Approximate Leave-One-Out with kernel SVM

Peng Xu

February 18, 2019

1 ALO with the Variational Problems

1.1 C-Support Vector Classification

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 SVC can be expressed in "loss + penalty" variational form:

$$\min_{\rho,\boldsymbol{\alpha}} \sum_{j=1}^{n} \max\left[0, 1 - y_j f(x_j)\right] + \frac{\lambda}{2}\boldsymbol{\alpha}^\top \boldsymbol{K}\boldsymbol{\alpha}, \qquad f(x_j) = \boldsymbol{K}_{\cdot,j}^\top \boldsymbol{\alpha} + \rho. \tag{1}$$

For simplicity we ignore the offset ρ for now. Let S and V be the smooth set and the set of singularities, respectively. For the j-th observation, $j \in S$, we have

$$\dot{\ell}(\boldsymbol{K}_{\cdot,j}^\top \boldsymbol{\alpha}) = -y_j \cdot \mathbf{1}\{y_j \boldsymbol{K}_{\cdot,j}^\top \boldsymbol{\alpha} < 1\}, \qquad \ddot{\ell}(\boldsymbol{K}_{\cdot,j}^\top \boldsymbol{\alpha}) = 0.$$

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 SVC:

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

where

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

and

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

1.2 ε-Support Vector Regression

The ε-SVR is associated with the ε-insensitive loss function. Its objective function can be written as:

$$\min_{\rho,\boldsymbol{\alpha}} \sum_{j=1}^{n} \max\left[0, |y_j - f(x_j)| - \varepsilon\right] + \frac{\lambda}{2}\boldsymbol{\alpha}^\top \boldsymbol{K}\boldsymbol{\alpha}, \qquad f(x_j) = \boldsymbol{K}_{\cdot,j}^\top \boldsymbol{\alpha} + \rho. \tag{2}$$

For the j-th observation, $j \in S$, we now have

$$\dot{\ell}(\boldsymbol{K}_{\cdot,j}^\top \boldsymbol{\alpha}) = -\operatorname{sgn}\left(\boldsymbol{K}_{\cdot,j}^\top \boldsymbol{\alpha}\right) \cdot \mathbf{1}\left\{\left|y_j - \boldsymbol{K}_{\cdot,j}^\top \boldsymbol{\alpha}\right| \geq \varepsilon\right\}, \qquad \ddot{\ell}(\boldsymbol{K}_{\cdot,j}^\top \boldsymbol{\alpha}) = 0.$$

Thus, our ALO recipe will be exactly the same as in C-SVC, except

$$g_{\ell,S} = -\operatorname{sgn}\left(\boldsymbol{K}_{\cdot,S}^\top \boldsymbol{\alpha}\right) \odot \mathbf{1}\left\{\left|y_j - \boldsymbol{K}_{\cdot,S}^\top \boldsymbol{\alpha}\right| \geq \varepsilon\right\},$$

and

$$g_{\ell,V} = \left(\boldsymbol{K}_{\cdot,V}^\top \boldsymbol{K}_{\cdot,V}\right)^{-1} \boldsymbol{K}_{\cdot,V}^\top \left[\sum_{j\in S:\left|y_j - \boldsymbol{K}_{\cdot,j}^\top \boldsymbol{\alpha}\right| \geq \varepsilon} \operatorname{sgn}\left(\boldsymbol{K}_{\cdot,j}^\top \boldsymbol{\alpha}\right) \boldsymbol{K}_{\cdot,j} - \lambda \boldsymbol{K}\alpha\right].$$

1.3 ν-SVC and ν-SVR

ν-SVC and ν-SVR are introduced in [Sch+00], with $\nu \in (0, 1]$ a new parameter that can control the number of support vectors. Notably, ν replaces C in C-SVC and ε in ε-SVR, and in latter case ε itself becomes an optimization argument.

Because of the following equivalencies, there is no need to figure out the variational problem in order to derive ALO for ν-SVC and ν-SVR:

- Let α^* be the dual solution of ν-SVC with parameters (C, ν), then α^*/ρ^* is an optimal solution of C-SVC with $C = 1/(n\rho^*)$. See [CL01].
- Let α^* be the dual solution of ν-SVR with parameters (C, ν), then α^* is also the solution of ε-SVR with parameters $(C/n, \varepsilon^*)$. See [CL02].

Here ρ^* and ε^* are the corresponding primal solutions, and can be worked out quickly from α^* through equations derived from KKT condition. See [CL11, Section 4.2].

References

[CL01] Chih-Chung Chang and Chih-Jen Lin. "Training v-Support Vector Classifiers: Theory and Algorithms". In: *Neural Computation* 13.9 (2001), pp. 2119–2147. DOI: 10.1162/089976601750399335.

[CL02] Chih-Chung Chang and Chih-Jen Lin. "Training v-Support Vector Regression: Theory and Algorithms". In: *Neural Computation* 14.8 (2002), pp. 1959–1977. DOI: 10.1162/089976602760128081.

[CL11] Chih-Chung Chang and Chih-Jen Lin. "LIBSVM: A Library for Support Vector Machines". In: *ACM Trans. Intell. Syst. Technol.* 2.3 (May 2011), 27:1–27:27. ISSN: 2157-6904. DOI: 10.1145/1961189.1961199. URL: https://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf.

[Sch+00] Bernhard Schölkopf et al. "New Support Vector Algorithms". In: *Neural Computation* 12.5 (2000), pp. 1207–1245. DOI: 10.1162/089976600300015565.