Complexity of the PSVM

Sepp Hochreiter

Institute of Bioinformatics,
Johannes Kepler Universität Linz
4040 Linz, Austria

hochreit@bioinf.jku.at

$$\boldsymbol{w}^\top \boldsymbol{X} \ \boldsymbol{X}^\top \boldsymbol{w} \ = \ \left\|\boldsymbol{X}^\top \boldsymbol{w}\right\|^2 \tag{1}$$

$$\boldsymbol{K} \ = \ \boldsymbol{X}^\top \ \boldsymbol{Z} \ . \tag{2}$$

The data vectors $(K_{1j}, K_{2j}, \ldots, K_{Lj})$ are normalized to zero mean and unit variance:

$$\frac{1}{L}\sum_{i=1}^{L} \left(K_{ij} \ - \ \bar{K}_j\right)^2 \ = \ 1 \quad \text{and} \quad \bar{K}_j \ = \ \frac{1}{L} \sum_{i=1}^{L} K_{ij} \ = \ 0 \ , \tag{3}$$

$$\boldsymbol{K}^\top \ \mathbf{1} \ = \ \mathbf{0} \quad \text{and} \quad \mathrm{diag}(\boldsymbol{K}^\top \boldsymbol{K}) \ = \ L \ \mathbf{1}. \tag{4}$$

The labels are normalized to zero mean:

$$b \ = \ \frac{1}{L} \ \sum_{i=1}^{L} y_i \ = \ 0 \ . \tag{5}$$

If the P-SVM is used for classification. Primal:

$$\begin{aligned} \min_{\boldsymbol{w}, \boldsymbol{\xi}^+, \boldsymbol{\xi}^-} \quad & \frac{1}{2} \ \|\boldsymbol{X}^\top \ \boldsymbol{w}\|^2 \ + \ C \mathbf{1}^\top \left(\boldsymbol{\xi}^+ \ + \ \boldsymbol{\xi}^-\right) \\ \text{s.t.} \quad & \boldsymbol{K}^\top \ \left(\boldsymbol{X}^\top \ \boldsymbol{w} \ - \ \boldsymbol{y}\right) \ + \ \boldsymbol{\xi}^+ \ \geq \ \mathbf{0} \\ & \boldsymbol{K}^\top \ \left(\boldsymbol{X}^\top \ \boldsymbol{w} \ - \ \boldsymbol{y}\right) \ - \ \boldsymbol{\xi}^- \ \leq \ \mathbf{0} \\ & \mathbf{0} \ \leq \ \boldsymbol{\xi}^+, \boldsymbol{\xi}^- \end{aligned} \tag{6}$$

Lagrangian L:

$$L = \frac{1}{2}\, \boldsymbol{w}^\top \boldsymbol{X}\, \boldsymbol{X}^\top \boldsymbol{w} + C\boldsymbol{1}^\top \left(\boldsymbol{\xi}^+ + \boldsymbol{\xi}^-\right) \tag{7}$$

$$- \left(\boldsymbol{\alpha}^+\right)^\top \left(\boldsymbol{K}^\top \left(\boldsymbol{X}^\top \boldsymbol{w} - \boldsymbol{y}\right) + \boldsymbol{\xi}^+\right)$$

$$+ \left(\boldsymbol{\alpha}^-\right)^\top \left(\boldsymbol{K}^\top \left(\boldsymbol{X}^\top \boldsymbol{w} - \boldsymbol{y}\right) - \boldsymbol{\xi}^-\right)$$

$$- \left(\boldsymbol{\mu}^+\right)^\top \boldsymbol{\xi}^+ - \left(\boldsymbol{\mu}^-\right)^\top \boldsymbol{\xi}^- \ . \tag{8}$$

Dual:

$$\min_{\boldsymbol{\alpha}} \quad \frac{1}{2}\boldsymbol{\alpha}^\top \boldsymbol{K}^\top \boldsymbol{K}\, \boldsymbol{\alpha} - \boldsymbol{y}^\top \boldsymbol{K}\, \boldsymbol{\alpha} \tag{9}$$

$$\text{s.t.} \quad -C\, \boldsymbol{1} \leq \boldsymbol{\alpha} \leq C\boldsymbol{1} \ ,$$

We know that

$$\boldsymbol{w} = \boldsymbol{Z}\, \boldsymbol{\alpha} \ . \tag{10}$$

P-SVM feature selection primal optimization problem:

$$\min_{\boldsymbol{w}} \quad \frac{1}{2} \left\|\boldsymbol{X}^\top \boldsymbol{w}\right\|^2 \tag{11}$$

$$\text{s.t.} \quad \boldsymbol{K}^\top \left(\boldsymbol{X}^\top \boldsymbol{w} - \boldsymbol{y}\right) + \epsilon\, \boldsymbol{1} \geq \boldsymbol{0}$$

$$\boldsymbol{K}^\top \left(\boldsymbol{X}^\top \boldsymbol{w} - \boldsymbol{y}\right) - \epsilon\, \boldsymbol{1} \leq \boldsymbol{0} \ .$$

Lagrangian:

$$L = \frac{1}{2}\, \boldsymbol{w}^\top \boldsymbol{X}\, \boldsymbol{X}^\top \boldsymbol{w} \tag{12}$$

$$- \left(\boldsymbol{\alpha}^+\right)^\top \left(\boldsymbol{K}^\top \left(\boldsymbol{X}^\top \boldsymbol{w} - \boldsymbol{y}\right) + \epsilon\, \boldsymbol{1}\right)$$

$$+ \left(\boldsymbol{\alpha}^-\right)^\top \left(\boldsymbol{K}^\top \left(\boldsymbol{X}^\top \boldsymbol{w} - \boldsymbol{y}\right) - \epsilon\, \boldsymbol{1}\right) \ ,$$

$$\boldsymbol{w} = \boldsymbol{Z}\, \boldsymbol{\alpha} \ , \tag{13}$$

Dual:

$$\min_{\boldsymbol{\alpha}^+, \boldsymbol{\alpha}^-} \quad \frac{1}{2} \left(\boldsymbol{\alpha}^+ - \boldsymbol{\alpha}^-\right)^\top \boldsymbol{K}^\top \boldsymbol{K} \left(\boldsymbol{\alpha}^+ - \boldsymbol{\alpha}^-\right) \tag{14}$$

$$- \boldsymbol{y}^\top \boldsymbol{K} \left(\boldsymbol{\alpha}^+ - \boldsymbol{\alpha}^-\right) + \epsilon\, \boldsymbol{1}^\top \left(\boldsymbol{\alpha}^+ + \boldsymbol{\alpha}^-\right)$$

$$\text{s.t.} \quad \boldsymbol{0} \leq \boldsymbol{\alpha}^+ \ , \quad \boldsymbol{0} \leq \boldsymbol{\alpha}^- \ .$$

$$\epsilon\, \boldsymbol{1}^\top \left(\boldsymbol{\alpha}^+ + \boldsymbol{\alpha}^-\right) \tag{15}$$

$$\boldsymbol{\alpha} = \left(\boldsymbol{\alpha}^+ - \boldsymbol{\alpha}^-\right) \tag{16}$$

The Karush-Kuhn-Tucker conditions require:

$$\left(\boldsymbol{\alpha}^{+}\right)^{\top} \left(\boldsymbol{K}^{\top}\left(\boldsymbol{X}^{\top} \boldsymbol{w} - \boldsymbol{y}\right) + \boldsymbol{\xi}^{+}\right) = 0 \tag{17}$$
$$\left(\boldsymbol{\alpha}^{-}\right)^{\top} \left(\boldsymbol{K}^{\top}\left(\boldsymbol{X}^{\top} \boldsymbol{w} - \boldsymbol{y}\right) - \boldsymbol{\xi}^{-}\right) = 0 \tag{18}$$

or

$$\left(\boldsymbol{\alpha}^{+}\right)^{\top} \left(\boldsymbol{K}^{\top}\left(\boldsymbol{X}^{\top} \boldsymbol{w} - \boldsymbol{y}\right) + \epsilon\, \mathbf{1}\right) = 0 \tag{19}$$
$$\left(\boldsymbol{\alpha}^{-}\right)^{\top} \left(\boldsymbol{K}^{\top}\left(\boldsymbol{X}^{\top} \boldsymbol{w} - \boldsymbol{y}\right) - \epsilon\, \mathbf{1}\right) = 0 \tag{20}$$

We know that

$$\frac{\partial L}{\partial \xi_i^{+}} = C - \alpha_i^{+} - \mu_i^{+} = 0 \tag{21}$$
$$\frac{\partial L}{\partial \xi_i^{-}} = C - \alpha_i^{-} - \mu_i^{-} = 0 \tag{22}$$
$$\tag{23}$$

KKT conditions require

$$\xi_i^{+} \mu_i^{+} = 0 \tag{24}$$
$$\xi_i^{-} \mu_i^{-} = 0 \tag{25}$$
$$\tag{26}$$

Optimality requires

$$\alpha_i^{+} \alpha_i^{-} = 0 \tag{27}$$

because otherwise the term $\epsilon\, \mathbf{1}^{\top} \left(\boldsymbol{\alpha}^{+} + \boldsymbol{\alpha}^{-}\right)$ can be decreased without changing the other terms in the objective.

$$\alpha_i^{+} = 0 \Longrightarrow \mu_i^{+} = C \Longrightarrow \xi_i^{+} = 0 \tag{28}$$
$$\alpha_i^{-} = 0 \Longrightarrow \mu_i^{-} = C \Longrightarrow \xi_i^{-} = 0 \tag{29}$$
$$\tag{30}$$

Therefore

$$\alpha_i^{+} > 0 \Longrightarrow \alpha_i^{-} = 0 \quad \text{and} \quad \xi_i^{-} = 0 \tag{31}$$
$$\alpha_i^{-} > 0 \Longrightarrow \alpha_i^{+} = 0 \quad \text{and} \quad \xi_i^{+} = 0 \tag{32}$$

It follows that

$$\left(\alpha_i^{+} + \alpha_i^{-}\right)\left(\xi_i^{+} + \xi_i^{-}\right) = \alpha_i^{+}\xi_i^{+} + \alpha_i^{-}\xi_i^{-} \tag{33}$$

$$
\begin{aligned}
&\boldsymbol{w}^\top \boldsymbol{X}\ \boldsymbol{X}^\top \boldsymbol{w}\ = \\
&\boldsymbol{\alpha}^\top \boldsymbol{Z}^\top \boldsymbol{X}\ \boldsymbol{X}^\top \boldsymbol{w}\ = \\
&\boldsymbol{\alpha}^\top \boldsymbol{K}^\top \boldsymbol{X}^\top \boldsymbol{w}\ = \\
&\boldsymbol{\alpha}^\top \boldsymbol{K}^\top \boldsymbol{X}^\top \boldsymbol{w}\ -\ \boldsymbol{\alpha}^\top \boldsymbol{K}^\top \boldsymbol{y}\ +\ \boldsymbol{\alpha}^\top \boldsymbol{K}^\top \boldsymbol{y} \\
&-\ \left(\boldsymbol{\alpha}^+\right)^\top \boldsymbol{\xi}^+\ +\ \left(\boldsymbol{\alpha}^+\right)^\top \boldsymbol{\xi}^+ -\ \left(\boldsymbol{\alpha}^-\right)^\top \boldsymbol{\xi}^-\ +\ \left(\boldsymbol{\alpha}^-\right)^\top \boldsymbol{\xi}^-\ = \\
&\left(\boldsymbol{\alpha}^+\right)^\top\ \left(\boldsymbol{K}^\top \left(\boldsymbol{X}^\top\ \boldsymbol{w}\ -\ \boldsymbol{y}\right)\ +\ \boldsymbol{\xi}^+\right)\ + \\
&\left(\boldsymbol{\alpha}^-\right)^\top\ \left(\boldsymbol{K}^\top \left(\boldsymbol{X}^\top\ \boldsymbol{w}\ -\ \boldsymbol{y}\right)\ -\ \boldsymbol{\xi}^-\right)\ + \\
&\boldsymbol{\alpha}^\top \boldsymbol{K}^\top \boldsymbol{y}\ +\ \left(\boldsymbol{\alpha}^+\right)^\top \boldsymbol{\xi}^+\ +\ \left(\boldsymbol{\alpha}^-\right)^\top \boldsymbol{\xi}^-\ = \\
&\boldsymbol{\alpha}^\top \boldsymbol{K}^\top \boldsymbol{y}\ +\ \left(\boldsymbol{\alpha}^+\right)^\top \boldsymbol{\xi}^+\ +\ \left(\boldsymbol{\alpha}^-\right)^\top \boldsymbol{\xi}^-\ = \\
&\boldsymbol{\alpha}^\top \boldsymbol{K}^\top \boldsymbol{y}\ +\ \left(\boldsymbol{\alpha}^+\ +\ \boldsymbol{\alpha}^-\right)^\top \left(\boldsymbol{\xi}^+\ +\ \boldsymbol{\xi}^-\right)\ = \\
&\boldsymbol{\alpha}^\top \boldsymbol{K}^\top \boldsymbol{y}\ +\ |\boldsymbol{\alpha}|^\top \boldsymbol{\xi}\ ,
\end{aligned} \tag{34}
$$

where

$$|\boldsymbol{\alpha}|\ =\ \boldsymbol{\alpha}^+\ +\ \boldsymbol{\alpha}^- \tag{35}$$

$$\boldsymbol{\xi}\ =\ \boldsymbol{\xi}^+\ +\ \boldsymbol{\xi}^- \tag{36}$$

We have for the optimal values:

$$\boldsymbol{w}^\top \boldsymbol{X}\ \boldsymbol{X}^\top \boldsymbol{w}\ =\ \boldsymbol{\alpha}^\top \boldsymbol{K}^\top \boldsymbol{y}\ +\ |\boldsymbol{\alpha}|^\top \boldsymbol{\xi} \tag{37}$$

Similarly it follows for feature selection

$$\boldsymbol{w}^\top \boldsymbol{X}\ \boldsymbol{X}^\top \boldsymbol{w}\ =\ \boldsymbol{\alpha}^\top \boldsymbol{K}^\top \boldsymbol{y}\ +\ \epsilon\ |\boldsymbol{\alpha}|^\top \mathbf{1}\ . \tag{38}$$

$$\mathrm{var}(\boldsymbol{K})\ =\ \frac{1}{L}\ \mathrm{diag}(\boldsymbol{K}^\top \boldsymbol{K})\ =\ \mathbf{1} \tag{39}$$

$$\mathrm{covar}(\boldsymbol{K}, \boldsymbol{y})\ =\ \frac{1}{L} \boldsymbol{K}^\top \boldsymbol{y} \tag{40}$$

$$\mathrm{var}(\boldsymbol{y})\ =\ \frac{1}{L} \boldsymbol{y}^\top \boldsymbol{y}\ =\ \frac{1}{L} \|\boldsymbol{y}\|^2 \tag{41}$$

$$\mathrm{corel}(\boldsymbol{K}, \boldsymbol{y})\ =\ \frac{\mathrm{covar}(\boldsymbol{K}, \boldsymbol{y})}{\sqrt{\mathrm{var}(\boldsymbol{K})}\ \sqrt{\mathrm{var}(\boldsymbol{y})}}\ =\ \frac{\boldsymbol{K}^\top \boldsymbol{y}}{\sqrt{L}\ \|\boldsymbol{y}\|} \tag{42}$$

We have

$$\boldsymbol{K}^\top \boldsymbol{y}\ =\ \sqrt{L}\ \|\boldsymbol{y}\|\ \mathrm{corel}(\boldsymbol{K}, \boldsymbol{y}) \tag{43}$$

and

$$\boldsymbol{\alpha}^\top \boldsymbol{K}^\top \boldsymbol{y}\ =\ \sqrt{L}\ \|\boldsymbol{y}\|\ \boldsymbol{\alpha}^\top \mathrm{corel}(\boldsymbol{K}, \boldsymbol{y}) \tag{44}$$

For positive correlation coefficient the α is positive, otherwise it is negative, so that

$$\boldsymbol{\alpha}^\top \boldsymbol{K}^\top \boldsymbol{y} \;=\; \sqrt{L}\; \|\boldsymbol{y}\|\; |\boldsymbol{\alpha}|^\top |\mathrm{corel}(\boldsymbol{K}, \boldsymbol{y})| \tag{45}$$

$$\boldsymbol{w}^\top \boldsymbol{X}\; \boldsymbol{X}^\top \boldsymbol{w} \;=\; |\boldsymbol{\alpha}|^\top \left(\sqrt{L}\; \|\boldsymbol{y}\|\; |\mathrm{corel}(\boldsymbol{K}, \boldsymbol{y})| \;+\; \boldsymbol{\xi} \right) \tag{46}$$

and for feature selection

$$\boldsymbol{w}^\top \boldsymbol{X}\; \boldsymbol{X}^\top \boldsymbol{w} \;=\; |\boldsymbol{\alpha}|^\top \left(\sqrt{L}\; \|\boldsymbol{y}\|\; |\mathrm{corel}(\boldsymbol{K}, \boldsymbol{y})| \;+\; \epsilon\; \mathbf{1} \right) \tag{47}$$

The complexity depends on $\|\boldsymbol{y}\|$ which can be used to bring the model into a canonical form, like scaling y until we have the canonical form of the classical C-SVM.

Then the complexity is the values $|\boldsymbol{\alpha}|$ weighted by correlation coefficient between the features given by $\boldsymbol{K}$ and the labels/targets $\boldsymbol{y}$ and the error given by ϵ or $\boldsymbol{\xi}$.

The complexity in the classical SVM is $\boldsymbol{\alpha}^\top\; \mathbf{1} \;=\; |\boldsymbol{\alpha}|^\top\; \mathbf{1}$ where all components of $\boldsymbol{\alpha}$ are equally weighted.