Rethinking multiple kernel learning under the lenses of stochastic variational inference

Davide Adamo [1,2] & Marco Corneli [1,2]

[1] *Université Côte d'Azur, Inria, CNRS, Laboratoire J.A.Dieudonné, Maasai team, Nice, France.*

[2] *Université Côte d'Azur, Laboratoire CEPAM, Nice, France.*
davide.adamo@univ-cotedazur.fr

Résumé. Les méthodes de noyaux sont largement utilisées dans l'apprentissage automatique car elles constituent un outil puissant pour cartographier implicitement les données dans des espaces à haute dimension, permettant la découverte de modèles complexes qui pourraient être difficiles à capturer dans l'espace de caractéristiques d'origine. Bien que de nombreux problèmes de classification et de régression puissent être résolus avec succès à l'aide d'un seul noyau, il arrive que les ensembles de données du monde réel présentent des structures diverses et qu'il soit nécessaire d'utiliser plusieurs types de noyaux (un pour chaque notion de similarité que l'on souhaite prendre en compte). C'est là que l'apprentissage multi-noyaux (MKL) entre en jeu.
Cet article revisite la classification multi-noyaux en mettant l'accent sur la sélection des noyaux à la lumière des développements récents de l'inférence variationnelle stochastique (SVI). Dans le cadre de la régression logistique à noyaux, nous considérons des combinaisons linéaires semi-définies positives de noyaux et nous traitons les poids des noyaux comme des variables aléatoires. Des choix appropriés de distributions préalables font naturellement émerger une pénalité Lasso, tandis que la puissance du SVI nous permet d'estimer le modèle et les paramètres variationnels dans un contexte entièrement différentiable, ainsi que de construire des intervalles de confiance pour les poids des noyaux. Des exemples numériques illustrent notre approche.

Mots-clés. Apprentissage à noyaux multiples, Inférence variationnelle stochastique, sélection du modèle.

Abstract. Kernel methods are widely employed in machine learning as they are a powerful tool to implicitly map data into high-dimensional spaces, enabling the discovery of complex patterns that might be challenging to capture in the original feature space. Although many classification and regression problems can be successfully attacked with a single kernel, sometimes real-world datasets exhibit diverse structures and employing several kernel types (one for each notion of similarity that we aim to take into account) is necessary. This is where multi-kernel learning (MKL) comes into play.
This paper revisits multi-kernel classification with a specific focus on kernel(s) selection in the light of the recent developments in stochastic variational inference (SVI). In the framework of kernelized logistic regression, we consider positive semi-definite linear combitinations of kernels and we treat the kernel weights as random variables. Proper choices of prior distributions make naturally emerge a Lasso penalty, whereas the power of SVI allows us to

estimate the model and variational parameters in a fully differentiable context as well as to build confidence intervals for the kernel weights. Numerical examples highlight our approach.

Keywords. Multiple kernel learning, Stochastic variational inference, Model selection.

1 Introduction

Kernel methods provide a powerful framework for capturing complex relationships in data by implicitly mapping input features into higher-dimensional spaces, the reproducing kernel Hilbert spaces (RKHS). This transformation both enables linear models to operate effectively despite non-linear feature domains and boost their ability to model intricate patterns and structures. The key ingredient in kernel methods is the kernel (Vapnik, 1999) itself, a symmetric positive semi-definite function that determines the similarity between pairs of data points. Standard kernels are the linear, polynomial, and radial basis function (RBF) one, however an infinite number of kernels can be manufactured and choosing the "right" one(s) is crucial to boost the performance of a classification or regression model. Multiple Kernel Learning (MKL) (Sonnenburg et al., 2006; Gönen and Alpaydın, 2011) emerges as an attempt to both exploit the information coming from different kernels and (possibly) select the relevant kernels for a given machine learning task. This challenge is addressed by allowing the integration of different input kernels and finding an optimal linear or non-linear combination of such kernels.

In (Rakotomamonjy et al., 2008) was developed Simple MKL, using a sub-gradient descent (SD) approach to handle high-dimensional data. This method involves the strategic selection of a subset of pertinent kernels, leading to an improvement in computational efficiency while preserving the advantages associated with integrating multiple kernels. These methods tend to produce sparse kernel combinations ($l1$ penalty on kernel weights) although, being pure optimization any form of inference on the kernel weighs is prevented. In the Bayesian universe, (Girolami and Rogers, 2005) introduced a Bayesian MKL approach for binary classification using hierarchical models and (Damoulas and Girolami, 2008) extended it to a multiclass formulation. In those works, a convex sum of input kernels was enforced using a Dirichlet prior on the kernel weights. Given the high computational cost of training these algorithms, (Gönen, 2012) opted for a fully conjugate Bayesian formulation (BEMKL), involving Gaussian processes. A recent review of the existing techniques that bridge MKL and deep learning approaches can be found in (Wang et al., 2021).

In the light of some (relatively) recent developments in stochastic variational inference (Blei et al., 2017; Naesseth et al., 2017) this paper revisits MKL in the context of supervised classification with the aim to preserve and possibly improve the simplicity of the approach described in (Rakotomamonjy et al., 2008) for the kernel(s) selection, whereas exploiting the versatility of Bayesian approaches, in particular allowing one to perform posterior inference on the kernel weights.

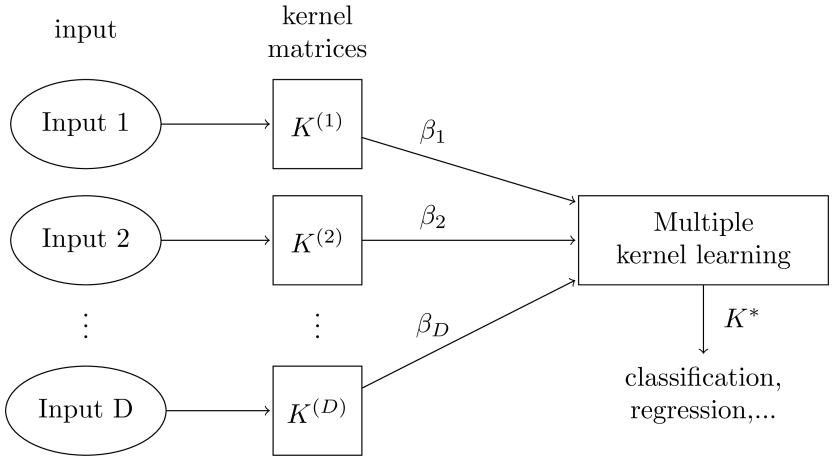

Figure 1: General multiple kernel learning pipeline.

2 Multiple kernel learning and logistic regression

In this section we state a binary classification problem by means of a convex combination of multiple kernels. The binary case is considered for simplicity, however the multiclass framework can be treated analogously. Instead of considering a single feature kernel matrix K given by a pre-defined kernel function $k(\cdot,\cdot)$ we propose to resume MKL approaches (see Figure 1). Given a set of training data $\{(x_i, y_i)\}_{i=1}^N$ with $N \in \mathbb{N}$, x_i represents a P-dimensional input vector and $y_i \in \{0, 1\}$ its label.

Let $D \in \mathbb{N}$ be the number of different sources: we define a set of kernel matrices $\mathcal{K} := \{K^{(1)}, K^{(2)}, \ldots, K^{(D)}\}$, of sizes $N \times N$, each defined by $K^{(d)} := [k^{(d)}(x_i, x_j)]_{i,j=1}^N$, where $k^{(d)}(\cdot,\cdot)$ is a positive semi-definite kernel function. Note that $K^{(1)}, \ldots, K^{(D)}$ could either arise from different kernel functions (e.g Gaussian, RBF or polynomial kernels) or the same functions with different parameters.
In order to define a linear combination of kernels still being a kernel, let us define a new matrix $K \in \mathbb{R}^{N\times N}$

$$K := \sum_{d=1}^{D} \beta_d K^{(d)} \tag{1}$$

where $\beta_1, \ldots, \beta_D$ are assumed to be non-negative weights. Although in the literature $\beta_1, \ldots, \beta_D$ are usually assumed to take values in the $D-1$ simplex, this property is not needed in order for K to be a kernel and this is why we remove that constraint in a first time. That case, however, will be treated in the long version of this paper.

We define then $K_i^{(d)}$ to be the i-th row of the matrix $K^{(d)}$. The i-th observation y_i is then described by row vectors $K_i^{(d)}$ for $d \in \{1, \ldots, D\}$. We see y_i as the outcome of a Bernoulli random variable Y_i, whose probability of success is denoted by p_i. The N random variables

$Y_1, ..., Y_N$ are assumed to be independent (not identically distributed) and

$$\log\left(\frac{p_i}{1-p_i}\right) = \left[\sum_{d=1}^{D} \beta_d K_i^{(d)}\right]\alpha, \tag{2}$$

where α is an unknown vector parameter in $\mathbb{R}^N$. We can rewrite the right-hand side of eq. (2) as

$$\left[\sum_{d=1}^{D} \beta_d K_i^{(d)}\right]\alpha = \sum_{n=1}^{N}\sum_{d=1}^{D} \beta_d K_{in}^{(d)} \alpha_n = \beta^T L_i \alpha,$$

where $\beta := [\beta_1, \ldots, \beta_D]^T \in \mathbb{R}^D$ and $L_i := [K_i^{(1)}, \ldots, K_i^{(D)}]^T \in \mathbb{R}^{D\times N}$ (the d-th row of L_i is $K_i^{(d)}$).

The likelihood of the observed data is

$$p(y|\mathcal{K}, \alpha, \beta) = \prod_{i=1}^{N} p_i^{y_i}(1-p_i)^{1-y_i}$$

and log-likelihood

$$\log p(y|\mathcal{K}, \alpha, \beta) = \sum_{i=1}^{N}\left[y_i \log\left(\frac{p_i}{1-p_i}\right) + \log(1-p_i)\right] = \sum_{i=1}^{N}\left[y_i \beta^T L_i \alpha - \log\left(1 + e^{\beta^T L_i \alpha}\right)\right]$$

where we used

$$p_i = \frac{e^{\beta^T L_i \alpha}}{1 + e^{\beta^T L_i \alpha}}.$$

In order to reduce the risk of over fitting and possibly achieve a better generalization, it is common to consider an $l2$ penalised log-likelihood for the logistic regression model (Cessie and Houwelingen, 1992). One major advantage of such choice is that the log-likelihood remains differentiable and concave and this is why we do adopt that choice too. Now, knowing that, under our assumptions for a given β, K in eq. (1) is a positive semi-definite kernel matrix, the above log-likelihood is a kernelized one and via the Representer Theorem (Schölkopf et al., 2001) the $l2$ penalty in the feature space translates into $\alpha^T K \alpha$. Then, the penalised log-likelihood is

$$l_R(\alpha, \beta) = \log p(y|\mathcal{K}, \alpha, \beta) - \lambda \alpha^T K \alpha = \sum_{i=1}^{N}\left[y_i \beta^T L_i \alpha - \log\left(1 + e^{\beta^T L_i \alpha}\right)\right] - \lambda \alpha^T K \alpha. \tag{3}$$

Optimizing the above quantity jointly in α and β is not trivial and no close formula exists, not even when one fixes α (respectively β) and optimizes with respect to β (α) although in that case our problem reduces to a standard (kernelized) logistic regression. Instead to directly attack the above problem by numerical maximization, we introduce an additional assumption whose benefits will be clear in a while.

Let us assume that $\beta_1, \ldots, \beta_D \overset{\text{i.i.d.}}{\sim} \text{Exp}(\delta)$ with δ a fixed positive scalar:

$$p(\beta|\delta) = \prod_{d=1}^{D} p(\beta_d|\delta) = \delta^D e^{-\delta \sum_{d=1}^{D} \beta_d},$$

thus

$$\log p(\beta|\delta) = D \log \delta - \delta \sum_{d=1}^{D} \beta_d = D \log \delta - \delta ||\beta||_1 \tag{4}$$

where the last equality comes from the fact that each β_d is non-negative. Thus the log-likelihood of the complete data (y, β) reads

$$\log p(y, \beta|\mathcal{K}, \alpha, \delta) = l_R(\alpha, \beta) + \log p(\beta|\delta) = l_R(\alpha, \beta) - \delta \|\beta\|_1 + D \log \delta. \tag{5}$$

As it can be seen a Lasso penalty naturally arises on β, making this objective function very interesting to optimize since it allow to select the relevant kernels via shrinking. A similar optimization problem was attacked from a pure optimization perspective in (Rakotomamonjy et al., 2008) for SVMs and (He et al., 2021) for logistic regression. However, the interest of the proposed probabilistic formulation is that it opens the doors to stochastic variational inference which allows us to optimize a lower bound of the the log-likelihood of the observed data

$$\log p(y|\mathcal{K}, \alpha, \delta) = \log \int_\beta p(y, \beta|\mathcal{K}, \alpha, \delta) d\beta$$

in an incredibly simple way (next section). A final remark before moving to the inference of the model parameters. Similarly to what we did for β, a prior distribution could be attached to α in order to avoid any estimation and adopt a fully Bayesian perspective, similar to what is done in (Gönen, 2012). However we prefer to maintain the hybrid approach detailed so far in order to keep the model as simple as possible.

3 Stochastic variational inference for MKL

Focusing now on the integrated log-likelihood with respect to β, we have that for any distribution $q(\cdot)$ on β, with the same support of the prior p_δ, it holds that

$$\begin{aligned}
\log p(y|\mathcal{K}, \alpha) &= \log \int_\beta p(y, \beta|\mathcal{K}, \alpha) d\beta \\
&= \log \int_\beta \frac{p(y, \beta|\mathcal{K}, \alpha)}{q(\beta)} q(\beta) d\beta \\
&= \log \mathbb{E}_q \left[\frac{p(y, \beta|\mathcal{K}, \alpha)}{q(\beta)} \right] \geq \mathbb{E}_q \left[\log \frac{p(y, \beta|\mathcal{K}, \alpha)}{q(\beta)} \right],
\end{aligned} \tag{6}$$

by Jensen inequality. It is also well known that the above inequality turns into an equality when $q(\beta) = p(\beta|y, K, \alpha)$ which is however not analytically tractable here. Thus, we call the lower-bound of eq. (6) $\mathcal{L}(q, \alpha)$ and we set the *variational* (approximate) posterior distribution such as

$$q(\beta) = \prod_{d=1}^{D} q(\beta_d)$$

with $q(\beta_d) := q(\beta_d|\lambda_d)$ assumed to follow an Exponential distribution of parameter $\lambda_d > 0$[1].

By denoting $\lambda := (\lambda_1, \ldots, \lambda_D)$, the lower-bound can be developed as

$$\begin{aligned}
\mathcal{L}(q, \alpha) &= \mathbb{E}_q\big[\log p(y, \beta|\mathcal{K}, \alpha)\big] - \mathbb{E}_q\big[\log q(\beta|\lambda)\big] = \\
&= \mathbb{E}_q\big[l_R(\alpha, \beta)\big] + \mathbb{E}_q\big[\log p(\beta|\delta)\big] - \mathbb{E}_q\big[\log q(\beta|\lambda)\big] = \\
&= \mathbb{E}_q\big[l_R(\alpha, \beta)\big] + \mathbb{E}_q\left[\log \frac{p(\beta|\delta)}{q(\beta|\lambda)}\right] = \\
&= \mathbb{E}_q\big[l_R(\alpha, \beta)\big] - \sum_{d=1}^{D}\left(\frac{\delta}{\lambda_d} + \log \lambda_d\right) + D(\log \delta + 1)
\end{aligned} \tag{7}$$

where the last two terms come from the explicit calculation of the negative Kullback-Leibler divergence between the approximate posterior $q(\cdot|\lambda)$ and the prior distribution $p(\cdot|\delta)$. Since for the Exponential distribution we can adopt the reparametrization trick, i.e.

$$\beta_d = \frac{1}{\lambda_d}\eta,$$

with $\eta \sim$ Exponential(1), following (Kingma and Welling, 2014) (in the context of variational auto-encoders) we can reparametrize the above lower bound in such a way that it is differentiable with respect to the model parameters α and the variational parameters λ and obtain an unbiased estimate of its gradient through sampling from independent Exponential distritubions of parameter 1. The model parameters are then optimized by stochastic gradient ascent. More details are in Pseudo-code 1.

Algorithm 1 SVIMKL

Require: $\alpha_0 \in \mathbb{R}^N$, $\ell_0 = [\log D, \ldots, \log D] \in \mathbb{R}^D$
Ensure: solution: (α^*, λ^*)
  $\alpha_c \leftarrow \alpha_0$
  $\ell_c \leftarrow \ell_0$
  while $\mathcal{L}(q, \alpha)$ not converged **do**
    $\lambda_c \leftarrow \exp(\ell_c)$
    $\beta_c \leftarrow \frac{1}{\lambda_c}\text{Exp}(1)$
    evaluate $\tilde{\mathcal{L}}(q, \alpha)$ in β_c ▷ $\tilde{\mathcal{L}}$: no more $\mathbb{E}_q$ in Eq. (7) and β_c in place of β
    $\alpha_c \leftarrow \alpha_{c+1} = \alpha_c + \nabla_\alpha \tilde{\mathcal{L}}(q, \alpha_c)$
    $\ell_c \leftarrow \ell_{c+1} = \ell_c + \nabla_\ell \tilde{\mathcal{L}}(q, \alpha_c)$
  end while
  $(\alpha^*, \lambda^*) \leftarrow (\alpha_c, \lambda_c)$

[1]Variational distributions other than the Exponential are also considered and these choices will discussed at the oral.

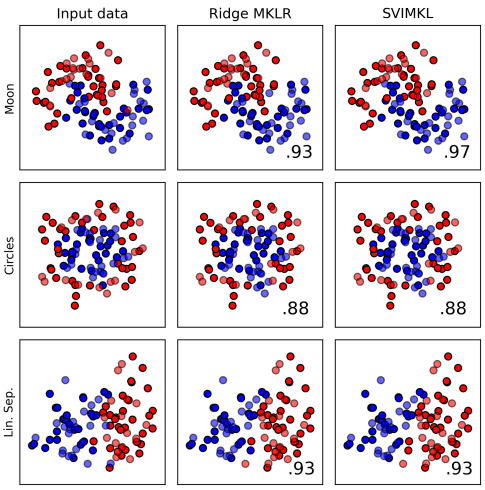

Figure 2: Binary classification results. The bottom right shows the accuracy on the test set.

4 Numerical examples

In this section, we present a toy example of binary classification. We consider three 2 dimensional datasets taken from the `sklearn` library: moon, circles and linearly separable data. For each dataset we consider 100 number of points.
In order to perform kernel(s) selection, we consider four ($D = 4$) different kernels. A linear kernel, a sigmoid kernel, an RBF kernel with $\sigma = 0.1$ and a Laplacian kernel with $\sigma = 0.1$.
We also compare our method with a multiple kernel logistic regression with a pure optimization approach (MKLR) and an $l2$ penalty on kernel weights.

Input Data	Ridge MKLR	SVIMKL
Moon	1.0244/0.2397/1.1187/3.7014	0.0338/0.0899/0.2191/0.6546
Circles	0.0277/0.0605/1.4507/5.7608	0.0964/0.0272/0.1329/0.8220
Lin. sep. data	1.1034/0.1429/0.8841/2.0121	0.2948/0.0853/0.2520/0.1871

Table 1: Table of optimal kernel weights ($\beta_1^*,\ \beta_2^*,\ \beta_3^*,\ \beta_4^*$) associated to the linear, sigmoid, RBF and Laplacian kernel respectively.

Table 1 reports the optimal weights associated to the four kernels. In general, our method tends to perform more kernel selection than Ridge MKLR. For example, in the case of the Moon dataset, it can be seen that the weights associated to linear and sigmoid kernels are closer to zero with SVIMKL than those computed with Ridge MKLR. Figure 2 illustrates the accuracy results on the test set.

(More examples will be presented at the oral)

References

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877.

Cessie, S. l. and Houwelingen, J. V. (1992). Ridge estimators in logistic regression. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 41(1):191–201.

Damoulas, T. and Girolami, M. A. (2008). Probabilistic multi-class multi-kernel learning: on protein fold recognition and remote homology detection. *Bioinformatics*, 24(10):1264–1270.

Girolami, M. and Rogers, S. (2005). Hierarchicsbayesian models for kernel learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 241–248.

Gönen, M. (2012). Bayesian efficient multiple kernel learning. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 91–98.

Gönen, M. and Alpaydın, E. (2011). Multiple kernel learning algorithms. *The Journal of Machine Learning Research*, 12:2211–2268.

He, X., Huang, J., and Zeng, Z. (2021). Logistic regression based multi-task, multi-kernel learning for emotion recognition. In *2021 6th IEEE International Conference on Advanced Robotics and Mechatronics (ICARM)*, pages 572–577. IEEE.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational bayes. *stat*, 1050:10.

Naesseth, C., Ruiz, F., Linderman, S., and Blei, D. (2017). Reparameterization gradients through acceptance-rejection sampling algorithms. In *Artificial Intelligence and Statistics*, pages 489–498. PMLR.

Rakotomamonjy, A., Bach, F., Canu, S., and Grandvalet, Y. (2008). Simplemkl. *Journal of Machine Learning Research*, 9:2491–2521.

Schölkopf, B., Herbrich, R., and Smola, A. J. (2001). A generalized representer theorem. In *International conference on computational learning theory*, pages 416–426. Springer.

Sonnenburg, S., Rätsch, G., Schäfer, C., and Schölkopf, B. (2006). Large scale multiple kernel learning. *The Journal of Machine Learning Research*, 7:1531–1565.

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999.

Wang, T., Zhang, L., and Hu, W. (2021). Bridging deep and multiple kernel learning: A review. *Information Fusion*, 67:3–13.