SMM Kernel method

Chanwoo Lee, April 12, 2020

1 Linear case

First, consider linear case of SMM. We can generalize our previous approach as

$$
\begin{aligned}
(P) \quad & \min_{B,U,V,b,\boldsymbol{\xi}} \frac{1}{2}\|B\|^2 + C\sum_{i=1}^N \xi_i, \\
& \text{subject to } y_i\left(\langle B, H_U X_i H_V\rangle + b\right) \geq 1-\xi_i \\
& \xi_i \geq 0 \quad i = 1,\ldots,N.
\end{aligned} \tag{1}
$$

We can have Laglangian equation as

$$
L(B,U,V,v,\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\mu}) = \frac{1}{2}\|B\|^2 + C\sum_{i=1}^N \xi_i - \sum_{i=1}^N \alpha_i\left(y_i\left(\langle B, H_U X_i H_V\rangle + b\right) - (1-\xi_i)\right) - \sum_{i=1}^N \mu_i\xi_i.
$$

By the first order necessity condition, we have

$$
\begin{aligned}
B &= \sum_{i=1}^N \alpha_i y_i H_U X_i H_V \\
&= UV^T \quad \text{(Special case).}
\end{aligned}
$$

By setting $B = UV^T$ we could have easier optimization such as

$$
\begin{aligned}
(P') \quad & \min_{U,V,b,\boldsymbol{\xi}} \frac{1}{2}\|UV^T\|^2 + C\sum_{i=1}^N \xi_i, \\
& \text{subject to } y_i\left(\langle UV^T, H_U X_i H_V\rangle + b\right) \geq 1-\xi_i \\
& \xi_i \geq 0 \quad i = 1,\ldots,N.
\end{aligned}
$$

Using the fact $\langle UV^T, H_U X_i H_V\rangle = \langle UV^T, X_i\rangle$, we could successfully find the algorithm using alternating update approach. To be specific, we could handle derivative of the inner product $\langle UV^T, H_U X_i H_V\rangle$ fixing the other matrix. This gives us update direction of U and V and makes it possible to take alternating update approach.

What if we stick to find B without having the structure $B = UV^T$ in (1)? I could not find a good algorithm to find optimizer. The main reason for this is that the derivatives of $\langle B, H_U X_i H_V\rangle$ with respect to U and V are formidable. Nonlinear kernel case also experiences the same trouble if we do not have good structure to avoid the derivatives.

2 Non linear case

We can interpret nonlinear kernel case as a generalization of linear case in (1). Suppose we have feature map such as $h: \mathbb{R}^{m\times n} \to \mathbb{R}^{m'\times n'}$, then we have SMM kernel method as

$$
(P) \quad \min_{B\in\mathbb{R}^{m'\times n'},U,V,b,\boldsymbol{\xi}} \frac{1}{2}\|B\|^2 + C\sum_{i=1}^N \xi_i, \tag{2}
$$

$$\text{subject to } y_i\left(\langle B, h(H_U X_i H_V)\rangle + b\right) \geq 1 - \xi_i$$
$$\xi_i \geq 0 \quad i = 1, \ldots, N.$$

When h is an identity map, then we have linear case SMM. We have the Laglangian equation as

$$L(B, U, V, v, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\mu}) = \frac{1}{2}\|B\|^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i\left(y_i\left(\langle B, h(H_U X_i H_V)\rangle + b\right) - (1 - \xi_i)\right) - \sum_{i=1}^{N}\mu_i\xi_i.$$

By the first order necessary condition, we have

$$B = \sum_{i=1}^{N}\alpha_i y_i h(H_U X_i H_V).$$

By the same reason on the linear case, we cannot use gradient based method without additional assumption on B or h.

Let us define good kernel function to handle this issue.

Definition 1. $K(X_i, X_j)$ is a divisible kernel if there exist h such that

1. $K(X_i, X_j) = \langle h(X_i), h(X_j)\rangle$.
2. There exists a function g such that $\langle h(H_U X_i), h(H_U X_j)\rangle = \langle H_{g(U)}h(X_i), H_{g(U)}h(H_j)\rangle$.
3. There exists a function g' such that $\langle h(X_i H_V), h(X_j H_V)\rangle = \langle h(X_i)H_{g'(V)}, h(H_j)H_{g'(V)}\rangle$.

If given kernel K is a divisible kernel, we can restrict coefficient space for B to have tractable algorithm.

$$\begin{aligned} B &= \sum_{i=1}^{N}\alpha_i y_i h(H_U X_i H_V) = \sum_{i=1}^{N}\alpha_i y_i H_{g(U)}h(X_i)H_{g'(V)} \\ &= g(U)g'(V)^T \quad \text{(Special case).} \end{aligned}$$

By setting $B = g(U)g'(V)^T$ (we do not have to know what exactly g, g', and h are), we have

$$(P') \qquad \min_{g(U), g'(V), b, \boldsymbol{\xi}} \frac{1}{2}\|g(U)g'(V)^T\|^2 + C\sum_{i=1}^{N}\xi_i, \tag{3}$$
$$\text{subject to } y_i\left(\langle g(U)g'(V)^T, h(H_U X_i H_V)\rangle + b\right) \geq 1 - \xi_i$$
$$\xi_i \geq 0 \quad i = 1, \ldots, N.$$

We can rewrite $\langle g(U)g'(V)^T, h(H_U X_i H_V)\rangle$ as

$$\langle g(U)g'(V)^T, h(X_i)\rangle = \langle g(U), h(X_i H_V)\rangle = \langle g'(V)^T, h(H_U X_i)\rangle$$

Therefore, we can use alternating update approach fixing the other matrix. One thing to note is in dual problem for (3), the knowledge of the kernel function K is enough. For example, if we fix V to update U, the dual problem is

$$(D') \quad \min_{\boldsymbol{\alpha} \geq 0} -\sum_{i=1}^{N}\alpha_i + \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j\langle h(X_i H_V), h(X_j H_V)\rangle,$$

$$\text{subject to } \sum_{i=1}^{N} y_i \alpha_i = 0, \quad 0 \le \alpha_i \le C, \quad i = 1, \dots, N.$$

Notice $\langle h(X_i H_V), h(X_j H_V) \rangle = K(X_i H_V, X_j H_V)$.

3 Limits

We have to find a criterion of the existence of h for a given kernel function as SVM kernel method shows that the positive definite kernel has feature mapping h. But it might be a harder problem than finding a tractable algorithm for (2).