Compatibility between kernel SMM and SVM

Miaoyan Wang, May 18, 2020

Theorem 0.1 (Compatibility with classical SVM). *Let $X \in \mathbb{R}^d$ denote a column vector and $X^* = \begin{bmatrix} 0 & X^T \\ X & 0 \end{bmatrix} \in \mathbb{R}^{(d+1)\times(d+1)}$ be the symmetrized matrix. Then, under our proposed column-wise kernels, the SMM classifier trained on X^* equals the SVM classifier trained on X.*

Remark 1. The decision boundaries generated by two classifiers are the same, but the optimal objective value may not be the same.

Proof. We prove the following stronger result.

(**Set-up**.) Let $X \in \mathbb{R}^{m\times n}$ be the original matrix feature,

$$X = \begin{bmatrix} | & | & | & | \\ x_1 & x_2 & \vdots & x_n \\ | & | & | & | \end{bmatrix} = \begin{bmatrix} — & y_1^T & — \\ — & y_2^T & — \\ — & \cdots & — \\ — & y_m^T & — \end{bmatrix},$$

where x_i and y_j^T denotes the i-th column and j-th row of the matrix X, respectively. Let $X^* \in \mathbb{R}^{(m+n)\times(m+n)}$ be the symmetrized matrix feature. Based on the assumption (of sufficiently valid kernels), the feature mapping h is applied to matrix X^* in a column-wise fashion; that is,

$$X^* = \begin{bmatrix} 0 & X^T \\ X & 0 \end{bmatrix} \Rightarrow h(X^*) = \left[\begin{array}{ccc:ccc} & & & | & | & | \\ & 0 & & h(y_1^T) & \ldots & h(y_m^T) \\ & & & | & | & | \\ \hdashline | & | & | & & & \\ h(x_1) & \ldots & h(x_n) & & 0 & \\ | & | & | & & & \end{array}\right] =: \begin{bmatrix} 0 & A \\ C & 0 \end{bmatrix}, \quad (1)$$

where, without a little abuse of notation, we have used $h(x_i)$ and $h(y_i^T)$ to denote the non-zero coordinates in the image vector $h((0,\ldots,0,x_i^T))$ and $h((y_j^T,0,\ldots,0)$, respectively.

(**Conclusion.**) Assume the set of elements in C contains all elements in A. Then, the classifier trained on X equals to the classifier trained on X^*.

(**Proof of the above conclusion.**). Because the elements in C contains all elements in A, we re-arange the elements in C into a possibly larger matrix:

$$C = \begin{bmatrix} 0 & A^T \\ D & 0 \end{bmatrix}. \quad (2)$$

Such reshaping is possible by, for example, setting $D = \mathrm{diag}(\{C\}/\{A\})$. (The choice of D is non-important.) To prove the equivalence between X- and X^*-trained classifiers, it suffices to prove the equivalence between

$$h(X^*) = \begin{bmatrix} 0 & 0 & A \\ 0 & A^T & 0 \\ D & 0 & 0 \end{bmatrix}, \quad \text{and} \quad h(X) = \begin{bmatrix} 0 & A^T \\ D & 0 \end{bmatrix} \quad \text{(c.f. (1) and (2))}.$$

Now we apply the same proof techniques as in 051820_SMMK_modification.pdf. Specifically, consider the primal problems with $h(X^*)$ and $h(X)$. The relevant linear predictors are

$$\langle B^*,\ h(X^*)\rangle = \left\langle \begin{bmatrix} 0 & 0 & B_3^* \\ 0 & B_2^* & 0 \\ B_1^* & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & A \\ 0 & A^T & 0 \\ D & 0 & 0 \end{bmatrix} \right\rangle, \quad \text{and} \quad \langle B,\ X\rangle = \left\langle \begin{bmatrix} 0 & B_2 \\ B_1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & A^T \\ D & 0 \end{bmatrix} \right\rangle.$$

The optimizations are

$$\begin{aligned} (P1) \quad & \min_{B_1^*, B_2^*, B_3^*} \|B_1^*\|_F^2 + \|B_2^*\|_F^2 + \|B_3^*\|_F^2 + C\sum_i \xi_i, \\ & \text{s.t. .... where } \langle B^*,\ h(X^*)\rangle = \langle B_1^*, D\rangle + \langle B_2^*, A^T\rangle + \langle B_3^{*T}, A\rangle \\ & \qquad\qquad\qquad\qquad\qquad\quad = \langle B_1^*, D\rangle + \langle B_2^{*T} + B_3^*, A\rangle, \end{aligned}$$

and

$$\begin{aligned} (P2) \quad & \min_{B_1, B_2} \|B_1\|_F^2 + \|B_2\|_F^2 + C\sum_i \xi_i, \\ & \text{s.t. ... where } \langle B,\ h(X)\rangle = \langle B_1, D\rangle + \langle B_2, A\rangle. \end{aligned}$$

It is easy to see that the optimal solution to (P1) is achieved at $B_2^* = B_3^{*T}$ because

$$\|B_2^*\|_F^2 + \|B_3^*\|_F^2 \geq \frac{1}{2}\|B_2^* + B_3^*\|_F = \frac{1}{2}\|B_2\|_F^2,$$

where the equality holds when $B_2^{*T} = B_3^*$. Moreover, under this choice, the optimal objective values

$$(P1) \ \min_{B_1^*, B_2^*} \|B_1\|_F + \frac{1}{2}\|B_2\|_F^2 + C\sum_i \xi_i, \quad \text{(P2)} \ \min_{B_1, B_2} \|B_1\|_F + \|B_2\|_F^2 + C\sum_i \xi_i,$$

are not necessarily equal unless $B_1 = 0$. $\square$