Two mappings comparison

Chanwoo Lee, August 9, 2020

- Concatenated mapping

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

  and decision function is $f(\boldsymbol{X}) = \langle \boldsymbol{C}_1\boldsymbol{P}_1^T,\Phi_1(\boldsymbol{X})\rangle + \langle \boldsymbol{C}_2\boldsymbol{P}_2^T,\Phi_2(\boldsymbol{X})\rangle$

- Bilinear mapping

$$\begin{aligned}\Phi_{\text{bi}}\colon \mathbb{R}^{d_1\times d_2} &\to (\mathcal{H}_r\times\mathcal{H}_c)^{d_1\times d_2}\\ \boldsymbol{X} &\mapsto [\![\Phi_{\text{bi}}(\boldsymbol{X})_{ij}]\!],\ \text{where } \Phi_{\text{bi}}(\boldsymbol{X})_{ij}\stackrel{\text{def}}{=}(\phi_r(\boldsymbol{X}_{i:}),\phi_c(\boldsymbol{X}_{j:})).\end{aligned}$$

  and decision function is $f(\boldsymbol{X}) = \langle \boldsymbol{P}_1\boldsymbol{C}\boldsymbol{P}_2^T,\Phi_{\text{bi}}(\boldsymbol{X})\rangle$ where $\boldsymbol{C} = [\![\boldsymbol{c}_{ij}]\!] \in (\mathcal{H}_1\times\mathcal{H}_2)^{r\times r}$ with $\boldsymbol{c}_{ij} = (\boldsymbol{c}_i^{(1)},\boldsymbol{c}_j^{(2)})$.

Define an isomorphism

$$\begin{aligned}\mathcal{T}\colon \mathcal{H}_1^{d_1}\times\mathcal{H}_2^{d_2} &\to (\mathcal{H}_1\times\mathcal{H}_2)^{d_1\times d_2}\\ (\boldsymbol{a},\boldsymbol{b}) &\mapsto [\![\mathcal{T}(\boldsymbol{a},\boldsymbol{b})_{ij}]\!],\ \text{where } \mathcal{T}(\boldsymbol{a},\boldsymbol{b})_{ij} = (\boldsymbol{a}_i,\boldsymbol{b}_j) \text{ for all } (i,j)\in[d_1]\times[d_2].\end{aligned}$$

From this mapping we can re-express $\Phi_{\text{bi}} = \mathcal{T}(\Phi_r(\boldsymbol{X}),\Phi_c(\boldsymbol{X})) = \mathcal{T}(\Phi_{\text{con}})$. Therefore, it seems that bilinear mapping is one step more from concatenated mapping.

1. When ϕ_r and ϕ_c are identiy maps (linear case):

   In concatenated mapping case,

$$\begin{aligned} f(\boldsymbol{X}) &= \langle \boldsymbol{C}\boldsymbol{P}^T,\Phi_{\text{con}}(\boldsymbol{X})\rangle\\ &= \langle (\boldsymbol{C}_r\boldsymbol{P}_r^T,\boldsymbol{C}_c\boldsymbol{P}_c^T),(\Phi_r(\boldsymbol{X}),\Phi_c(\boldsymbol{X}))\rangle\\ &= \langle \boldsymbol{C}_r\boldsymbol{P}_r^T,\boldsymbol{X}^T\rangle + \langle \boldsymbol{C}_c\boldsymbol{P}_c^T,\boldsymbol{X}\rangle\end{aligned}$$

   In bilinear mapping case,

$$\begin{aligned} f(\boldsymbol{X}) &= \langle \boldsymbol{P}^{\text{row}}\boldsymbol{C}(\boldsymbol{P}^{\text{col}})^T,\Phi_{\text{bi}}(\boldsymbol{X})\rangle\\ &= \sum_{i,j,s,s'}\boldsymbol{P}^{\text{row}}_{si}\boldsymbol{P}^{\text{col}}_{s',j}\langle(\boldsymbol{c}_i^{\text{row}},\boldsymbol{c}_j^{\text{col}}),(\boldsymbol{X}_{s:},\boldsymbol{X}_{:s'})\rangle\\ &= \Big(\sum_{s',j}\boldsymbol{P}^{\text{col}}_{s',j}\Big)\sum_{s,i}\boldsymbol{P}^{\text{row}}_{s,i}\langle \boldsymbol{c}_i^{\text{row}},\boldsymbol{X}_{s:}\rangle + \Big(\sum_{s,i}\boldsymbol{P}^{\text{row}}_{s,i}\Big)\sum_{s',j}\boldsymbol{P}^{\text{col}}_{s',j}\langle \boldsymbol{c}_j^{\text{col}},\boldsymbol{X}_{:s'}\rangle\\ &= \langle \boldsymbol{C}^{\text{row}}(\boldsymbol{P}^{\text{row}})^T,\boldsymbol{X}^T\rangle + \langle \boldsymbol{C}^{\text{col}}(\boldsymbol{P}^{\text{col}})^T,\boldsymbol{X}\rangle,\end{aligned}$$

   Without loss of generality, we could set these two scaling factors to be 1.
   (Otherwise, absorb the scaling factors into C)

   where $\boldsymbol{C}^{\text{row}} = (\sum_{s',j}\boldsymbol{P}^{\text{col}}_{s',j})(\boldsymbol{c}_1^{\text{row}},\ldots,\boldsymbol{c}_{d_1}^{\text{row}})$ and$\boldsymbol{C}^{\text{col}} = (\sum_{s,i}\boldsymbol{P}^{\text{row}}_{s,i})(\boldsymbol{c}_1^{\text{col}},\ldots,\boldsymbol{c}_{d_2}^{\text{col}})$.

   In both cases, f is successfully reduced down to linear case decision function with low-rank $2r$. But concatenated mapping has consistent formula whereas bilinear mapping does not.

Remark 1. Notice that in both cases, the rank of coefficient becomes $2r$.

$$f(\boldsymbol{X}) = \langle \boldsymbol{C}_1\boldsymbol{P}_1^T, \boldsymbol{X}\rangle + \langle \boldsymbol{C}_2\boldsymbol{P}_2^T, \boldsymbol{X}^T\rangle \text{ where } \boldsymbol{C}_i \in \mathbb{R}^{d_1\times r}, \boldsymbol{P}_i \in \mathbb{R}^{d_2\times r} \text{ for } i = 1, 2.$$
$$= \langle \boldsymbol{C}_3\boldsymbol{P}_3, \boldsymbol{X}\rangle \text{ where } \boldsymbol{C}_3 \in \mathbb{R}^{d_1\times 2r}, \boldsymbol{P}_3 \in \mathbb{R}^{d_2\times 2r}.$$

why 2r? Can we set C3 = 2C1 (= 2P2) and set P3 = (P1= C2)?

We go back to same argument about difference of decision rule when we use $\boldsymbol{X}$ as input and $\tilde{\boldsymbol{X}} = \begin{pmatrix} 0 & \boldsymbol{X} \\ \boldsymbol{X}^T & 0 \end{pmatrix}$ in the notes 051820*.pdfs. But I found out previous argument (algorithm outputs are the same with input ($\tilde{\boldsymbol{X}}$,$2r$) and ($\boldsymbol{X}$, r)) is wrong because main inequality of the problem was

$$\left\| \begin{pmatrix} 0 & \boldsymbol{B}_1 \\ \boldsymbol{B}_2^T & 0 \end{pmatrix} \right\|^2 = \|\boldsymbol{B}_1\|^2 + \|\boldsymbol{B}_2\|^2 \geq \|\boldsymbol{B}_1 + \boldsymbol{B}_2\|^2 = \|\boldsymbol{B}\|^2.$$

and the equality condition was $\boldsymbol{B}_1 = \boldsymbol{B}_2$. However, $\boldsymbol{B}$ becomes less than rank $2r$ not less than r. Therefore, $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ is the same but under the different rank constraint. The argument should be changed to algorithm outputs are the same with input ($\tilde{\boldsymbol{X}}$,$2r$) and ($\boldsymbol{X}$, $2r$). And also this new argument does make sense considering the number of free parameters.

2. Reduction to vector case: Let $\boldsymbol{x} \in \mathbb{R}^{d_1}$

   In concatenated mapping case, vector feature mapping is

$$\begin{aligned} \Phi_{\text{con}} : \mathbb{R}^{d_1} &\to \mathcal{H}_1 \\ \boldsymbol{x} &\mapsto \phi(\boldsymbol{x}). \end{aligned}$$

   and $f(\boldsymbol{x}) = \langle \boldsymbol{b}, \phi(\boldsymbol{x})\rangle$ where $\boldsymbol{b} \in \mathcal{H}_1$.

   In bilinear mapping case, vector feature mapping is

$$\begin{aligned} \Phi_{\text{bi}} : \mathbb{R}^{d_1} &\to (\mathcal{H}_1)^{d_1} \\ \boldsymbol{x} &\mapsto (\phi(\boldsymbol{x}), \ldots, \phi(\boldsymbol{x}))^T. \end{aligned}$$

   and $f(\boldsymbol{x}) = \langle \boldsymbol{b}, \Phi_{\text{bi}}(\boldsymbol{x})\rangle = d_1\langle \boldsymbol{b}_1, \phi(\boldsymbol{x})\rangle$ where $\boldsymbol{b} = [\![\boldsymbol{b}_i]\!] \in \mathcal{H}_1^{d_1}$ with $\boldsymbol{b}_i = \boldsymbol{b}_1 \in \mathcal{H}_1$.

   Notice two functions are the same upto constant multiplication.

3. Generalization to tensor case: Let $\mathcal{X} \in \mathbb{R}^{d_1\times d_2\times d_3}$ (We can just comment the formulas for intuition and work in the next project) Agree. could be a future project on its own. Seems some deep connection with nonlinear Tucker model

   In concatenated mapping case, we define tensor feature mapping as

$$\begin{aligned} \Phi_{\text{con}} : \mathbb{R}^{d_1\times d_2\times d_3} &\to \mathcal{H}_1^{d_2d_3} \times \mathcal{H}_2^{d_1d_3} \times \mathcal{H}_3^{d_1d_2} \\ \mathcal{X} &\mapsto (\Phi_1(\mathcal{X}), \Phi_2(\mathcal{X}), \Phi_3(\mathcal{X})) \end{aligned}$$

   where $\Phi_1(\mathcal{X}) \stackrel{\text{def}}{=} [\phi_1(\mathcal{X}_{:jk})] \in \mathcal{H}_1^{d_2d_3}$, $\Phi_2(\mathcal{X}) \stackrel{\text{def}}{=} [\phi_2(\mathcal{X}_{i:k})] \in \mathcal{H}_2^{d_1d_3}$ and $\Phi_3(\mathcal{X}) \stackrel{\text{def}}{=} [\phi_3(\mathcal{X}_{ij:})] \in \mathcal{H}_3^{d_1d_2}$.

   and $f(\mathcal{X}) = \langle \boldsymbol{C}_1\boldsymbol{P}_1^T, \Phi_1(\mathcal{X})\rangle + \langle \boldsymbol{C}_2\boldsymbol{P}_2^T, \Phi_2(\mathcal{X})\rangle + \langle \boldsymbol{C}_3\boldsymbol{P}_3^T, \Phi_3(\mathcal{X})\rangle$, of which form I guess has to do with Tucker decomposition (derivation of Tucker decomposition requires unfolding tensor and performs SVD for each mode). Here what we really estimate is $\boldsymbol{P}_1, \boldsymbol{P}_2, \boldsymbol{P}_3$ and coefficients $\boldsymbol{\alpha}$.

In bilinear mapping case, we define tensor feature mapping as

$$\begin{aligned} \Phi_{\mathrm{bi}} \colon \mathbb{R}^{d_1 \times d_2 \times d_3} &\to (\mathcal{H}_1 \times \mathcal{H}_2 \times \mathcal{H}_3)^{d_1 \times d_2 \times d_3} \\ \mathcal{X} &\mapsto [\![\Phi_{\mathrm{bi}}(\mathcal{X})_{ijk}]\!] \in (\mathcal{H}_1 \times \mathcal{H}_2 \times \mathcal{H}_3)^{d_1 \times d_2 \times d_3} \\ &\text{where } \Phi_{\mathrm{bi}}(\mathcal{X})_{ijk} = (\phi_1(\mathcal{X}_{:jk}), \phi_2(\mathcal{X}_{i:k}), \phi_3(\mathcal{X}_{:jk})) \end{aligned}$$

and $f(\mathcal{X}) = \langle \boldsymbol{C} \times_1 \boldsymbol{P}_1 \times_2 \boldsymbol{P}_2 \times_3 \boldsymbol{P}_3, \Phi_{\mathrm{bi}}(\mathcal{X}) \rangle$. where $\boldsymbol{C} = [\![c_{ijk}]\!] \in (\mathcal{H}_1 \times \mathcal{H}_2 \times \mathcal{H}_3)^{r_1 \times r_2 \times r_3}$ with $c_{ijk} = (\boldsymbol{c}_i^{(1)}, \boldsymbol{c}_j^{(2)}, \boldsymbol{c}_k^{(3)}) \in \mathcal{H}_1 \times \mathcal{H}_2 \times \mathcal{H}_3$

I guess that those two functions are also equivalent as in vector and matrix case.