Connection between kernel SMM and SVM

Miaoyan Wang, May 18, 2020

Fact: 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 lifted matrix. Under general nonlinear kernels, we cannot guarantee the equal decision boundaries between X-trained and X^*-trained SMMs!

Where goes wrong? Fact: Repeated attributes are down-weighted in SVM.

Consider two SVMs, one trained on $X = (x_1, x_2)^T$, and another one trained on $X^* = (x_1, x_2, x_2)^T$. It turns out the decision boundary trained by X and X^* are different.

Reason: The primal problems for these two SVMs are

$$(P1) \qquad \min_{\beta_1,\beta_2} \left\{ \beta_1^2 + \beta_2^2 + C \sum_i \left[y_i(\beta_1 x_1 + \beta_2 x_2) \right]_+ \right\},$$

and

$$\begin{aligned} (P2) \qquad & \min_{\beta_1,\beta_2,\beta_3} \left\{ \beta_1^2 + \beta_2^2 + \beta_3^2 + C' sum_i \left[y_i(\beta_1 x_1 + (\beta_2+\beta_3) x_2) \right]_+ \right\} \\ &= \min_{\beta_1,\beta_2} \left\{ \beta_1^2 + \frac{1}{2}\beta_2^2 + C' \sum_i \left[y_i(\beta_1 x_1 + \beta_2 x_2) \right]_+ \right\}. \end{aligned}$$

The solutions to (P1) and (P2) are usually different unless $\beta_1 = 0$. In particular, the repeated attribute, x_2, is down-weighted in the cost function (P2).

Back to the kernel SMM problem. Under general nonlinear kernels, the attributes have different occurrences in $h(X^*)$ and $h(X)$. Therefore, the two decision boundaries might be different. Note that the two classifiers agree in the special case of linear kernels, since each attribute repeats precisely twice.

Implication: A weighted SVM

$$\min_{\boldsymbol{\beta}} \boldsymbol{\beta}^T [\mathrm{Cov}(X)]^{-1} \boldsymbol{\beta} + \sum_i [y_i \langle \boldsymbol{\beta},\ X \rangle]_+$$

will stabilize the contribution from repeated attributes, thereby maintaining the equivalence between X- and X^*-trained SVMs. In our context of kernel SMM, however, the unequal occurrences of attributes are actually okay. We can interpret the unequal occurrences as a prior knowledge of "importance" in the classifier.