Introduction to Machine Learning **Fall Semester, 2014/5**

Recitation 7: December 7

Lecturer: Regev Schweiger *Scribe: Yishay Mansour*

7.1 SVM optimization

In the lecture we saw the following optimization problem, for a maximum margin classifier.

$$\begin{aligned} &\min_{\mathbf{w},b} \frac{1}{2}\mathbf{w}^T\mathbf{w} \\ &\text{s.t. } y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1 \qquad \forall n = 1,\ldots,N \end{aligned}$$

where $\mathbf{w} \in \mathbb{R}^d$ is the weight vector, $b \in \mathbb{R}$ is the bias, and $(\mathbf{x}_n, y_n)$ are the examples and $\mathbf{x}_n \in \mathbb{R}^d$ and $y_n \in \{+1, -1\}$.

The first step is to write the Lagrangian. In general, for a program

$$\begin{aligned} &\min f(\mathbf{z}) \\ &\text{s.t. } g_i(\mathbf{z}) \leq 0 \qquad \forall i = 1,\ldots,N \end{aligned}$$

the Lagrangian is

$$L(\mathbf{z}, \boldsymbol{\alpha}) = f(\mathbf{z}) + \sum_{i=1}^{N} \alpha_i g_i(\mathbf{z})$$

where $\boldsymbol{\alpha}$ are called the *Lagrangian multipliers*. The KKT conditions (on which we do not elaborate here) tell us that the dual program:

$$\begin{aligned} &\max_{\boldsymbol{\alpha}} \min_{\mathbf{z}} L(\mathbf{z}, \boldsymbol{\alpha}) \\ &\text{s.t. } \alpha_i \geq 0 \qquad \forall i = 1,\ldots,N \end{aligned}$$

achieves the same optimal point for all the cases which we will consider.

For our SVM program we get

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{n=1}^{N} \alpha_n (y_n(\mathbf{w}^T\mathbf{x}_n + b) - 1)$$

In our case, we now take the derivative of L and equate it with zero to minimize over $\mathbf{w}$ and b.

$$\nabla_{\mathbf{w}} L = \mathbf{w} - \sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n = 0 \quad \Longrightarrow \quad \mathbf{w} = \sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n$$

this gives us a way to compute $\mathbf{w}$ given $\boldsymbol{\alpha}$. We call this the $\mathbf{w}$-constraint. For b we have

$$\frac{d}{db}L = -\sum_{n=1}^{N}\alpha_n y_n = 0 \quad \Longrightarrow \quad \sum_{n=1}^{N}\alpha_n y_n = 0$$

We call this the b-constraint. Plugging the constraints back in L we have

$$\begin{aligned}
L(\mathbf{w}, b, \boldsymbol{\alpha}) =& \frac{1}{2}\mathbf{w}^T\mathbf{w} - \mathbf{w}^T \underbrace{\left(\sum_{n=1}^{N}\alpha_n y_n \mathbf{x}_n\right)}_{\mathbf{w}} - b\underbrace{\left(\sum_{n=1}^{N}\alpha_n y_n\right)}_{0} + \left(\sum_{n=1}^{N}\alpha_n\right) \\
=& -\frac{1}{2}\mathbf{w}^T\mathbf{w} + \left(\sum_{n=1}^{N}\alpha_n\right) \\
=& -\frac{1}{2}\left(\sum_{i=1}^{N}\alpha_i y_i \mathbf{x}_i\right)^T\left(\sum_{j=1}^{N}\alpha_j y_j \mathbf{x}_j\right) + \sum_{n=1}^{N}\alpha_n \\
=& -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j + \sum_{n=1}^{N}\alpha_n
\end{aligned}$$

where we have the constraints $\sum_{n=1}^{N}\alpha_n y_n = 0$ and $\forall n$ we have $\alpha_n \geq 0$.

Formally, the dual problem is

$$\begin{aligned}
\max_{\boldsymbol{\alpha}} L(\mathbf{w}, b, \boldsymbol{\alpha}) = \min_{\boldsymbol{\alpha}} & \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j - \sum_{n=1}^{N}\alpha_n \\
\text{s.t.} \quad & \sum_{n=1}^{N}\alpha_n y_n = 0 \\
& \forall n \quad \alpha_n \geq 0
\end{aligned}$$

When $\alpha_n > 0$, this means the constraint $y_n(\mathbf{w}^T\mathbf{x}_n + b) = 1$ must be satisfied (otherwise, $\alpha_n = 0$ would give a smaller solution). Therefore, the support vectors are those with $\alpha_n > 0$.

7.2 Unrealizable Case

We add slack variables ξ_n to ensure feasibility. We have,

$$\begin{aligned}
\min_{\mathbf{w},b,\boldsymbol{\xi}} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{n=1}^{N}\xi_n \\
\text{s.t.} \quad & y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1 - \xi_n \qquad \forall n = 1, \ldots, N \\
& \xi_n \geq 0
\end{aligned}$$

We can now write the Lagrangian

$$L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \mathbf{r}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{n=1}^{N}\xi_n - \sum_{n=1}^{N}\alpha_n(y_n(\mathbf{w}^T\mathbf{x}_n + b) - 1 + \xi_n) - \sum_{n=1}^{N} r_n\xi_n$$

We now take the derivatives

$$\nabla_{\mathbf{w}} L = \mathbf{w} - \sum_{n=1}^{N}\alpha_n y_n \mathbf{x}_n = 0 \quad \Longrightarrow \quad \mathbf{w} = \sum_{n=1}^{N}\alpha_n y_n \mathbf{x}_n$$

identically as before. For b we have

$$\frac{d}{db}L = -\sum_{n=1}^{N}\alpha_n y_n = 0 \quad \Longrightarrow \quad \sum_{n=1}^{N}\alpha_n y_n = 0$$

also as before. For ξ_n we have

$$\frac{d}{d\xi_n}L = C - \alpha_n - r_n = 0 \quad \Longrightarrow \quad \alpha_n = C - r_n$$

Substituting the constraints in L, we get

$$\begin{aligned} L(\mathbf{w}, b, \boldsymbol{\alpha}) =& \frac{1}{2}\mathbf{w}^T\mathbf{w} - \mathbf{w}^T \underbrace{\left(\sum_{n=1}^{N}\alpha_n y_n \mathbf{x}_n\right)}_{\mathbf{w}} - b\underbrace{\left(\sum_{n=1}^{N}\alpha_n y_n\right)}_{0} + \left(\sum_{n=1}^{N}\alpha_n\right) + \sum_{n=1}^{N}\xi_n \underbrace{(C - \alpha_n - r_n)}_{0} \\ =& -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j + \sum_{n=1}^{N}\alpha_n \end{aligned}$$

identically to before. The only difference is that now we have two additional constraints, $r_n \geq 0$ and $\alpha_n = C - r_n$. Since r_n does not appear in the optimization, we can drop it, and join the two constraints to $\alpha_n \leq C$. (For any solution of α_n we can set $r_n = C - \alpha_n$.)

Note that when we have an error in classification or in the margin, then $\xi_n > 0$ and therefore $r_n = 0$, which implies that $\alpha_n = C$.

If $C > \alpha_n > 0$, this means as before that $y_n(\mathbf{w}^T\mathbf{x}_n + b) = 1$ and $\xi_n = 0$, and thus $\mathbf{x}_n$ is a support vector.

7.3 Importance SVM

We add a variable $0 \leq v_n \leq 1$ that defines the n-th sample's importance. The margin error is multiplied by this weight:

$$\begin{aligned}
&\min_{\mathbf{w},b,\boldsymbol{\xi}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{n=1}^{N}\xi_n v_n \\
&\text{s.t. } y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1 - \xi_n \qquad \forall n = 1, \ldots, N \\
&\quad \forall n \;\; \xi_n \geq 0
\end{aligned}$$

The solution of the dual program is the same as above, with the condition $0 \leq \alpha_n \leq Cv_n$. The importance thus limits the Lagrange multiplier from above.