An equivalent formulation of matrix kernels

Miaoyan Wang, July 30, 2020

For ease of notation, we assume the feature matrices are symmetric in that $\boldsymbol{X} = \boldsymbol{X}^T$ and $d_1 = d_2 = d$. The adaptation to non-symmetric matrices are easy to derive.

- Let $\text{Sym}_d(\mathbb{R}) = \{\boldsymbol{X} \mid \boldsymbol{X} = [\![x_{ij}]\!],\ x_{ij} = x_{ji} \in \mathbb{R},\ \text{for all } (i,j) \in [d] \times d\}$ denote the collection of d-by-d symmetric matrices with each entry taking values in $\mathbb{R}$.
- Let $\mathcal{H}$ denote a possibly infinite dimensional Hilbert space.
- Let $\mathcal{H}^{d\times d} = \{\boldsymbol{X} \mid \boldsymbol{X} = [\![x_{ij}]\!], x_{ij} \in \mathcal{H},\ \text{for all } (i,j) \in [d] \times d\}$ denote the collection of d-by-d matrices with each entry taking value in $\mathcal{H}$. Similarly, $\text{Sym}_d(\mathcal{H})$ denotes the collection of d-by-d symmetric matrices defined over $\mathcal{H}$.

Matrix algebraic operations are carried over from $\mathbb{R}^{d\times d}$ to $\mathcal{H}^{d\times d}$. The main difference is that the multiplication in $\mathbb{R}$ is replaced by inner product in $\mathcal{H}$.

Proposition 1. Let $\boldsymbol{B} = [\![b_{ij}]\!]$ and $\boldsymbol{B}' = [\![b'_{ij}]\!]$ be two matrices in $\mathcal{H}^{d\times d}$. Let $\boldsymbol{P} = [\![p_{ij}]\!] \in \mathbb{R}^{d\times r}$ be a real-valued matrix.

1. Sum: $\boldsymbol{B} + \boldsymbol{B}' = [\![b_{ij} + b'_{ij}]\!] \in \mathcal{H}^{d\times d}$.
2. Inner product: $\langle \boldsymbol{B}, \boldsymbol{B}' \rangle = \sum_{ij} \langle b_{ij},\ b'_{ij} \rangle \in \mathbb{R}$.
3. Linear combination: $\boldsymbol{BP} = [\![c_{ij}]\!] \in \mathcal{H}^{d\times r}$, where $c_{ij} = \sum_{s\in[d]} b_{is} p_{sj} \in \mathcal{H}^2$ for all $(i,j) \in [d] \times [r]$.
4. Matrix product: $\boldsymbol{BB}' = [\![c_{ij}]\!] \in \mathbb{R}^{d\times d}$, where $c_{ij} = \sum_{s\in[d]} \langle b_{is},\ b'_{sj} \rangle$.

Now we are ready to present the matrix kernel and associated feature mapping.

First viewpoint: feature mapping

Let $\phi\colon \mathbb{R}^d \to \mathcal{H}$ be a feature mapping associated with a classical kernel $K\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. Define a feature mapping Φ over the matrix space

$$\begin{aligned} \Phi : \text{Sym}_d(\mathbb{R}) &\to \text{Sym}_d(\mathcal{H}^2) \\ \boldsymbol{X} &\mapsto \Phi(\boldsymbol{X}) = [\![\Phi(\boldsymbol{X})_{ij}]\!], \end{aligned}$$

where each element of $\Phi(\boldsymbol{X})$ is a pair of possibly infinite dimensional features.

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

Note that the entry f_{ij} takes value in $\mathcal{H}^2$ for all $(i,j) \in [d] \times [d]$. Furthermore, $\Phi(\boldsymbol{X})$ is a symmetric matrix in the sense that $[\Phi(\boldsymbol{X})]_{ij} = [\Phi(\boldsymbol{X})]_{ji}$ for all $(i,j) \in [d] \times [d]$. We propose to consider decision functions $f\colon \mathrm{Sym}_d(\mathbb{R}) \to \mathbb{R}$ using the linear functions with respect to $\Phi(\boldsymbol{X}) \in \mathrm{Sym}_d(\mathcal{H}^2)$,

$$f(\boldsymbol{X}) \stackrel{\text{def}}{=} \langle \boldsymbol{B},\ \Phi(\boldsymbol{X})\rangle, \text{ where } \boldsymbol{B} \in \mathrm{Sym}_d(\mathcal{H}^2) \text{ and } \mathrm{rank}(\boldsymbol{B}) \le r. \tag{1}$$

Here the parameter matrix $\boldsymbol{B} = [\![b_{ij}]\!]$ is a d-by-d symmetric matrix with entries defined in $\mathcal{H}^2$. Suppose $\boldsymbol{B}$ admits low-rank decomposition, $\boldsymbol{B} = \boldsymbol{P}^T\boldsymbol{C}\boldsymbol{P}$. The class of functions (1) induced by all possible low-rank $\boldsymbol{B}$ is

$$\begin{aligned}
\mathcal{F} &= \left\{ f\colon \boldsymbol{X} \mapsto \langle \boldsymbol{C},\ \boldsymbol{P}^T\Phi(\boldsymbol{X})\boldsymbol{P}\rangle \ \middle|\ \boldsymbol{P}\boldsymbol{P}^T = \boldsymbol{I},\ \boldsymbol{P} \in \mathbb{R}^{d\times r},\ \boldsymbol{C} \in \mathrm{Sym}_d(\mathcal{H}^2) \right\} \\
&= \left\{ f \in \text{RKHS generated by } \mathcal{K}(\boldsymbol{P}) \ \middle|\ \boldsymbol{P}\boldsymbol{P}^T = \boldsymbol{I},\ \boldsymbol{P} \in \mathbb{R}^{d\times r} \right\} \\
&= \left\{ f \in \text{RKHS generated by } \mathcal{K}(\boldsymbol{W}) \ \middle|\ \boldsymbol{W} \succeq 0,\ \mathrm{rank}(\boldsymbol{W}) \le r \right\},
\end{aligned}$$

where $\boldsymbol{W} = \boldsymbol{P}^T\Lambda^2\boldsymbol{P}$ for some positive definite diagonal matrix $\Lambda \in \mathbb{R}^{r\times r}$ (see below).

Second viewpoint: matrix kernel

Definition 1 (Kernel defined in matrix space)**.** Let $K(\cdot,\ \cdot)$ be a classical kernel defined in vector space $\mathbb{R}^d$, and $\boldsymbol{W} = [\![w_{ij}]\!] \in \mathbb{R}^{d\times d}$ be a rank-r semi-positive definite matrix. Then $\boldsymbol{K}$ and $\boldsymbol{W}$ induce a matrix kernel $\mathcal{K}$:

$$\begin{aligned}
\mathcal{K}\colon \mathrm{Sym}_d(\mathbb{R}) \times \mathrm{Sym}_d(\mathbb{R}) &\mapsto \mathbb{R}, \\
(\boldsymbol{X}, \boldsymbol{X}') &\mapsto \mathcal{K}(\boldsymbol{X}, \boldsymbol{X}') = \sum_{i,j\in[d]} w_{ij} K(\boldsymbol{x}_i, \boldsymbol{x}'_j),
\end{aligned}$$

where $\boldsymbol{x}_i, \boldsymbol{x}'_j$ denote the i-th and j-th columns of $\boldsymbol{X}$, $\boldsymbol{X}'$, respectively.

Oftentime, the generator kernel K is specified by users, whereas the projection kernel $\boldsymbol{W}$ is learned by our algorithm. In particular $\boldsymbol{W} \propto \boldsymbol{I}_{d\times d}$ corresponds to the classical SVM. We use $\mathcal{K} = \mathcal{K}(\boldsymbol{W})$ to denote the reproducing kernel hilbert space (RKHS) induced by $\boldsymbol{W}$.

Connection to our learning algorithm

We consider the optimization over the union of RKHS induced by low rank $\boldsymbol{W}$:

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