

Algorithmic perspectives of kernel method

Chanwoo Lee, August 5, 2020

1 A choice of feature mapping Φ

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

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

We define decision function,

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

where $\mathbf{X}^1, \dots, \mathbf{X}^n$ are sampled matrix features and $\mathbf{W}^{\text{col}}, \mathbf{W}^{\text{row}}$ are some positive semi definite matrices with low rank. We estimate $\boldsymbol{\alpha}^r = (\alpha_1^r, \dots, \alpha_n^r)$, $\boldsymbol{\alpha}^c = (\alpha_1^c, \dots, \alpha_n^c)$, \mathbf{W}^{col} , and \mathbf{W}^{row} from the training data set.

2 Algorithm derivation

We solve an optimization problem

$$\begin{aligned}\min_{\mathbf{B}} \quad & \frac{1}{2} \|\mathbf{B}\|_F^2 + c \sum_{i=1}^n \xi_i, \\ \text{subject to} \quad & y_i \langle \mathbf{B}, \Phi(\mathbf{X}_i) \rangle \leq 1 - \xi_i \text{ and } \xi_i \geq 0, i = 1, \dots, n.\end{aligned} \tag{2}$$

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

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

where $\mathbf{C} = (\mathbf{C}_r, \mathbf{C}_c) \in \mathcal{H}_r^r \times \mathcal{H}_c^r$ and $\mathbf{P} = (\mathbf{P}_r, \mathbf{P}_c) \in \mathbb{R}^{d_1 \times r} \times \mathbb{R}^{d_2 \times r}$. We assume that $\mathbf{P}_r, \mathbf{P}_c$ are orthonormal matrices.

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

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

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

$$\mathbf{K}(i, j) = (\mathbf{K}_r(i, j), \mathbf{K}_c(i, j)) \stackrel{\text{def}}{=} \Phi(\mathbf{X}_i)^T \Phi(\mathbf{X}_j)$$

$$\begin{aligned} \text{where } [\mathbf{K}_r(i, j)]_{pq} &= K_r([\mathbf{X}_i]_p, [\mathbf{X}_j]_q) \stackrel{\text{def}}{=} \langle \phi_r([\mathbf{X}_i]_p), \phi_r([\mathbf{X}_j]_q) \rangle, \\ [\mathbf{K}_c(i, j)]_{pq} &= K_c([\mathbf{X}_i]_p, [\mathbf{X}_j]_q) \stackrel{\text{def}}{=} \langle \phi_c([\mathbf{X}_i]_p), \phi_c([\mathbf{X}_j]_q) \rangle. \end{aligned}$$

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

$$\mathbf{C} = \sum_{i=1}^n \alpha_i y_i \Phi(\mathbf{X}_i) \mathbf{P} \in \mathcal{H}_r^r \times \mathcal{H}_c^r. \quad (3)$$

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

$$\min_{\beta=(\beta_1, \dots, \beta_n)} - \sum_{i=1}^n \beta_i + \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \beta_i \beta_j y_i y_j \langle \mathbf{C} ((\mathbf{C}^T \mathbf{C})^{-1} \mathbf{C}^T \Phi(\mathbf{X}_i)), \mathbf{C} ((\mathbf{C}^T \mathbf{C})^{-1} \mathbf{C}^T \Phi(\mathbf{X}_j)) \rangle, \quad (4)$$

$$\text{subject to } \sum_{i=1}^n y_i \beta_i = 0, \text{ and } 0 \leq \beta_i \leq C, i = 1, \dots, n,$$

Notice $\mathbf{C} ((\mathbf{C}^T \mathbf{C})^{-1} \mathbf{C}^T \Phi(\mathbf{X}_i)) \in \mathcal{H}^{d_1} \times \mathcal{H}^{d_2}$ is well defined by matrix product: for $\mathbf{A}_1 \in \mathcal{H}^r$ and $\mathbf{A}_2 \in \mathcal{H}^d$, $\mathbf{A}_1^T \mathbf{A}_2 = \llbracket a_{ij} \rrbracket \in \mathbb{R}^{r \times d}$, where $a_{ij} = \langle [\mathbf{A}_1]_i, [\mathbf{A}_2]_j \rangle$. We can find an optimizer of (4) without the feature mapping. To show this, notice that by plugging (3) into (4), we have

$$\begin{aligned} \mathbf{C}^T \mathbf{C} &= \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \mathbf{P}^T \mathbf{K}(i, j) \mathbf{P} \in \mathbb{R}^{r \times r} \times \mathbb{R}^{r \times r}, \\ \mathbf{C}^T \Phi(\mathbf{X}_i) &= \sum_{j=1}^n \alpha_j y_j \mathbf{P}^T \mathbf{K}(i, j) \in \mathbb{R}^{r \times d_1} \times \mathbb{R}^{r \times d_2}. \end{aligned} \quad (5)$$

(5) makes inner product in (4) expressed in terms of only \mathbf{P} and $\{\mathbf{K}(i, j): i, j \in [n]\}$ by the following equation.

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

3 Relation with the previous algorithm symmetric trick

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

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

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

$$f(\mathbf{X}) = \langle \mathbf{B}, \Phi(\mathbf{X}) \rangle$$

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

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

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

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

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

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