}\|\boldsymbol{\theta}\|^2 - \sum_{i=1}^{N} \lambda_i \left[y_i(\boldsymbol{\theta}^T \boldsymbol{x}_i + \theta_0) - 1 \right].$$

Taking derivatives with respect to $\boldsymbol{\theta}$ and θ_0 and setting them to zero, gives

$$\frac{\partial L}{\boldsymbol{\theta}} = 0 \iff \boldsymbol{\theta} = \sum_{i=1}^{N} \lambda_i y_i \boldsymbol{x}_i \tag{1}$$

$$\frac{\partial L}{\theta_0} = 0 \iff \sum_{i=1}^{N} \lambda_i y_i = 0 \tag{2}$$

(b) Using (1),(2) to the Lagrangian function gives

$$\begin{aligned}L(\boldsymbol{\theta}, \theta_0, \lambda) = \frac{1}{2} \left(\sum_{i=1}^{N} \lambda_i y_i \boldsymbol{x}_i \right)^T \sum_{i=1}^{N} \lambda_i y_i \boldsymbol{x}_i - \sum_{i=1}^{N} \lambda_i \left[y_i \left(\left(\sum_{j=1}^{N} \lambda_j y_j \boldsymbol{x}_j \right)^T \boldsymbol{x}_i + \theta_0 \right) - 1 \right] \\ = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_i - \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_i - \sum_{i=1}^{N} \lambda_i y_i \theta_0 + \sum_{i=1}^{N} \lambda_i \\ = -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_i + \sum_{i=1}^{N} \lambda_i.\end{aligned}$$

(c) So, the Wolfe dual representation is

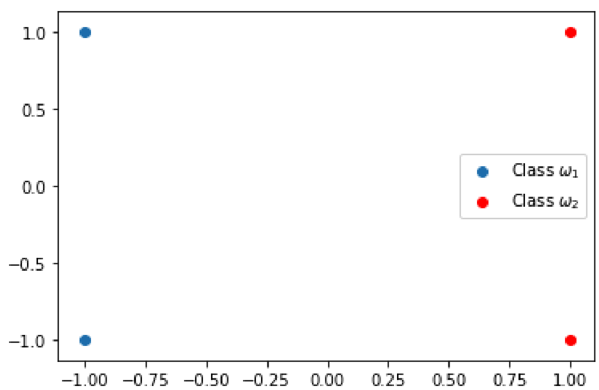
$$\max_{\lambda\geq 0} \sum_{i=1}^{N} \lambda_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_i\lambda_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_i$$

$$\text{subject to} \sum_{i=1}^{N} \lambda_i y_i = 0, \quad i = 1, \ldots, N.$$

Solution for exercise 2

We consider a two-class two-dim. problem where $\omega_1 = +1$ consists of $\boldsymbol{x}_1 = [-1, 1]^T, \boldsymbol{x}_2 = [-1, -1]^T$, while class $\omega_2 = -1$ consists of $\boldsymbol{x}_3 = [1, -1]^T, \boldsymbol{x}_4 = [1, 1]^T$.

(a) The data points are depicted bellow.



(b) Using the dual representation of the SVM problem, the lagrange function becomes

$$L = \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 - \lambda_1^2 - \lambda_2^2 - \lambda_3^2 - \lambda_4^2 - 2\lambda_1\lambda_3 - 2\lambda_2\lambda_4.$$

Taking the derivatives and setting them to zero, gives

$$\lambda_1 + \lambda_3 = \frac{1}{2}, \tag{3}$$

$$\lambda_2 + \lambda_4 = \frac{1}{2}. \tag{4}$$

From $\sum_{i=1}^{N} \lambda_i y_i = 0$, we have

$$\lambda_1 + \lambda_2 = \lambda_3 + \lambda_4. \tag{5}$$

From (3),(4),(5) we have

$$\lambda_1 = \lambda_4 \tag{6}$$

$$\lambda_2 = \lambda_3. \tag{7}$$

From (3),(4),(6),(7) and the fact that $\lambda \geq 0$ we have that $\lambda_1 = \lambda_4 = u$ and $\lambda_2 = \lambda_3 = 1 - u$. From (1), we can find θ as follows

$$\boldsymbol{\theta} = \sum_{i=1}^{N} \lambda_i y_i \boldsymbol{x}_i = u[-1, 1]^T + u[-1, -1]^T - (1 - u)[1, -1]^T - (1 - u)[1, 1]^T = [-1, 0]^T.$$

From KKT we have that

$$\lambda_i \left[y_i(\boldsymbol{\theta}^T \boldsymbol{x}_i + \theta_0) - 1 \right] = 0, \forall i$$

and substituting the values of $y_i, \boldsymbol{\theta}, \boldsymbol{x}_i$ gives $\theta_0 = 0$. So, the classification line is

$$x_0 = 0.$$

Solution for exercise 3

Assume that the points are linear separable between the two classes $\omega_1 = 1$ and $\omega_2 = 0$. The feature space is 3-dim, so the hyperplane equation is

$$f(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3.$$

For the data points belonging to ω_1 we have that $f(x) > 0$ and for ω_2 we have that $f(x) < 0$. Substituting x_1, x_2, x_3 gives

$$\begin{align}
\theta_0 &> 0 \tag{8}\\
\theta_0 + \theta_3 &< 0 \tag{9}\\
\theta_0 + \theta_2 &< 0 \tag{10}\\
\theta_0 + \theta_2 + \theta_3 &< 0 \tag{11}\\
\theta_0 + \theta_1 &< 0 \tag{12}\\
\theta_0 + \theta_1 + \theta_3 &< 0 \tag{13}\\
\theta_0 + \theta_1 + \theta_2 &< 0 \tag{14}\\
\theta_0 + \theta_1 + \theta_2 + \theta_3 &> 0. \tag{15}
\end{align}$$

From (9),(10),(12),(15) we have that $2\theta_0 < 0$ which contradicts (8). So, our data are not linear separable.