Algorithmic perspectives of kernel method

Chanwoo Lee, August 7, 2020

1 A choice of feature mapping Φ

To derive an algorithm, I choose to use Mapping 1 in the previous note for convenience.

$$\begin{aligned}
\Phi\colon \mathbb{R}^{d_1\times d_2} &\to \mathcal{H}_r^{d_1}\times\mathcal{H}_c^{d_2}\\
\boldsymbol{X} &\mapsto (\Phi_r(\boldsymbol{X})\Phi_c(\boldsymbol{X})) \stackrel{\text{def}}{=} (\phi_r([\boldsymbol{X}]_{1:}),\ldots,\phi_r([\boldsymbol{X}]_{d_1:}),\phi_c([\boldsymbol{X}]_{:1})\ldots,\phi([\boldsymbol{X}]_{:d_2})).
\end{aligned}$$

We define decision function,

$$\begin{aligned}
f(\boldsymbol{X}) &= \langle \boldsymbol{B},\Phi(\boldsymbol{X})\rangle,\text{ where } \boldsymbol{B}=(\boldsymbol{B}_r,\boldsymbol{B}_c)\in\mathcal{H}_r^{d_1}\times\mathcal{H}_c^{d_2} \qquad (1)\\
&= \langle \boldsymbol{B}_r,\Phi_r(\boldsymbol{X})\rangle + \langle \boldsymbol{B}_c,\Phi_c(\boldsymbol{X})\rangle\\
&= \sum_{k=1}^n \gamma_k\left(\sum_{i,j\in[d_2]} w_{ij}^{row}K([\boldsymbol{X}_k]_{i\cdot},[\boldsymbol{X}]_{j\cdot}) + \sum_{i,j\in[d_2]} w_{ij}^{col}K([\boldsymbol{X}_k]_{\cdot i},[\boldsymbol{X}]_{\cdot j})\right),
\end{aligned}$$

where $\boldsymbol{X}^1,\ldots\boldsymbol{X}^n$ are sampled matrix features and $\boldsymbol{W}^{\text{col}},\boldsymbol{W}^{\text{row}}$ are some positive semi definite matrices with low rank. We estimate $\boldsymbol{W}^{\text{col}}$, $\boldsymbol{W}^{\text{row}}$, and $\boldsymbol{\gamma}=(\gamma_1,\ldots,\gamma_n)$ from the training data set.

2 Algorithm derivation

We solve an optimization problem

$$\begin{aligned}
&\min_{\boldsymbol{B}} \frac{1}{2}\|\boldsymbol{B}\|_F^2 + c\sum_{i=1}^n \xi_i, \qquad (2)\\
&\text{subject to } y_i\langle \boldsymbol{B},\Phi(\boldsymbol{X}_i)\rangle \le 1-\xi_i \text{ and } \xi_i\ge 0, i=1,\ldots,n.
\end{aligned}$$

where $\|\boldsymbol{B}\|_F^2 = \|\boldsymbol{B}_r\|_F^2 + \|\boldsymbol{B}_c\|_F^2$. From the low rank assumption on $\boldsymbol{B}$ such that

$$\boldsymbol{B} = (\boldsymbol{B}_r,\boldsymbol{B}_c) = \boldsymbol{C}\boldsymbol{P}^T = (\boldsymbol{C}_r,\boldsymbol{C}_c)(\boldsymbol{P}_r,\boldsymbol{P}_c)^T,$$

where $\boldsymbol{C}=(\boldsymbol{C}_r,\boldsymbol{C}_c)\in\mathcal{H}_r^r\times\mathcal{H}_c^r$ and $\boldsymbol{P}=(\boldsymbol{P}_r,\boldsymbol{P}_c)\in\mathbb{R}^{d_1\times r}\times\mathbb{R}^{d_2\times r}$.

1. First we update $\boldsymbol{C}$ holding $\boldsymbol{P}$ fixed. The dual problem of Equation (2) is

$$\begin{aligned}
&\min_{\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_n)} -\sum_{i=1}^n\alpha_i + \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n \alpha_i\alpha_j y_i y_j\langle \Phi(\boldsymbol{X}_i)\boldsymbol{P}(\boldsymbol{P}^T\boldsymbol{P})^{-1}\boldsymbol{P}^T,\Phi(\boldsymbol{X}_j)\boldsymbol{P}(\boldsymbol{P}^T\boldsymbol{P})^{-1}\boldsymbol{P}^T\rangle \quad (3)\\
&\text{subject to } 0\le\alpha_i\le C, i=1,\ldots,n.
\end{aligned}$$

Define $\boldsymbol{K}(i,j)\in\mathbb{R}^{d_1\times d_1}\times\mathbb{R}^{d_2\times d_2}$ as

$$\begin{aligned}
&\boldsymbol{K}(i,j) = (\boldsymbol{K}_r(i,j),\boldsymbol{K}_c(i,j)) \stackrel{\text{def}}{=} \Phi(\boldsymbol{X}_i)^T\Phi(\boldsymbol{X}_j)\\
&\text{where } [\boldsymbol{K}_r(i,j)]_{pq} = K_r([\boldsymbol{X}_i]_{p:},[\boldsymbol{X}_i]_{q:}), \stackrel{\text{def}}{=} \langle\phi_r([\boldsymbol{X}_i]_{p:}),\phi_r([\boldsymbol{X}_j]_{q:})\rangle,
\end{aligned}$$

$$[\boldsymbol{K}_c(i,j)]_{pq} = K_c([\boldsymbol{X}_i]_{:p},[\boldsymbol{X}_i]_{:q}) \stackrel{\text{def}}{=} \langle \phi_c([\boldsymbol{X}_i]_{:p}), \phi_c([\boldsymbol{X}_j]_{:q})\rangle.$$

Therefore, we can successfully estimate $\boldsymbol{\alpha}$ with quadratic programming based on $\boldsymbol{K}$ without description of feature mapping ϕ_r, ϕ_c. We update $\boldsymbol{C}$ as

$$\boldsymbol{C} = \sum_{i=1}^n \alpha_i y_i \Phi(\boldsymbol{X}_i)\boldsymbol{P}(\boldsymbol{P}^T\boldsymbol{P})^{-1} \in \mathcal{H}_r^r \times \mathcal{H}_c^r. \tag{4}$$

2. Second, we update $\boldsymbol{P}$ holding $\boldsymbol{C}$ fixed. The dual problem of Equation (2) is

$$\min_{\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_n)} -\sum_{i=1}^n \alpha_i + \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n \alpha_i\alpha_j y_i y_j \langle \boldsymbol{C}\left((\boldsymbol{C}^T\boldsymbol{C})^{-1}\boldsymbol{C}^T\Phi(\boldsymbol{X}_i)\right), \boldsymbol{C}\left((\boldsymbol{C}^T\boldsymbol{C})^{-1}\boldsymbol{C}^T\Phi(\boldsymbol{X}_j)\right)\rangle, \tag{5}$$

subject to $0 \leq \alpha_i \leq C, i = 1,\ldots,n,$

Notice $\boldsymbol{C}\left((\boldsymbol{C}^T\boldsymbol{C})^{-1}\boldsymbol{C}^T\Phi(\boldsymbol{X}_i)\right) \in \mathcal{H}^{d_1}\times\mathcal{H}^{d_2}$ is well defined by matrix product: for $\boldsymbol{A}_1 \in \mathcal{H}^r$ and $\boldsymbol{A}_2 \in \mathcal{H}^d$, $\boldsymbol{A}_1^T\boldsymbol{A}_2 = [\![a_{ij}]\!] \in \mathbb{R}^{r\times d}$, where $a_{ij} = \langle[\boldsymbol{A}_1]_i, [\boldsymbol{A}_2]_j\rangle$. We can find an optimizer of (5) without the feature mapping. To show this, notice that by plugging (4) into (5), we have

$$\begin{aligned}\boldsymbol{C}^T\boldsymbol{C} &= \sum_{i=1}^n\sum_{j=1}^n \alpha_i\alpha_j y_i y_j (\boldsymbol{P}^T\boldsymbol{P})^{-1}\boldsymbol{P}^T\boldsymbol{K}(i,j)\boldsymbol{P}(\boldsymbol{P}^T\boldsymbol{P})^{-1} \in \mathbb{R}^{r\times r}\times\mathbb{R}^{r\times r}, \\ \boldsymbol{C}^T\Phi(\boldsymbol{X}_i) &= \sum_{j=1}^n \alpha_i y_i (\boldsymbol{P}^T\boldsymbol{P})^{-1}\boldsymbol{P}^T\boldsymbol{K}(i,j) \in \mathbb{R}^{r\times d_1}\times\mathbb{R}^{r\times d_2}.\end{aligned} \tag{6}$$

(6) makes inner product in (5) expressed in terms of only $\boldsymbol{P}$ and $\{\boldsymbol{K}(i,j) \colon\ i,j\in[n]\}$ by the following equation.

$$\langle \boldsymbol{C}\left((\boldsymbol{C}^T\boldsymbol{C})^{-1}\boldsymbol{C}^T\Phi(\boldsymbol{X}_i)\right), \boldsymbol{C}\left((\boldsymbol{C}^T\boldsymbol{C})^{-1}\boldsymbol{C}^T\Phi(\boldsymbol{X}_j)\right)\rangle = \operatorname{tr}\left(\left(\boldsymbol{C}^T\Phi(\boldsymbol{X}_i)\right)^T(\boldsymbol{C}^T\boldsymbol{C})^{-1}\left(\boldsymbol{C}^T\Phi(\boldsymbol{X}_i)\right)\right).$$

We update $\boldsymbol{P}$ from an optimal coefficient $\boldsymbol{\alpha}$ of (5) and the formulas in (6).

$$\boldsymbol{P} = \sum_{i=1}^n \alpha_i y_i (\boldsymbol{C}^T\boldsymbol{C})^{-1}\boldsymbol{C}^T\Phi(\boldsymbol{X}_i).$$

Based on the procedure, we can obtain estimated $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{P}} = (\hat{\boldsymbol{P}}_r, \hat{\boldsymbol{P}}_c)$, Therefore, our estimated classifier is

$$\hat{f}(\boldsymbol{X}) = \sum_{k=1}^n \hat{\alpha}_k y_k \left(\sum_{i=1}^{d_1}\sum_{j=1}^{d_1}[\hat{\boldsymbol{P}}_r(\hat{\boldsymbol{P}}_r^T\hat{\boldsymbol{P}}_r)^{-1}\hat{\boldsymbol{P}}_r^T]_{ij}K_r\left([\boldsymbol{X}_k]_{i:},[\boldsymbol{X}]_{j:}\right) + \sum_{i=1}^{d_2}\sum_{j=1}^{d_2}[\hat{\boldsymbol{P}}_c(\hat{\boldsymbol{P}}_c^T\hat{\boldsymbol{P}}_c)^{-1}\hat{\boldsymbol{P}}_c^T]_{ij}K_c\left([\boldsymbol{X}_k]_{:i},[\boldsymbol{X}]_{:j}\right)\right).$$

Notice $\hat{\boldsymbol{W}}_r = \hat{\boldsymbol{P}}_r(\hat{\boldsymbol{P}}_r^T\hat{\boldsymbol{P}}_r)^{-1}\hat{\boldsymbol{P}}_r^T, \hat{\boldsymbol{W}}_c = \hat{\boldsymbol{P}}_c(\hat{\boldsymbol{P}}_c^T\hat{\boldsymbol{P}}_c)^{-1}\hat{\boldsymbol{P}}_c^T$ and $\hat{\boldsymbol{\gamma}} = \boldsymbol{y}\circ\hat{\boldsymbol{\alpha}}$.

Remark 1. From derivation of algorithm, we are looking for parameters $\boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{P}, \boldsymbol{C}$ which optimize

$$L_P = \frac{1}{2}\|\boldsymbol{C}\boldsymbol{P}^T\|_F^2 + c\sum_{i=1}^n \xi_i - \sum_{i=1}^n \alpha_i \left(y_i\langle \boldsymbol{C}\boldsymbol{P}^T, \Phi(\boldsymbol{X}_i)\rangle - (1-\xi_i)\right) - \sum_{i=1}^n \mu_i\xi_i,$$

where $\boldsymbol{\alpha}, \boldsymbol{\mu}$ are Laglange multiplies. So dual problem (3) updates $(\boldsymbol{C}, \boldsymbol{\alpha})$ and (5) updates $(\boldsymbol{P}, \boldsymbol{\alpha})$ and $\boldsymbol{\mu} = C - \boldsymbol{\alpha}$. Therefore, we do not have to distinguish coefficients $\boldsymbol{\alpha}, \boldsymbol{\beta}$ in two dual problems as I specified before.

3 Relation with the previous algorithm symmetric trick

Define symmetric feature matrix $\tilde{\boldsymbol{X}} = \begin{pmatrix} 0_{d_1\times d_2} & \boldsymbol{X} \\ \boldsymbol{X}^t & 0_{d_2\times d_1}\end{pmatrix} \in \mathbb{R}^{(d_1+d_2)\times(d_1+d_2)}$. Feature mapping 3 is defined as

$$\begin{aligned}\tilde{\Phi}\colon \mathbb{R}^{d_1\times d_2} &\to \mathcal{H}^{d_1+d_2} \\ \boldsymbol{X} &\mapsto \left(\phi([\tilde{\boldsymbol{X}}]_{1:}), \ldots, \phi([\tilde{\boldsymbol{X}}]_{d_1+d_2:})\right)\end{aligned}$$

where ϕ is induced by kernel $K\colon \mathbb{R}^{(d_1+d_2)\times(d_1+d_2)} \times \mathbb{R}^{(d_1+d_2)\times(d_1+d_2)} \to \mathbb{R}$. Since all entries of $\Phi_r(\boldsymbol{X})$ are corresponding to $[\tilde{\Phi}(\boldsymbol{X})]_{1:d_1}$ and $\Phi_c(\boldsymbol{X})$ to $[\tilde{\Phi}(\tilde{\boldsymbol{X}})]_{d_1+1:d_1+d_2}$, we have an equivalent representation of (1)

$$\begin{aligned}f(\boldsymbol{X}) &= \langle \boldsymbol{B}, \Phi(\boldsymbol{X})\rangle \\ &= \langle \boldsymbol{B}_r, \Phi_r(\boldsymbol{X})\rangle + \langle \boldsymbol{B}_c, \Phi_c(\boldsymbol{X})\rangle \\ &= \langle \tilde{\boldsymbol{B}}_r, [\tilde{\Phi}(\boldsymbol{X})]_{1:d_1}\rangle + \langle \tilde{\boldsymbol{B}}_c, [\tilde{\Phi}(\boldsymbol{X})]_{d_1+1:d_1+d_2}\rangle, \text{ where } \tilde{\boldsymbol{B}}_r \in \mathcal{H}^{d_1}, \tilde{\boldsymbol{B}}_c \in \mathcal{H}^{d_2} \\ &= \langle \tilde{\boldsymbol{B}}, \tilde{\Phi}(\boldsymbol{X})\rangle, \text{ where } \tilde{\boldsymbol{B}} = (\tilde{\boldsymbol{B}}_r, \tilde{\boldsymbol{B}}_c) \in \mathcal{H}^{d_1+d_2}.\end{aligned}$$

Assume that $\tilde{\boldsymbol{B}} = \tilde{\boldsymbol{C}}\tilde{\boldsymbol{P}}^T$ where $\tilde{\boldsymbol{C}} \in \mathcal{H}^r$, $\tilde{\boldsymbol{P}} = (\tilde{\boldsymbol{P}}_r, \tilde{\boldsymbol{P}}_c) \in \mathbb{R}^{(d_1+d_2)\times r}$ and $\tilde{\boldsymbol{P}}_r \in \mathbb{R}^{d_1\times r}, \tilde{\boldsymbol{P}}_c \in \mathbb{R}^{d_2\times r}$. Let Π_r, Π_c are permutation operators such that

$$\begin{aligned}\mathrm{Proj}_{\mathcal{H}_r}\left(\Pi_r[\tilde{\Phi}(\boldsymbol{X})]_{1:d_1}\right) &= \Phi_r(\boldsymbol{X}) \\ \mathrm{Proj}_{\mathcal{H}_c}\left(\Pi_c[\tilde{\Phi}(\boldsymbol{X})]_{d_1+1:d_1+d_2}\right) &= \Phi_c(\boldsymbol{X}).\end{aligned}$$

Here, we denote $\mathrm{Proj}_{\mathcal{H}_c}\colon \mathcal{H} \to \mathcal{H}_r$ and $\mathrm{Proj}_{\mathcal{H}_r}\colon \mathcal{H} \to \mathcal{H}_c$ as entry-wise projection mappings. Then the following holds

$$\begin{aligned}\langle \tilde{\boldsymbol{B}}, \tilde{\Phi}(\boldsymbol{X})\rangle &= \langle \tilde{\boldsymbol{C}}(\tilde{\boldsymbol{P}}_r, \tilde{\boldsymbol{P}}_c)^T, \tilde{\Phi}(\boldsymbol{X})\rangle \\ &= \langle \tilde{\boldsymbol{C}}\tilde{\boldsymbol{P}}_r^T, [\tilde{\Phi}(\boldsymbol{X})]_{1:d_1}\rangle + \langle \tilde{\boldsymbol{C}}\tilde{\boldsymbol{P}}_C^T, [\tilde{\Phi}(\boldsymbol{X})]_{d_1+1:d_1+d_2}\rangle \\ &= \langle \Pi_r\tilde{\boldsymbol{C}}\tilde{\boldsymbol{P}}_r^T, \Pi_r[\tilde{\Phi}(\boldsymbol{X})]_{1:d_1}\rangle + \langle \Pi_c\tilde{\boldsymbol{C}}\tilde{\boldsymbol{P}}_c^T, \Pi_c[\tilde{\Phi}(\boldsymbol{X})]_{d_1+1:d_1+d_2}\rangle \\ &= \langle \tilde{\boldsymbol{C}}_r\tilde{\boldsymbol{P}}_r^T, \Phi_r(\boldsymbol{X})\rangle + \langle \tilde{\boldsymbol{C}}_c\tilde{\boldsymbol{P}}_c^T, \Phi_c(\boldsymbol{X})\rangle,\end{aligned}$$

where $\tilde{\boldsymbol{C}}_r = \mathrm{Proj}_{\mathcal{H}_r}(\Pi_r\tilde{\boldsymbol{C}})$ and $\tilde{\boldsymbol{C}}_c = \mathrm{Proj}_{\mathcal{H}_c}(\Pi_c\tilde{\boldsymbol{C}})$. Therefore, we can conclude that the low rankness of the coefficient on the feature image of $\tilde{\Phi}(\boldsymbol{X})$ implies the same low rankness of the coefficient of the feature image of $\Phi(\boldsymbol{X})$. The other direction is also true.