SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

01.112 Machine Learning, Fall 2019
Supplementary Lecture Notes

8.5. Support Vector Machines (III)

Thursday, 10 October, 2019

1 Soft-Margin Support Vector Machines

The primal form of the problem in the notes is as follows:

$$\min \frac{\lambda}{2} \|\theta\|^2 + \sum_{t=1}^{n} \xi_t \tag{1}$$

$$\text{subject to } y^{(t)}(\theta \cdot x^{(t)} + \theta_0) \geq 1 - \xi_t, \xi_t \geq 0, \text{ for } t = 1, \ldots, n \tag{2}$$

which is equivalent to the following:

$$\min \frac{1}{2} \|\theta\|^2 + C \sum_{t=1}^{n} \xi_t \tag{3}$$

$$\text{subject to } y^{(t)}(\theta \cdot x^{(t)} + \theta_0) \geq 1 - \xi_t, \xi_t \geq 0, \text{ for } t = 1, \ldots, n \tag{4}$$

where $C = 1/\lambda$.

Okay, now let us focus on the last optimization problem. How do we optimize this problem? Remember we introduced Lagrangian multipliers to convert the original constrained optimization problem into a new problem. Here, similarly, we introduce the following Lagrangian:

$$\mathcal{L}(\theta, \theta_0, \xi, \alpha, \beta) = \frac{1}{2} \|\theta\|^2 + C \sum_{t=1}^{n} \xi_t + \sum_{t=1}^{n} \alpha_t (1 - \xi_t - y^{(t)}(\theta \cdot x^{(t)} + \theta_0)) + \sum_{t=1}^{n} \beta(-\xi_t) \tag{5}$$

where $\alpha_t \geq 0, \beta_t \geq 0$ for $t = 1, \ldots, n$. As we have discussed in class, the optimization problem is now as follows (why?):

$$\min_{\theta, \theta_0, \xi,} \max_{\alpha \geq 0, \beta \geq 0} \mathcal{L}(\theta, \theta_0, \xi, \alpha, \beta) \tag{6}$$

Note that here θ, α, ξ and β are vectors. $\alpha \geq 0$ means all elements in the vector α are non-negative. This problem is in fact equivalent to:

$$\max_{\alpha \geq 0, \beta \geq 0} \min_{\theta, \theta_0, \xi,} \mathcal{L}(\theta, \theta_0, \xi, \alpha, \beta) \tag{7}$$

Why do we do this? Because we hope that we can rewrite the following term as a function $f(\alpha, \beta)$, i.e., a function that involves only α and β. If we could do so, then we can just optimze such a function $f(\alpha, \beta)$ by tuning α and β:

$$\min_{\theta,\theta_0,\xi,} \mathcal{L}(\theta, \theta_0, \xi, \alpha, \beta) \tag{8}$$

Okay. Now let's look at the above sub-problem. First of all, this is an optimization problem. α and β are now free variables, so we could regard them as "constants". Let's try to see what conditions we could have if the minimum is reached. Naturally if we take the following gradients they shall be zeros when the minimum is reached:

$$\frac{\partial \mathcal{L}(\theta, \theta_0, \xi, \alpha, \beta)}{\partial \theta} = 0 \tag{9}$$

$$\frac{\partial \mathcal{L}(\theta, \theta_0, \xi, \alpha, \beta)}{\partial \theta_0} = 0 \tag{10}$$

$$\frac{\partial \mathcal{L}(\theta, \theta_0, \xi, \alpha, \beta)}{\partial \xi_t} = 0 \qquad \text{for each } t \tag{11}$$

These conditions give us the following facts:

$$\theta - \sum_{t=1}^{n} \alpha_t y^{(t)} x^{(t)} = 0 \tag{12}$$

$$-\sum_{t=1}^{n} \alpha_t y^{(t)} = 0 \tag{13}$$

$$C - \alpha_t - \beta_t = 0 \qquad \text{for each } t \tag{14}$$

Now, let's look at the $\mathcal{L}$ function again, and see what simplifications we can make based on the above facts:

$$\mathcal{L}(\theta, \theta_0, \xi, \alpha, \beta) \tag{15}$$

$$= \frac{1}{2} \|\theta\|^2 + C \sum_{t=1}^{n} \xi_t + \sum_{t=1}^{n} \alpha_t (1 - \xi_t - y^{(t)}(\theta \cdot x^{(t)} + \theta_0)) + \sum_{t=1}^{n} \beta_t (-\xi_t) \tag{16}$$

$$= \frac{1}{2} \|\theta\|^2 + \sum_{t=1}^{n} (C - \alpha_t - \beta_t) \xi_t + \sum_{t=1}^{n} \alpha_t (1 - y^{(t)}(\theta \cdot x^{(t)} + \theta_0)) \tag{17}$$

Since $C - \alpha_t - \beta_t = 0$ we can drop the 2nd term above:

$$\mathcal{L}(\theta, \theta_0, \xi, \alpha, \beta) \tag{18}$$

$$= \frac{1}{2} \|\theta\|^2 + \sum_{t=1}^{n} \alpha_t (1 - y^{(t)}(\theta \cdot x^{(t)} + \theta_0)) \tag{19}$$

$$= \frac{1}{2} \|\theta\|^2 + \sum_{t=1}^{n} \alpha_t - \sum_{t=1}^{n} \alpha_t y^{(t)}(\theta \cdot x^{(t)} + \theta_0) \tag{20}$$

$$= \frac{1}{2} \|\theta\|^2 + \sum_{t=1}^{n} \alpha_t - \sum_{t=1}^{n} \alpha_t y^{(t)}(\theta \cdot x^{(t)}) - \sum_{t=1}^{n} \alpha_t y^{(t)} \theta_0 \tag{21}$$

$$= \frac{1}{2} \|\theta\|^2 + \sum_{t=1}^{n} \alpha_t - \left(\sum_{t=1}^{n} \alpha_t y^{(t)} x^{(t)} \right) \cdot \theta - \left(\sum_{t=1}^{n} \alpha_t y^{(t)} \right) \theta_0 \tag{22}$$

Since $\theta = \sum_{t=1}^{n} \alpha_t y^{(t)} x^{(t)}$ and $\sum_{t=1}^{n} \alpha_t y^{(t)} = 0$ we have:

$$\mathcal{L}(\theta, \theta_0, \xi, \alpha, \beta) \tag{23}$$

$$= \frac{1}{2} \|\theta\|^2 + \sum_{t=1}^{n} \alpha_t - \theta \cdot \theta \tag{24}$$

$$= -\frac{1}{2} \|\theta\|^2 + \sum_{t=1}^{n} \alpha_t \tag{25}$$

$$= \sum_{t=1}^{n} \alpha_t - \frac{1}{2} \sum_{t'=1}^{n} \sum_{t=1}^{n} \alpha_t \alpha_{t'} y^{(t)} y^{(t')} x^{(t)} \cdot x^{(t')} \tag{26}$$

Okay, now we have reached a form that involves only α (well, we expected to have a form that involves both α and β, but it turns out there is no β in this form, which is even better). Now let's go back to our original problem. We would like to do the following:

$$\max_{\alpha \geq 0, \beta \geq 0} \min_{\theta, \theta_0, \xi,} \mathcal{L}(\theta, \theta_0, \xi, \alpha, \beta) \tag{27}$$

$$= \max_{\alpha \geq 0, \beta \geq 0} \min_{\theta, \theta_0, \xi,} \sum_{t=1}^{n} \alpha_t - \frac{1}{2} \sum_{t'=1}^{n} \sum_{t=1}^{n} \alpha_t \alpha_{t'} y^{(t)} y^{(t')} x^{(t)} \cdot x^{(t')} \tag{28}$$

Since the above form only involves α, we can remove all other variables, leading to the following problem:

$$\max_{\alpha} \sum_{t=1}^{n} \alpha_t - \frac{1}{2} \sum_{t'=1}^{n} \sum_{t=1}^{n} \alpha_t \alpha_{t'} y^{(t)} y^{(t')} x^{(t)} \cdot x^{(t')} \tag{29}$$

or :

$$\min_{\alpha} \frac{1}{2} \sum_{t'=1}^{n} \sum_{t=1}^{n} \alpha_t \alpha_{t'} y^{(t)} y^{(t')} x^{(t)} \cdot x^{(t')} - \sum_{t=1}^{n} \alpha_t \tag{30}$$

What are the constraints? Here they are:

$$\alpha_t \geq 0, \beta_t \geq 0 \tag{31}$$

for $t = 1, \ldots, n$, and (from Equation 17, which involves α only):

$$\sum_{t=1}^{n} \alpha_t y^{(t)} = 0 \tag{32}$$

However the objective does not involve β, and $\beta_t = C - \alpha_t$. This means the second set of constraints are really $C - \alpha_t \geq 0$, or $\alpha_t \leq C$ for all t. This concludes that the constraints are:

$$0 \leq \alpha_t \leq C \text{ for all } t = 1, \ldots, n \tag{33}$$

$$\sum_{t=1}^{n} \alpha_t y^{(t)} = 0 \tag{34}$$

Okay. Finally here comes our dual form for the SVM with soft-margin:

$$\min_{\alpha} \frac{1}{2} \sum_{t'=1}^{n} \sum_{t=1}^{n} \alpha_t \alpha_{t'} y^{(t)} y^{(t')} x^{(t)} \cdot x^{(t')} - \sum_{t=1}^{n} \alpha_t \tag{35}$$

subject to :

$$0 \leq \alpha_t \leq C \text{ for all } t = 1, \ldots, n \tag{36}$$

$$\sum_{t=1}^{n} \alpha_t y^{(t)} = 0 \tag{37}$$

This is in fact a standard quadratic program. There are standard algorithms for solving such a quadratic program (in fact in the case of linear SVM, there exists very efficient algorithms that involve