A formulation of kernel matrix

Chanwoo Lee, August 2, 2020

1 Some thoughts about feature mapping Φ

In the last note, the feature mapping Φ is defined as

$$[\Phi(\boldsymbol{X})]_{ij} \stackrel{\text{def}}{=} \begin{cases} (\phi(\boldsymbol{X}_{i:}), \phi(\boldsymbol{X}_{j:}))\,, & \text{if } i \geq j \\ [\Phi(\boldsymbol{X})]_{ji} & \text{if } i < j. \end{cases} \tag{1}$$

Based on the feature mapping, we define decision function

$$f(\boldsymbol{X}) \stackrel{\text{def}}{=} \langle \boldsymbol{B}, \Phi(\boldsymbol{X})\rangle, \text{ where } \boldsymbol{B} \in \text{Sym}_d(\mathcal{H}^2) \text{ and } \text{rank}(\boldsymbol{B}) \leq r. \tag{2}$$

I have two remarks about this definition.

Remark 1. I cannot find this definition is equivalent to the previous decision function which is defined as

$$f(\boldsymbol{X}) = \langle \boldsymbol{B}, (\phi(\boldsymbol{X}_{1:}), \ldots, \phi(\boldsymbol{X}_{d:}))\rangle\,. \text{ where } \boldsymbol{B} \in (\mathcal{H})^d \text{ and } \boldsymbol{B} = \boldsymbol{C}\boldsymbol{P} \text{ for some } \boldsymbol{C} \in (\mathcal{H})^r, \boldsymbol{P} \in \mathbb{R}^{r\times d}. \tag{3}$$

Notice that the decision function (3) is what we use for optimization problem. One reason is that the free parameters in (2) is $\dim(\mathcal{H})^{r(r-1)} + dr$ while $\dim(\mathcal{H})^r + rd$ in (3). Even if we consider $\Phi(\boldsymbol{X})$ has $2d$ same variables for each $\phi(\boldsymbol{X}_{i:})$ for $i = 1, \ldots, d$, it cannot justify this big difference of the free parameters. Therefore, Equation (2) and (3) have similar formula but not exactly equivalent. In addition, I am not sure that newly defined matrix feature mapping $\Phi(\boldsymbol{X})$ has better representation than $(\phi(\boldsymbol{X}_{1:}), \ldots, \phi(\boldsymbol{X}_{d:}))$ where there is no replicates of features.

Remark 2. The decision function (2) is not equivalent to

$$f(\boldsymbol{X}) = \sum_{k=1}^{n} \alpha_k \sum_{i,j\in[d]} w_{i,j} K(\boldsymbol{X}^k_{i:}, \boldsymbol{X}_{j:}),$$

where $\boldsymbol{X}^1, \ldots, \boldsymbol{X}^n$ are n-samples, $\alpha_k \in \mathbb{R}$ and $w_{i,j} = [\boldsymbol{W}]_{ij}$ such that $\text{rank}(\boldsymbol{W}) \leq r$. To check this, let us consider simple case where $d = 2$ and $r = 1$. We consider $\langle \boldsymbol{P}\Phi(\boldsymbol{X}), \boldsymbol{P}\Phi(\boldsymbol{X}')\rangle$ where $\boldsymbol{P} \in \mathbb{R}^{2\times 1}$ which equals to

$$\left\langle \begin{pmatrix} p_1 \\ p_2 \end{pmatrix}^T \begin{pmatrix} (\phi(\boldsymbol{X}_{1:}), \phi(\boldsymbol{X}_{1:})) & (\phi(\boldsymbol{X}_{2:}), \phi(\boldsymbol{X}_{1:})) \\ (\phi(\boldsymbol{X}_{2:}), \phi(\boldsymbol{X}_{1:})) & (\phi(\boldsymbol{X}_{2:}), \phi(\boldsymbol{X}_{2:})) \end{pmatrix}, \begin{pmatrix} p_1 \\ p_2 \end{pmatrix}^T \begin{pmatrix} (\phi(\boldsymbol{X}'_{1:}), \phi(\boldsymbol{X}'_{1:})) & (\phi(\boldsymbol{X}'_{2:}), \phi(\boldsymbol{X}'_{1:})) \\ (\phi(\boldsymbol{X}'_{2:}), \phi(\boldsymbol{X}'_{1:})) & (\phi(\boldsymbol{X}'_{2:}), \phi(\boldsymbol{X}'_{2:})) \end{pmatrix} \right\rangle$$

$$= (3p_1^2 + 2p_1p_2 + p_2^2)K(\boldsymbol{X}_{1:}, \boldsymbol{X}'_{1:}) + 2p_1p_2K(\boldsymbol{X}_{1:}, \boldsymbol{X}'_{2:}) + 2p_1p_2K(\boldsymbol{X}_{1:}, \boldsymbol{X}'_{2:}) + (p_1^2 + 2p_1p_2 + 3p_2^2)K(\boldsymbol{X}_{2:}, \boldsymbol{X}'_{2:})$$

$$= \left\langle \begin{pmatrix} 3p_1^2 + 2p_1p_2 + p_2^2 & 2p_1p_2 \\ 2p_1p_2 & p_1^2 + 2p_1p_2 + 3p_2^2 \end{pmatrix} \boldsymbol{K}. \right\rangle$$

where $[\boldsymbol{K}]_{ij} = K(\boldsymbol{X}_{i:}, \boldsymbol{X}'_{j:}) = \langle \phi(\boldsymbol{X}_{i:}), \phi(\boldsymbol{X}'_{j:})\rangle$. Notice the matrix $\boldsymbol{W} = \begin{pmatrix} 3p_1^2 + 2p_1p_2 + p_2^2 & 2p_1p_2 \\ 2p_1p_2 & p_1^2 + 2p_1p_2 + 3p_2^2 \end{pmatrix}$ is not rank 1. The reason for $\boldsymbol{W}$ not being rank 1 is $\Phi(\boldsymbol{X})$ has many duplicates of elements of $\boldsymbol{K}$.

2 A formulation of feature mapping Φ

We go back to feature mapping with suggested Hilbert space notation. We define feature mapping $\Phi\colon \mathbb{R}^{d\times d} \to \mathcal{H}^d$ as

$$\Phi(\boldsymbol{X}) = \left(\phi(\boldsymbol{X}_{1:}), \ldots, \phi(\boldsymbol{X}_{d:})\right).$$

With this definition we have the the same decision function (3) that we can use our current optimization algorithm. In addition, we can address problem mentioned in Remark 2. From Equation (3), we have

$$f(\boldsymbol{X}) = \langle \boldsymbol{B}, \Phi(\boldsymbol{X})\rangle = \langle \boldsymbol{C}\boldsymbol{P}, \Phi(\boldsymbol{X})\rangle = \left\langle \boldsymbol{C}, \Phi(\boldsymbol{X})\boldsymbol{P}^T\right\rangle \text{ where } \boldsymbol{C} \in (\mathcal{H})^r, \boldsymbol{P} \in \mathbb{R}^{r\times d}. \tag{4}$$

Therefore, $f(\boldsymbol{X})$ is an element of RKHS induced by $\langle \Phi(\boldsymbol{X})\boldsymbol{P}^T, \Phi(\boldsymbol{X}')\boldsymbol{P}^T\rangle$. Let $\boldsymbol{P} = (P_1, \ldots, P_r)^T$.

$$\begin{aligned}
\langle \Phi(\boldsymbol{X})\boldsymbol{P}^T, \Phi(\boldsymbol{X}')\boldsymbol{P}^T\rangle &= \left\langle \left(\Phi(\boldsymbol{X})P_1, \ldots, \Phi(\boldsymbol{X})P_r\right), \left(\Phi(\boldsymbol{X}')P_1, \ldots, \Phi(\boldsymbol{X}')P_r\right)\right\rangle \\
&= \sum_{i=1}^{r} \langle \Phi(\boldsymbol{X})P_i, \Phi(\boldsymbol{X}')P_i\rangle \\
&= \sum_{i=1}^{r} \left\langle \sum_{j=1}^{d} P_{ij}\phi(\boldsymbol{X}_{j:}), \sum_{j=1}^{d} P_{ij}\phi(\boldsymbol{X}'_{j:})\right\rangle \\
&= \sum_{i=1}^{r} P_i^T \boldsymbol{K} P_i \text{ where } [\boldsymbol{K}]_{ij} = K(\boldsymbol{X}_{i:}, \boldsymbol{X}'_{j:}) = \left\langle \phi(\boldsymbol{X}_{i:}), \phi(\boldsymbol{X}'_{j:})\right\rangle \\
&= \left\langle \sum_{i=1}^{r} P_i P_i^T, \boldsymbol{K}\right\rangle \\
&= \sum_{i,j\in[d]} w_{ij} K(\boldsymbol{X}_{i:}, \boldsymbol{X}'_{j:}) \text{ where } \boldsymbol{W} = \sum_{i=1}^{r} P_i P_i^T.
\end{aligned}$$

Therefore, we can write considered decision function class as

$$\begin{aligned}
\mathcal{F} &= \left\{ f\colon \boldsymbol{X} \mapsto \langle \boldsymbol{C}, \Phi(\boldsymbol{X})\boldsymbol{P}^T\rangle | \boldsymbol{P}\boldsymbol{P}^T = \boldsymbol{I}, \boldsymbol{P} \in \mathbb{R}^{r\times d}, \boldsymbol{C} \in \mathcal{H}^r \right\} \\
&= \left\{ f \in \text{RKHS generated by } \langle \cdot, \cdot \rangle_{\boldsymbol{P}} | \boldsymbol{P}\boldsymbol{P}^T = \boldsymbol{I}, \boldsymbol{P} \in \mathbb{R}^{r\times d} \right\} \\
&= \left\{ f \in \text{RKHS generated by } \mathcal{K}(\boldsymbol{W}) | \boldsymbol{W} \succeq 0, \operatorname{rank}(\boldsymbol{W}) \le r \right\}.
\end{aligned}$$

Remark 3. Connection to our learning algorithm: We consider the optimization over the union of RKHS induced by low rank $\boldsymbol{W}$:

$$\max_{f\in\mathcal{F}(r)} L(f) = \max_{\operatorname{rank}(\boldsymbol{W})\le r, \boldsymbol{W}\succeq 0} \; \max_{f\in \text{RKHS}(\mathcal{K}(\boldsymbol{W}))} L(f).$$

Remark 4. In this case, feature matrix $\boldsymbol{X}$ does not have to be symmetric matrix. However, when $\boldsymbol{X} \in \mathbb{R}^{d_1\times d_2}$, we are considering only row-wise extended feature map images. This might be enough when the feature matrix is known to have only one of columnwise or rowwise correlation. In the following section, we propose method for more general feature matrix case where we need to analyze both column and row structures.

3 Case for non symmetric feature matrices

We consdier the case where feature matrices are not symmetric $\boldsymbol{X} \in \mathbb{R}^{d_1 \times d_2}$. Define feature mapping $\Phi_r(\boldsymbol{X}) \overset{\text{def}}{=} (\phi_r(\boldsymbol{X}_{1:}), \ldots, \phi_r(\boldsymbol{X}_{d_1:}))$ and $\Phi_c(\boldsymbol{X}) \overset{\text{def}}{=} (\phi_c(\boldsymbol{X}_{:1}), \ldots, \phi_c(\boldsymbol{X}_{:d_2}))$ where ϕ_c, ϕ_r is feature map induced by kernel K. Notice we use the same kernel K in the sense that if we use Gaussian kernel on row vectors, we use the same kernel on column vectors so that the formula of ϕ_r, ϕ_c are the same except the dimension. We define comprehensive feature map $\tilde{\Phi} \colon \mathbb{R}^{d_1 \times d_2} \to \mathcal{H}_r^{d_1} \times \mathcal{H}_c^{d_2}$ where $\mathcal{H}_r$ and $\mathcal{H}_c$ are image space of ϕ_c and ϕ_r.

$$\tilde{\Phi}(\boldsymbol{X}) = (\Phi_r(\boldsymbol{X}), \Phi_c(\boldsymbol{X})).$$

We define decision function,

$$\begin{aligned} f(\boldsymbol{X}) &= \langle \tilde{\boldsymbol{B}}, \tilde{\Phi}(\boldsymbol{X}) \rangle, \text{ where } \tilde{\boldsymbol{B}} = (\tilde{\boldsymbol{B}}_r, \tilde{\boldsymbol{B}}_c) \in \mathcal{H}_r^{d_1} \times \mathcal{H}_c^{d_2} \qquad (5) \\ &= \langle \tilde{\boldsymbol{B}}_r, \Phi_r(\boldsymbol{X}) \rangle + \langle \tilde{\boldsymbol{B}}_c, \Phi_c(\boldsymbol{X}) \rangle \\ &= \sum_{k=1}^{n} \beta_k \sum_{i,j \in [d_2]} w_{ij}^{row} K(\boldsymbol{X}_{i\cdot}^k, \boldsymbol{X}_{j\cdot}) + \sum_{k=1}^{n} \alpha_k \sum_{i,j \in [d_2]} w_{ij}^{col} K(\boldsymbol{X}_{\cdot i}^k, \boldsymbol{X}_{\cdot j}), \end{aligned}$$

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. We estimate $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n), \boldsymbol{\beta} = (\beta_1, \ldots, \beta_n), \boldsymbol{W}^{\text{col}}$, and $\boldsymbol{W}^{\text{row}}$ from the training data set. Define new symmetric feature matrix $\tilde{\boldsymbol{X}} = \begin{pmatrix} 0_{d_1 \times d_2} & \boldsymbol{X} \\ \boldsymbol{X}^t & 0_{d_2 \times d_1} \end{pmatrix} \in \mathbb{R}^{(d_1+d_2) \times (d_1+d_2)}$ and feature map $\Phi \colon \mathbb{R}^{(d_1+d_2) \times (d_1+d_2)} \to \mathcal{H}^{d_1+d_2}$ as in symmetric case (1),

$$\Phi(\tilde{\boldsymbol{X}}) = \left(\phi(\tilde{\boldsymbol{X}}_{1:}), \ldots, \phi(\tilde{\boldsymbol{X}}_{d_1+d_2:}) \right),$$

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 $[\Phi(\tilde{\boldsymbol{X}})]_{1:d_1}$ and $\Phi_c(\boldsymbol{X})$ to $[\Phi(\tilde{\boldsymbol{X}})]_{d_1+1:d_1+d_2}$, we have an equivalent representation of (5)

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

Therefore, we can apply low-rank structure on $\boldsymbol{B} = \boldsymbol{C}\boldsymbol{P}$ where $\boldsymbol{C} \in \mathcal{H}^r$ and $\boldsymbol{P} \in \mathbb{R}^{r \times (d_1+d_2)}$ and use the same optimization procedure with Section 2.