Classification algorithm based on matrix kernels

Miaoyan Wang, Aug 18, 2020

1 Notation

1. $\mathbb{O}(d,r) := \{\boldsymbol{P} \in \mathbb{R}^{d\times r} \colon \boldsymbol{P}^T\boldsymbol{P} = \boldsymbol{I}_r\}$, the collection of d-by-r matrices whose columns are orthonormal. When no confusion arises, I use the term "projection matrix" to denote either the matrix $\boldsymbol{P}\boldsymbol{P}^T \in \mathbb{R}^{d\times d}$ or the matrix $\boldsymbol{P} \in \mathbb{R}^{d\times r}$.
2. $\mathcal{K}^{\text{row}}(i,j,\boldsymbol{X},\boldsymbol{X}') := \langle \Phi(\boldsymbol{X}_{i:}),\ \Phi(\boldsymbol{X}'_{j:})\rangle$ denotes the value of row kernel evaluated at the vector pair, (i-th row of matrix $\boldsymbol{X}$, j-th row of matrix $\boldsymbol{X}'$).
3. I sometimes use the shorthand $\mathcal{K}^{\text{row}}(i,j)$ to denote $\mathcal{K}^{\text{row}}(i,j,\boldsymbol{X},\boldsymbol{X}')$, when the feature pair $(\boldsymbol{X},\boldsymbol{X}')$ is clear given the contexts. Note that $\mathcal{K}^{\text{row}}(i,j)$ can be calcualted without explicit feature mapping. Similar convention for $\mathcal{K}^{\text{col}}(i,j,\boldsymbol{X},\boldsymbol{X}')$.
4. Let $\boldsymbol{W}^{\text{row}} = \boldsymbol{P}_r\boldsymbol{P}_r^T = [\![w_{ij}^{\text{row}}]\!] \in \mathbb{R}^{d_1\times d_1}$ and $\boldsymbol{W}^{\text{col}} = \boldsymbol{P}_c\boldsymbol{P}_c^T = [\![w_{ij}^{\text{col}}]\!] \in \mathbb{R}^{d_2\times d_2}$ denote the row- and column-wise projection matrices, respectively.

2 Algorithm based on bilinear maps

Consider the bilinear mapping,

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

Primal problem:

$$\boxed{\begin{aligned}&\min_{\boldsymbol{P}_r,\boldsymbol{P}_c}\min_{\boldsymbol{C}} \quad \frac{1}{2}\|\boldsymbol{C}\|_F^2 + c\sum_{i=1}^n \xi_i,\\ &\text{subject to} \quad y_i\langle \boldsymbol{P}_r\boldsymbol{C}\boldsymbol{P}_c^T,\ \Phi(\boldsymbol{X}_i)\rangle \leq 1-\xi_i \text{ and } \xi_i \geq 0,\ i=1,\ldots,n.\end{aligned}} \tag{1}$$

Parameters in the primal problem: $(\boldsymbol{P}_r,\boldsymbol{P}_c,\boldsymbol{C})$, where $\boldsymbol{P}_r \in \mathbb{O}(d_1,r_1)$, $\boldsymbol{P}_c \in \mathbb{O}(d_2,r_2)$, and $\boldsymbol{C} = [\![(\boldsymbol{c}_i^{\text{row}},\ \boldsymbol{c}_j^{\text{col}})]\!] \in (\mathcal{H}_r\times\mathcal{H}_c)^{r_1\times r_2}$ is the "core matrix" consisting of linear coefficients.

1. Update $\boldsymbol{C}$, given $(\boldsymbol{P}_r,\boldsymbol{P}_c)$.

$$\boxed{\text{implicit update } \boldsymbol{C} \leftarrow \sum_i \alpha_i y_i \boldsymbol{P}_r^T\Phi(\boldsymbol{X}_i)\boldsymbol{P}_c.}$$

Saved quantities: dual variables $\boldsymbol{\alpha} \in \mathbb{R}^n$.
Intermediate quantities: kernel $\mathcal{K}(\boldsymbol{X}, \boldsymbol{X}')$ specified below.
Objective value: objective value for the dual problem.

Details: We use kernel trick to solve for $\boldsymbol{\alpha}$ without explicit feature mapping. Given the projections $(\boldsymbol{P}_r, \boldsymbol{P}_c)$, the optimization (1) is a standard SVM with kernel

$$\begin{aligned}\mathcal{K}(\boldsymbol{X}, \boldsymbol{X}') &= \langle \boldsymbol{P}_r^T \Phi(\boldsymbol{X}) \boldsymbol{P}_c,\ \boldsymbol{P}_r^T \Phi(\boldsymbol{X}') \boldsymbol{P}_c \rangle \\ &= (\sum_{i,j} w_{ij}^{\text{col}})(\sum_{i,j} w_{ij}^{\text{row}} K^{\text{row}}(i,j)) + (\sum_{i,j} w_{ij}^{\text{row}})(\sum_{i,j} w_{ij}^{\text{col}} K^{\text{col}}(i,j)).\end{aligned} \tag{2}$$

for all feature pairs $(\boldsymbol{X}, \boldsymbol{X}')$. Here I have used the shorthand $K^{\text{row}}(i,j)$ to denote the value of row kernel evaluated on the i-th row of $\boldsymbol{X}$ and j-th row of $\boldsymbol{X}'$.

Remark 1 (Computational consideration)**.** We can compute the summations in (2) without explicit loop. In particular, both identities hold: $\sum_{i,j} w_{ij}^{\text{col}} = \|\boldsymbol{1}^T \boldsymbol{P}_c\|_2^2$ and $\sum_{i,j} w_{ij}^{\text{row}} K^{\text{row}}(i,j) = \text{trace}(\boldsymbol{W}^T \boldsymbol{K})$, where $\boldsymbol{K} \leftarrow [\![K^{\text{row}}(i,j,\boldsymbol{X},\boldsymbol{X}')]\!]$ is a pre-stored matrix (or array, if we go through all possible feature pairs $(\boldsymbol{X}, \boldsymbol{X}')$).

2. Update $\boldsymbol{P}_r$, given $(\boldsymbol{C}, \boldsymbol{P}_c)$.

$$\boxed{\begin{aligned}&\text{explicit update } \tilde{\boldsymbol{P}}_r^{\text{new}} \leftarrow \sum_i \beta_i y_i \Phi(\boldsymbol{X}_i)\boldsymbol{P}_c \boldsymbol{C}^T = \sum_{i,j} \beta_i \alpha_j y_i y_j \underbrace{\Phi(\boldsymbol{X}_i)\boldsymbol{P}_c\boldsymbol{P}_c^T \Phi^T(\boldsymbol{X}_j)}_{d_1\text{-by-}d_1 \text{ matrix over } \mathbb{R}} \boldsymbol{P}_r \\ &\qquad \text{normalize } \boldsymbol{P}_r^{\text{new}} \leftarrow \text{QR decomposition of } \tilde{\boldsymbol{P}}_r^{\text{new}}.\end{aligned}}$$

Saved quantities: $\boldsymbol{P}^{\text{new}} \in \mathbb{O}(d_1, r_1)$.
Intermediate quantities: matrix $\Phi(\boldsymbol{X}_i)\boldsymbol{P}_c\boldsymbol{P}_c^T\Phi^T(\boldsymbol{X}_j)$ and its trace, dual variables $\boldsymbol{\beta} \in \mathbb{R}^n$.
Objective value: objective value for the dual problem.

Details: for each feature pair $(i,j) \in [n]^2$, we compute the matrix $\Phi(\boldsymbol{X}_i)\boldsymbol{P}_c\boldsymbol{P}_c^T\Phi^T(\boldsymbol{X}_j)$ without explicit feature mapping,

$$\begin{aligned}\Phi(\boldsymbol{X}_i)\boldsymbol{P}_c\boldsymbol{P}_c^T\Phi^T(\boldsymbol{X}_j) = (\sum_{s,s'} w_{ss'}^{\text{col}}) &\begin{bmatrix} K^{\text{row}}(1,1,\boldsymbol{X}_i,\boldsymbol{X}_j) & \cdots & K^{\text{row}}(1,d_1,\boldsymbol{X}_i,\boldsymbol{X}_j) \\ \vdots & \vdots & \vdots \\ K^{\text{row}}(d_1,1,\boldsymbol{X}_i,\boldsymbol{X}_j) & \cdots & K^{\text{row}}(d_1,d_1,\boldsymbol{X}_i,\boldsymbol{X}_j) \end{bmatrix} + \\ &\left(\sum_{s,s'} w_{ss'}^{\text{col}} K^{\text{col}}(s,s',\boldsymbol{X}_i,\boldsymbol{X}_j)\right) \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix},\end{aligned} \tag{3}$$

where $K^{\text{row}}(s,s',\boldsymbol{X}_i,\boldsymbol{X}_j)$ denotes the value of row kernel value evaluated on the s-th row of $\boldsymbol{X}_i$ and s'-th row of $\boldsymbol{X}_j$, and likewise for $K^{\text{col}}(s,s',\boldsymbol{X}_i,\boldsymbol{X}_j)$.

The coefficient $\boldsymbol{\beta}$ is obtained from a standard SVM with kernel

$$\mathcal{K}(\boldsymbol{X}, \boldsymbol{X}') = \text{trace}(\Phi(\boldsymbol{X})\boldsymbol{P}_c\boldsymbol{P}_c^T\Phi^T(\boldsymbol{X}')) = \text{trace of matrix specified in (3)},$$

for all feature pairs $(\boldsymbol{X}, \boldsymbol{X}')$.

3. Update $\boldsymbol{P}_c$, given $(\boldsymbol{C}, \boldsymbol{P}_r)$.

$$\boxed{\begin{aligned} &\text{explicitle update } \boldsymbol{P}_c^{\text{new}} \leftarrow \sum_i \gamma_i y_i \Phi^T(\boldsymbol{X}_i)\boldsymbol{P}_r\boldsymbol{C} = \sum_{i,j} \gamma_i\alpha_j y_i y_j \underbrace{\Phi^T(\boldsymbol{X}_i)\boldsymbol{P}_r\boldsymbol{P}_r^T\Phi(\boldsymbol{X}_j)}_{d_2\text{-by-}d_2 \text{ matrix over } \mathbb{R}}\boldsymbol{P}_c \\ &\qquad\text{normalize } \boldsymbol{P}_c^{\text{new}} \leftarrow \text{QR decomposition of } \tilde{\boldsymbol{P}}_c^{\text{new}}. \end{aligned}}$$

The intermediate quantities, $\Phi^T(\boldsymbol{X}_i)\boldsymbol{P}_r\boldsymbol{P}_r^T\Phi(\boldsymbol{X}_j)$ and $\boldsymbol{\gamma}$, are calculated similarly as in step 2.

3 Outputs

1. Convergence criterum? Objective value in the dual problem.

2. How to read off the decision function from the algorithm?

$$\begin{aligned} f(\boldsymbol{X}_{\text{new}}) &= \langle \boldsymbol{P}_r^T\Phi(\boldsymbol{X}_{\text{new}})\boldsymbol{P}_c, \ \sum_i \alpha_i y_i \boldsymbol{P}_r^T\Phi(\boldsymbol{X}_i)\boldsymbol{P}_c\rangle \\ &= \sum_i \alpha_i y_i \left\{ \left(\sum_{s,s'} w_{ss'}^{\text{col}}\right)\left(\sum_{s,s'} w_{ss'}^{\text{row}} K^{\text{row}}(s, s', \boldsymbol{X}_i, \boldsymbol{X}_{\text{new}})\right) + \right. \\ &\qquad \left. \left(\sum_{s,s'} w_{ss'}^{\text{row}}\right)\left(\sum_{s,s'} w_{ss'}^{\text{col}} K^{\text{col}}(s, s', \boldsymbol{X}_i, \boldsymbol{X}_{\text{new}})\right)\right\}. \end{aligned}$$

3. How to estimate the intercept in the primal problem?

$$\hat{b}_0 = \arg\min_{b_0\in\mathbb{R}} \left\{ \sum_{i=1}^n (1 - y_i f(\boldsymbol{X}_i) - y_i b_0)_+ \right\}.$$

4. The objective value in the primal problem? The primal objective is $\frac{1}{2}\|\boldsymbol{C}\|_F^2 + c\sum_{i=1}^n(1 - y_i f(\boldsymbol{X}_i) - y_i b_0)_+$, where

$$\|\boldsymbol{C}\|_F^2 = \sum_{i,j} \alpha_i\alpha_j y_i y_j \text{trace}(\Phi(\boldsymbol{X}_i)\boldsymbol{P}_c\boldsymbol{P}_c^T\Phi^T(\boldsymbol{X}_j)\boldsymbol{P}_r\boldsymbol{P}_r^T).$$

We omit the explicit expression of $\boldsymbol{C}$ because it is not needed in the algorithm.