Possible formulation for Kernel SMM (updated on 04/12/2020)

Let $\boldsymbol{U} \in \mathbb{R}^{m\times r}, \boldsymbol{V} \in \mathbb{R}^{n\times r}$ denote the factor matrices of interest in the low-rank kernel.

- Dual problem

$$(D) \quad f(\boldsymbol{U},\boldsymbol{V}) := \max_{\boldsymbol{\alpha}\geq 0} \left\{ \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i\alpha_j y_i y_j K(\boldsymbol{X}_i, \boldsymbol{X}_j) \right\}, \tag{1}$$

$$\text{subject to} \quad \sum_{i=1}^{N} y_i\alpha_i = 0, \quad 0 \leq \alpha_i \leq C, \quad i = 1,\ldots,N.$$

where the Kernel

$$K(\boldsymbol{X}_i, \boldsymbol{X}_j) = \exp\left(-\frac{\|P_{\boldsymbol{U}}\boldsymbol{X}_i P_{\boldsymbol{V}} - P_{\boldsymbol{U}}\boldsymbol{X}_j P_{\boldsymbol{V}}\|_2^2}{\sigma^2}\right),$$

implicitly depends on $\boldsymbol{U}$ and $\boldsymbol{V}$.

- Primal problem:

$$(P) \quad g(\boldsymbol{U},\boldsymbol{V}) := \min_{\boldsymbol{\xi}\in\mathbb{R}^N, \boldsymbol{D}\in\mathbb{R}^{r\times r}, b} \left\{ \|\boldsymbol{D}\|_F^2 + C\sum_{i=1}^{N} \xi_i \right\}, \tag{2}$$

$$\begin{aligned} \text{subject to} \quad & y_i\left(\langle \boldsymbol{D},\ h(P_{\boldsymbol{U}}\boldsymbol{X}_i P_{\boldsymbol{V}})\rangle + b\right) \geq 1 - \xi_i, \\ & \xi_i \geq 0,\ i = 1,\ldots,N. \end{aligned}$$

Remark 1. The duality gap between (1) and (2) is zero for all $(\boldsymbol{U},\boldsymbol{V})$. Therefore,

$$\min_{(\boldsymbol{U},\boldsymbol{V})\in \text{ feasible domain}} g(\boldsymbol{U},\boldsymbol{V}) = \min_{(\boldsymbol{U},\boldsymbol{V})\in \text{ feasible domain}} f(\boldsymbol{U},\boldsymbol{V}).$$

Recall that we have proposed the non-linear SMM problem as $\min_{\boldsymbol{U},\boldsymbol{V}} g(\boldsymbol{U},\boldsymbol{V})$. Equivalently, we seek the solution to the following (strong) dual problem:

$$(D') \quad \min_{\substack{\boldsymbol{\alpha}\geq 0,\\ \boldsymbol{U}\in\mathbb{R}^{m\times r}, \boldsymbol{V}\in\mathbb{R}^{n\times r}}} \left\{ -\sum_{i=1}^{N} \alpha_i + \frac{1}{2}\sum_{i,j} \alpha_i\alpha_j y_i y_j K(\boldsymbol{X}_i, \boldsymbol{X}_j) \right\},$$

$$\text{subject to} \quad \sum_{i=1}^{N} y_i\alpha_i = 0, \quad 0 \leq \alpha_i \leq C, \quad i = 1,\ldots,N.$$

Note that no orthogonality constraints are imposed on $\boldsymbol{U}$ and $\boldsymbol{V}$. This is because only the column spaces of $\boldsymbol{U}$ and $\boldsymbol{V}$ are relevant in the dual problem, and their actual values are non-important.

Remark 2 (Implementation)**.** For linear kernel $K(\boldsymbol{X}_i, \boldsymbol{X}_j) = \langle \boldsymbol{X}_i P_{\boldsymbol{V}}, P_{\boldsymbol{U}}\boldsymbol{X}_j\rangle$, the above optimization can be easily solved using alternating SVM (see earlier notes). How about non-liner kernel? In particular, how to update $\boldsymbol{U}, \boldsymbol{V}$? Gradient descent?

Remark 3 (Rank-1 + vector case)**.** Let $\boldsymbol{X}_i = [\![x_{ip}]\!], \boldsymbol{X}_j = [\![x_{jp}]\!] \in \mathbb{R}^m$ be two vectors, where m is the number of features. Then the rank-1 Gaussian kernel can be represented as

$$K(\boldsymbol{X}_i, \boldsymbol{X}_j) = \exp\left(\frac{-\sum_{p\in[m]} w_p(x_{ip} - x_{jp})^2}{\sigma^2}\right),$$

where $\boldsymbol{U} = (w_1, \dots, w_m)^T \in \mathbb{R}^m_{\geq 0}$ are unknown weights. This kernel was popularly used for the task of feature selection in SVM (add references..).