

Classification algorithm with matrix kernels

Miaoyan Wang, Aug 17, 2020

Notation:

1. $\mathbb{O}(d, r) := \{\mathbf{P} \in \mathbb{R}^{d \times r} : \mathbf{P}^T \mathbf{P} = \mathbf{I}_r\}$, 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 $\mathbf{P}\mathbf{P}^T \in \mathbb{R}^{d \times d}$ or the matrix $\mathbf{P} \in \mathbb{R}^{d \times r}$.
2. $\mathcal{K}^{\text{row}}(i, j, \mathbf{X}, \mathbf{X}') := \langle \Phi(\mathbf{X}_{i:}), \Phi(\mathbf{X}'_{j:}) \rangle$ denotes the value of row kernel evaluated at the vector pair, (i -th row of matrix \mathbf{X} , j -th row of matrix \mathbf{X}').
3. I sometimes use the shorthand $\mathcal{K}^{\text{row}}(i, j)$ to denote $\mathcal{K}^{\text{row}}(i, j, \mathbf{X}, \mathbf{X}')$, when the feature pair $(\mathbf{X}, \mathbf{X}')$ is clear given the contexts. Note that $\mathcal{K}^{\text{row}}(i, j)$ can be calculated without explicit feature mapping.
4. Similar convention for $\mathcal{K}^{\text{col}}(i, j, \mathbf{X}, \mathbf{X}')$.

1 Optimization formulation with bilinear mapping

Consider the bilinear mapping,

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

Primal problem:

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

Parameters in the primal problem: $(\mathbf{P}_r, \mathbf{P}_c, \mathbf{C})$, where $\mathbf{P}_r \in \mathbb{O}(d_1, r_1)$, $\mathbf{P}_c \in \mathbb{O}(d_2, r_2)$, and $\mathbf{C} = \llbracket (\mathbf{c}_i^{\text{row}}, \mathbf{c}_j^{\text{col}}) \rrbracket \in (\mathcal{H}_r \times \mathcal{H}_c)^{r_1 \times r_2}$ is the low-rank “core matrix” consisting of linear coefficients.

The equivalent dual problem for (1) is

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

The optimization (2) is further equivalent to

$$\begin{aligned}
& \max_{\mathbf{P}_r, \mathbf{P}_c} \min_{\boldsymbol{\alpha}} \quad - \sum_{i=1}^n \alpha_i + \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \langle \mathbf{P}_r^T \Phi(\mathbf{X}_i) \mathbf{P}_c, \mathbf{P}_r^T \Phi(\mathbf{X}_j) \mathbf{P}_c \rangle, \\
& \text{subject to} \quad \sum_i y_i \alpha_i = 0, \text{ and } 0 \leq \alpha_i \leq c, \ i = 1, \dots, n, \\
& \quad \mathbf{P}_r \in \mathbb{O}(d_1, r_1), \ \mathbf{P}_c \in \mathbb{O}(d_2, r_2).
\end{aligned} \tag{3}$$

Our goal is to solve (3). The unknown parameters are $(\mathbf{P}_r, \mathbf{P}_c, \boldsymbol{\alpha})$.

2 Algorithm for problem (3)

1. Update $\boldsymbol{\alpha}$, while holding $(\mathbf{P}_r, \mathbf{P}_c)$ fixed.

Preparation: Let $\mathbf{W}^{\text{row}} = \mathbf{P}_r \mathbf{P}_r^T = \llbracket w_{ij}^{\text{row}} \rrbracket \in \mathbb{R}^{d_1 \times d_1}$ and $\mathbf{W}^{\text{col}} = \mathbf{P}_c \mathbf{P}_c^T = \llbracket w_{ij}^{\text{col}} \rrbracket \in \mathbb{R}^{d_2 \times d_2}$ denote the row- and column-wise projection matrices, respectively.

We use kernel trick to solve for $\boldsymbol{\alpha}$ without explicit feature mapping. Given the projections $(\mathbf{P}_r, \mathbf{P}_c)$, the optimization (3) is a standard SVM with kernel $\mathcal{K}(\mathbf{X}, \mathbf{X}')$ defined as follows,

$$\begin{aligned}
\mathcal{K}(\mathbf{X}, \mathbf{X}') &= \langle \mathbf{P}_r^T \Phi(\mathbf{X}) \mathbf{P}_c, \mathbf{P}_r^T \Phi(\mathbf{X}') \mathbf{P}_c \rangle \\
&= \left(\sum_{i,j} w_{ij}^{\text{col}} \right) \left(\sum_{i,j} w_{ij}^{\text{row}} K^{\text{row}}(i, j) \right) + \left(\sum_{i,j} w_{ij}^{\text{row}} \right) \left(\sum_{i,j} w_{ij}^{\text{col}} K^{\text{col}}(i, j) \right).
\end{aligned} \tag{4}$$

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 \mathbf{X} and j -th row of \mathbf{X}' .

Remark 1 (Computational consideration). We can compute the summations in (4) without explicit loop. In particular, both identities hold: $\sum_{i,j} w_{ij}^{\text{col}} = \|\mathbf{1}^T \mathbf{P}_c\|_2^2$ and $\sum_{i,j} w_{ij}^{\text{row}} K^{\text{row}}(i, j) = \text{trace}(\mathbf{W}^T \mathbf{K})$, where $\mathbf{K} \leftarrow \llbracket K^{\text{row}}(i, j, \mathbf{X}, \mathbf{X}') \rrbracket$ is a pre-stored matrix (or array, if we go through all possible feature pairs $(\mathbf{X}, \mathbf{X}')$).

2. Update \mathbf{P}_r , while holding $(\boldsymbol{\alpha}, \mathbf{P}_c)$ fixed.

Denote the matrix $\mathbf{M} = \sum_i \alpha_i y_i \Phi(\mathbf{X}_i) \mathbf{P}_c \in (\mathcal{H}_1 \times \mathcal{H}_2)^{d_1 \times r_2}$. The problem (3) reduces to

$$\begin{aligned}
& \max_{\mathbf{P}_r \in \mathbb{O}(d_1, r_1)} \sum_{i,j} \alpha_i \alpha_j y_i y_j \langle \mathbf{P}_r^T \Phi(\mathbf{X}_i) \mathbf{P}_c, \mathbf{P}_r^T \Phi(\mathbf{X}_j) \mathbf{P}_c \rangle \\
&= \max_{\mathbf{P}_r \in \mathbb{O}(d_1, r_1)} \langle \mathbf{P}_r^T \mathbf{M}, \mathbf{P}_r^T \mathbf{M} \rangle \\
&= \max_{\mathbf{P}_r \in \mathbb{O}(d_1, r_1)} \left\langle \underbrace{\mathbf{P}_r \mathbf{P}_r^T}_{\text{rank-}r_1 \text{ projection}}, \underbrace{\mathbf{M} \mathbf{M}^T}_{\substack{d_1\text{-by-}d_1 \text{ p.s.d. matrix over } \mathbb{R}}} \right\rangle.
\end{aligned} \tag{5}$$

By the property of low-rank projection (c.f. Lemma 1), the optimization in the last line has a closed-form solution,

$$\mathbf{P}_r \leftarrow \text{top } r_1 \text{ eigenvectors of the matrix } \mathbf{M}\mathbf{M}^T.$$

It remains to compute the matrix $\mathbf{M}\mathbf{M}^T$ without explicit feature mapping. Write

$$\begin{aligned} \mathbf{M}\mathbf{M}^T &= \left(\sum_i \alpha_i y_i \Phi(\mathbf{X}_i) \mathbf{P}_c \right) \left(\sum_i \alpha_i y_i \Phi(\mathbf{X}_i) \mathbf{P}_c \right)^T \\ &= \sum_{i,j} \alpha_i \alpha_j y_i y_j \underbrace{\Phi(\mathbf{X}_i) \mathbf{P}_c \mathbf{P}_c^T \Phi^T(\mathbf{X}_j)}_{d_1\text{-by-}d_1 \text{ matrix over } \mathbb{R}}. \end{aligned} \quad (6)$$

The summand (6) involves the matrix of the type $\Phi(\mathbf{X}_i) \mathbf{P}_c \mathbf{P}_c^T \Phi^T(\mathbf{X}_j)$, for all feature pairs $(i, j) \in [n]^2$. Each of these matrices can be obtained without explicit feature mapping,

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

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

3. Update \mathbf{P}_c , while holding $(\boldsymbol{\alpha}, \mathbf{P}_r)$ fixed. Similar as step 2 but switching the role of rows and columns.

Lemma 1 (Best rank- r projection). *Let $\mathbf{A} \in \mathbb{R}^{d \times d}$ be a positive semi-definite matrix. Let $(\lambda_i, \mathbf{p}_i) \in \mathbb{R} \times \mathbb{R}^d$ denote the i -th singular-value-singularvector pair of \mathbf{A} , and assume that eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d \geq 0$ are sorted in non-increasing order. Consider an optimization problem specified as*

$$\max_{\mathbf{P} \in \mathbb{O}(d, r)} f(\mathbf{P}), \quad \text{where } f(\mathbf{P}) = \langle \mathbf{P}\mathbf{P}^T, \mathbf{A} \rangle.$$

Then, the leading rank- r singular space of \mathbf{A} , denoted $\mathbf{P}^ = \text{Span}(\mathbf{p}_1, \dots, \mathbf{p}_r)$, optimizes the objective $f(\mathbf{P})$. In particular, $f(\mathbf{P}^*) = \sum_{i=1}^r \lambda_i(\mathbf{A})$.*

Proof. The positive semi-definiteness of \mathbf{A} implies the existence of a symmetric matrix $\mathbf{B} \in \mathbb{R}^{d \times d}$ such that $\mathbf{A} = \mathbf{B}^2$. Furthermore, the singular values satisfy $\lambda_i^2(\mathbf{B}) = \lambda_i(\mathbf{A})$ for all $i \in [d]$. Notice

that

$$f(\mathbf{P}) = \langle \mathbf{P}\mathbf{P}^T, \mathbf{B}^2 \rangle = \|\mathbf{B}\|_F^2 - \underbrace{\|\mathbf{B}(\mathbf{I} - \mathbf{P}\mathbf{P}^T)\|_F^2}_{\text{rank-}(d-r) \text{ approximation of } \mathbf{B}} \leq \sum_{i=1}^r \lambda_i^2(\mathbf{B})$$

holds for all matrices $\mathbf{P} \in \mathbb{O}(d, r)$. Therefore,

$$\max_{\mathbf{P} \in \mathbb{O}(d, r)} f(\mathbf{P}) \leq \sum_{i=1}^r \lambda_i(\mathbf{A}),$$

where equality is attained if $\mathbf{P} = \text{Span}(\mathbf{p}_1, \dots, \mathbf{p}_r)$. □

3 Outputs

How to read off the decision function from the algorithm outputs?

$$\begin{aligned} f(\mathbf{X}_{\text{new}}) &= \langle \mathbf{P}_r^T \Phi(\mathbf{X}_{\text{new}}) \mathbf{P}_c, \sum_i \alpha_i y_i \mathbf{P}_r^T \Phi(\mathbf{X}_i) \mathbf{P}_c \rangle \\ &= \langle \Phi(\mathbf{X}_{\text{new}}), \underbrace{\mathbf{P}_r \mathbf{P}_r^T \left(\sum_i \alpha_i y_i \Phi(\mathbf{X}_i) \right) \mathbf{P}_c \mathbf{P}_c^T}_{\text{core tensor } \mathbf{C} \text{ in the primal problem}} \rangle \\ &= \sum_i \alpha_i y_i \left\{ \left(\sum_{s, s'} w_{ss'}^{\text{col}} \right) \left(\sum_{s, s'} w_{ss'}^{\text{row}} K^{\text{row}}(s, s', \mathbf{X}_i, \mathbf{X}_{\text{new}}) \right) + \right. \\ &\quad \left. \left(\sum_{s, s'} w_{ss'}^{\text{row}} \right) \left(\sum_{s, s'} w_{ss'}^{\text{col}} K^{\text{col}}(s, s', \mathbf{X}_i, \mathbf{X}_{\text{new}}) \right) \right\}. \end{aligned} \quad (7)$$

How to estimate the intercept in the primal problem?

$$\hat{b}_0 = \arg \min_{b_0 \in \mathbb{R}} \left\{ \frac{1}{2} \|\mathbf{C}\|_F^2 + c \sum_{i=1}^n (1 - y_i f(\mathbf{X}_i) - y_i b_0)_+ \right\},$$

where $\|\mathbf{C}\|_F^2 = \sum_{i,j} \alpha_i \alpha_j y_i y_j \langle \mathbf{P}_r^T \Phi(\mathbf{X}_i) \mathbf{P}_c, \mathbf{P}_r^T \Phi(\mathbf{X}_j) \mathbf{P}_c \rangle = \sum_{i=1}^r \lambda_i(\mathbf{M}\mathbf{M}^T)$, and $\lambda_i(\cdot)$ denotes the i -th eigenvalue of the matrix. The formula for $\|\mathbf{C}\|_F^2$ follows from the second line of (7) and the optimization (5).

4 Further thoughts

The dual optimization (3) yields a neater algorithm than previous approaches. Recall that, in the note *0620.pdf, we have derived the alternating optimization algorithm for the primal prob-

lem,

$$\begin{aligned}
& \min_{\mathbf{B}} \quad \frac{1}{2} \|\mathbf{B}\|_F^2 + c \sum_{i=1}^n \xi_i, \\
& \text{subject to} \quad y_i \langle \mathbf{B}, \Phi(\mathbf{X}_i) \rangle \leq 1 - \xi_i \text{ and } \xi_i \geq 0, i = 1, \dots, n, \\
& \quad \text{where } \mathbf{B} = \mathbf{C}\mathbf{P}^T = (\mathbf{C}_r, \mathbf{C}_c)(\mathbf{P}_r, \mathbf{P}_c)^T, \mathbf{P}_r \in \mathbb{O}(d_1, r), \mathbf{P}_c \in \mathbb{O}(d_2, r).
\end{aligned} \tag{8}$$

The block variable \mathbf{P} has explicit update, whereas the other block \mathbf{C} has implicit update. Here we give a different perspective on the algorithm derivation. Notice that the primal problem (8) is equivalent to the dual problem,

$$\begin{aligned}
& \max_{\mathbf{P}} \min_{\boldsymbol{\alpha}} \quad - \sum_{i=1}^n \alpha_i + \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \langle \Phi(\mathbf{X}_i) \mathbf{P}, \Phi(\mathbf{X}_j) \mathbf{P} \rangle, \\
& \text{subject to} \quad \sum_i y_i \alpha_i = 0, \text{ and } 0 \leq \alpha_i \leq c, i = 1, \dots, n, \\
& \quad \mathbf{P} = (\mathbf{P}_r, \mathbf{P}_c) \in \mathbb{O}(d_1, r_1) \times \mathbb{O}(d_2, r_2).
\end{aligned} \tag{9}$$

(For notational convenience, I will drop the column-wise projection \mathbf{P}_c and consider row-wise projection \mathbf{P}_r only. In such a case, $\Phi(\mathbf{X}) \in \mathcal{H}^{d_1}$.)

Algorithm for optimization (9) over parameters $(\mathbf{P}, \boldsymbol{\alpha})$.

1. Update $\boldsymbol{\alpha}$ holding \mathbf{P} fixed. \implies same as in the note *0620.pdf.
2. Update \mathbf{P} holding $\boldsymbol{\alpha}$ fixed.

$$\begin{aligned}
\mathbf{P} & \leftarrow \arg \max_{\mathbf{P} \in \mathbb{O}(d, r)} \sum_{i,j} \alpha_i \alpha_j y_i y_j \langle \Phi(\mathbf{X}_i) \mathbf{P}, \Phi(\mathbf{X}_j) \mathbf{P} \rangle \\
& \stackrel{\text{c.f. Lemma 1}}{=} \text{top } r \text{ singular vectors of matrix } \mathbf{B}\mathbf{B}^T, \quad \text{where } \mathbf{B} = \underbrace{\sum_{i=1}^n \alpha_i y_i \Phi(\mathbf{X}_i)}_{\mathcal{H}^{d_1}}.
\end{aligned}$$

Notice that $\mathbf{B}\mathbf{B}^T$ can be obtained without explicit feature mapping,

$$\mathbf{B}\mathbf{B}^T = \left(\sum_{i=1}^n \alpha_i y_i \Phi(\mathbf{X}_i) \right) \left(\sum_{i=1}^n \alpha_i y_i \Phi(\mathbf{X}_i) \right)^T = \sum_{i,j} \alpha_i \alpha_j y_i y_j \Phi(\mathbf{X}_i) \Phi^T(\mathbf{X}_j).$$

As a by-product, the dual formulation (9) also justifies the same treatment to coefficients $\boldsymbol{\alpha}, \boldsymbol{\beta}$ in the previous algorithm.