Approximate Leave-One-Out with kernel SVM

Linyun He Wanchao Qin Peng Xu Yuze Zhou

September 11, 2018

1 ALO with Representer Theorem

Let $\boldsymbol{K}$ denote the positive-definite kernel matrix (hence invertible), with $\boldsymbol{K}_{i,j} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$. By the Representer Theorem, the dual problem to kernel SVM can be expressed in "loss + penalty" form:

$$\min_{\rho,\boldsymbol{\alpha}} \sum_{j=1}^{n} \left(1 - y_j f(x_j)\right)_+ + \frac{\lambda}{2}\boldsymbol{\alpha}^\top \boldsymbol{K}\boldsymbol{\alpha}, \qquad f(x_j) = \boldsymbol{K}_{\cdot,j}^\top \boldsymbol{\alpha} + \rho. \tag{1}$$

For simplicity we ignore the offset for now. Let S and V be the smooth set and the singularities set, respectively. For the j-th observation, $j \in S$ and that $f(x_j) < 1$, we have

$$\dot{\ell}(\boldsymbol{K}_{\cdot,j}^\top \boldsymbol{\alpha}) = -y_j, \qquad \ddot{\ell}(\boldsymbol{K}_{\cdot,j}^\top \boldsymbol{\alpha}) = 0.$$

Additionally,

$$\nabla R(\boldsymbol{\alpha}) = \lambda \boldsymbol{K}\boldsymbol{\alpha}, \qquad \nabla^2 R(\boldsymbol{\alpha}) = \lambda \boldsymbol{K}.$$

Substitute corresponding terms in Thm. 4.1, we deduce the ALO formula for kernel SVM:

$$\boldsymbol{K}_{\cdot,i}^\top \tilde{\boldsymbol{\alpha}}^{\setminus i} = \boldsymbol{K}_{\cdot,i}^\top \hat{\boldsymbol{\alpha}} + a_i g_{\ell,i},$$

where

$$a_i = \begin{cases} \dfrac{1}{\lambda}\boldsymbol{K}_{\cdot,i}^\top \left[\boldsymbol{K}^{-1} - \boldsymbol{K}^{-1}\boldsymbol{K}_{\cdot,V}\left(\boldsymbol{K}_{\cdot,V}^\top \boldsymbol{K}^{-1}\boldsymbol{K}_{\cdot,V}\right)^{-1} \boldsymbol{K}_{\cdot,V}^\top \boldsymbol{K}^{-1}\right] \boldsymbol{K}_{\cdot,i} & i \in S, \\ \left[\lambda \left(\boldsymbol{K}_{\cdot,V}^\top \boldsymbol{K}^{-1}\boldsymbol{K}_{\cdot,V}\right)_{ii}^{-1}\right]^{-1} & i \in V, \end{cases}$$

and

$$g_{\ell,S} = -y_S \odot \mathbf{1}\left\{y_S \boldsymbol{K}_{\cdot,S}^\top \boldsymbol{\alpha} < 1\right\}, \qquad g_{\ell,V} = \left(\boldsymbol{K}_{\cdot,V}\boldsymbol{K}_{\cdot,V}^\top\right)^{-1} \boldsymbol{K}_{\cdot,V}\left[\lambda \boldsymbol{K}\alpha - \sum_{j:y_j \boldsymbol{K}_{\cdot,j}^\top \boldsymbol{\alpha} < 1} y_j \boldsymbol{K}_{\cdot,j}\right].$$

2 ALO with Approximate Explicit Feature Maps

In non-linear SVM, kernel trick is employed to avoid the explicit computation of feature maps, which sometime is impossible since the feature space can be infinite-dimensional. However, when sample size n is large, the kernel matrices become quite expensive to handle. Methods such as the Nyström approximation are used in order to retain the benefit of features mapping whilst retaining the speed of linear SVM. We may adopt a similar idea to help the ALO computation.

Let $\boldsymbol{X}$ be the data matrix and $\boldsymbol{K}$ be the corresponding kernel matrix. An approximation $\hat{\boldsymbol{\Phi}}$ to the feature maps $\boldsymbol{\Phi}(\boldsymbol{X})$ can be constructed as following (procedure adopted from `scikit-learn`):

1. Perform SVD: $\boldsymbol{K} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$;
2. Clamp the singular values: $\tilde{\boldsymbol{S}} = \max(\boldsymbol{S}, 10^{-12})$;
3. Construct the approximate map as $\hat{\boldsymbol{\Phi}} = \boldsymbol{K}\boldsymbol{U}\tilde{\boldsymbol{S}}^{-1/2}\boldsymbol{V}^\top \approx \boldsymbol{K}^{1/2}$.

To compute ALO, we then simply replace the data matrix $\boldsymbol{X}$ with $\hat{\boldsymbol{\Phi}}$ in the linear SVM formula.