Learning the kernel matrix by predictive low-rank approximations

Martin Stražar and Tomaž Curk

Bioinformatics Laboratory, University of Ljubljana, Faculty of Computer and Information Science, Večna pot 113, 1000 Ljubljana

Motivations

Many representaions for the same objects are often available: vectors, strings, graphs, time series, etc.

Kernel methods enable learning independently of representation. Contemporary multiple kernel learning (MKL) algorithms are stated as optimization problems and require full kernel matrices.

Low-rank approximations are essential for efficient large scale kernel learning, but are rarely learned simultaneously with the combined kernel matrix.

Highlights

The algorithm `mklaren` (Multiple kernel learning with least-angle regression) learns low-rank approximations to kernels **simultaneously** including the information on targets.

Relevant kernels are selected using a heuristic and approximated using a numerical algorithm in $O(K^3 + Kpn\delta^2)$.

L2-regularized regression (ridge) in the combined feature space.

Prediction in a transductive and/or inductive setting.

article

github.com/mstrazar/mklaren

Inputs

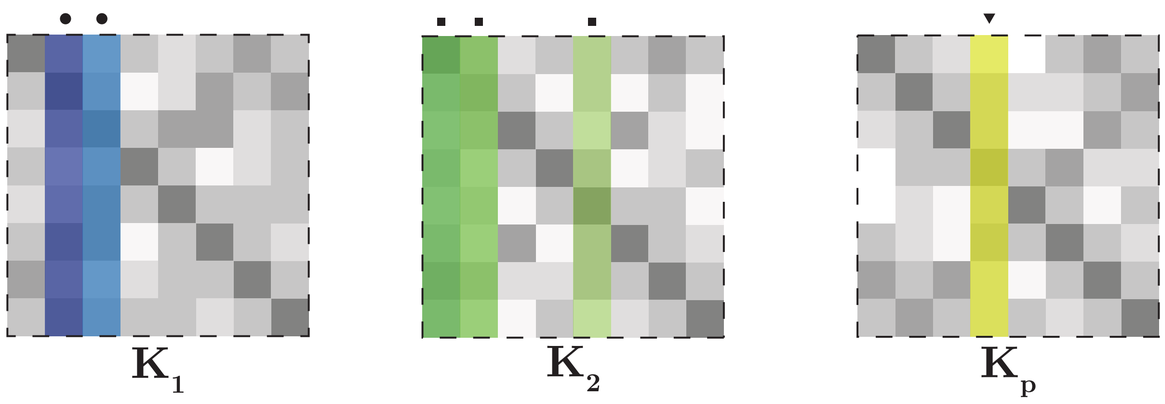
$\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ objects
$k_1, k_2, \ldots, k_p$ kernels
$\mathbf{y}$ regression targets
K maximum rank
δ no. look-ahead columns
λ regularization parameter

Results

$\mathbf{G}_1, \mathbf{G}_2, \ldots, \mathbf{G}_p$ approximations
$\mathbf{H}$ combined feature matrix
$\boldsymbol{\mu}$ regression line
$\boldsymbol{\beta}$ regression coefficients

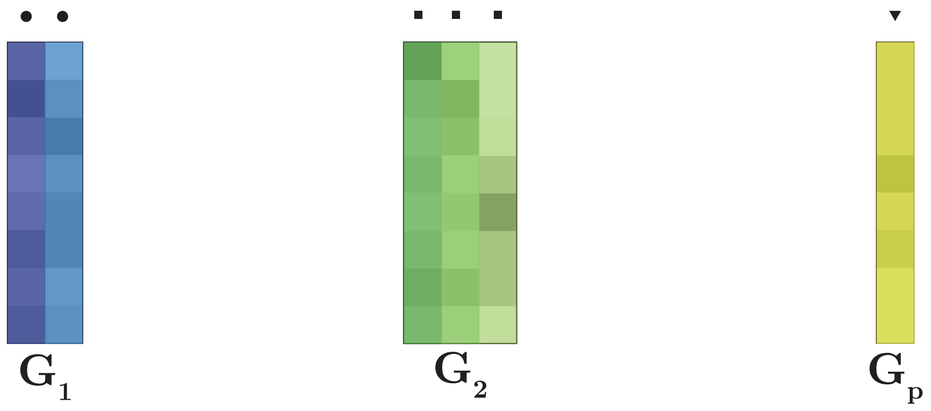
kernel functions model different input representations (vectors, strings, structures, ...)

kernel matrices approximated rather than computed explicitly

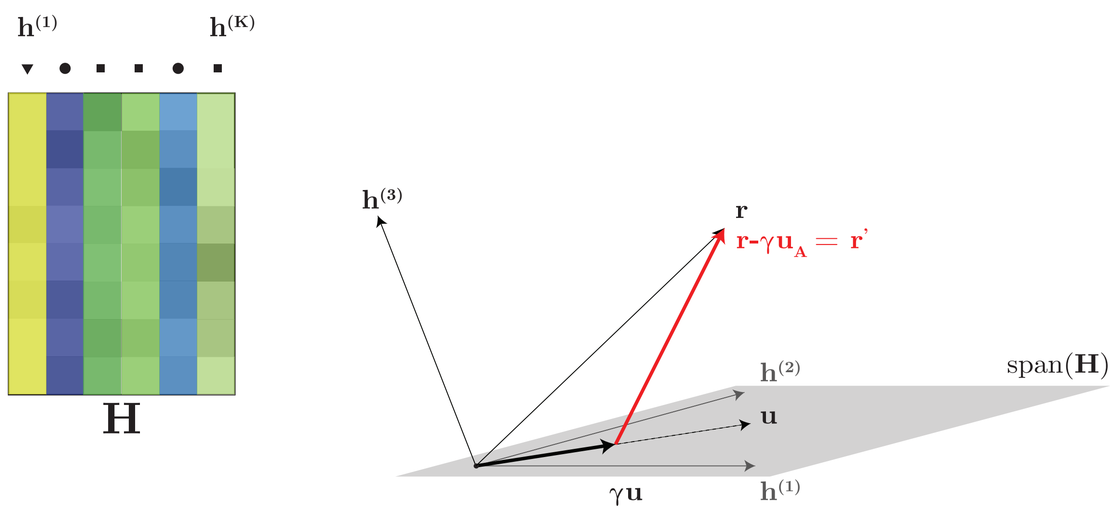


Cholesky factors represent implicit features

more relevant kernels wrt. **targets** get more columns



column selection via least angle regression in a combined feature space



Combined kernel $\mathbf{HH}^T$

mklaren pseudocode

```
Compute standard ICD for each G_q for δ
lookahead columns

while dim(H) < K:
   Select k_q and pivot i using LAR
   Compute column with Cholesky step g_qi
   G_q ← [G_q g_qi]
   h_j ← standardize(g_qi)
   H ← [H h_j]

   Compute bisector u
      ∡(h_1, u) = ∡(h_2, u) = ∡(h_j-1, u)
   Compute γ s.t.
      r' = r - γu
      ∡(h_1, r') = ∡(h_2, r') = ∡(h_j, r')
   Update μ and r = r'
      μ = μ + γu

Solve Hβ=μ  for regression coefficients β
```

Incomplete Cholesky Decomposition

A greedy approach to column sampling of $\mathbf{K}_q$
No explicit evaluation of full $\mathbf{K}_q$ required
Novel approach to pivot selection using LAR

... with look-ahead columns [1]
Evaluate gain with respect to $\boldsymbol{\mu}$, $\mathbf{r}$

Use δ look-ahead columns

$$\mathbf{K}_q \approx \mathbf{L}_q = \mathbf{G}_q^{\delta}\mathbf{G}_q^{\delta T}$$

Pivot selection in $O(pn\delta^2)$

Least-angle regression

Alternative to step-, stage- wise feature selection in combined feature space spanned by $\mathbf{H}$.

Select a column and update along the bisector $\mathbf{u}$ such that correlations (angles) with residual $\mathbf{r}$ are equal for all active columns.

Step size γ is determined such that a new pivot column is added to a $\mathbf{G}_q$ and $\mathbf{H}$.

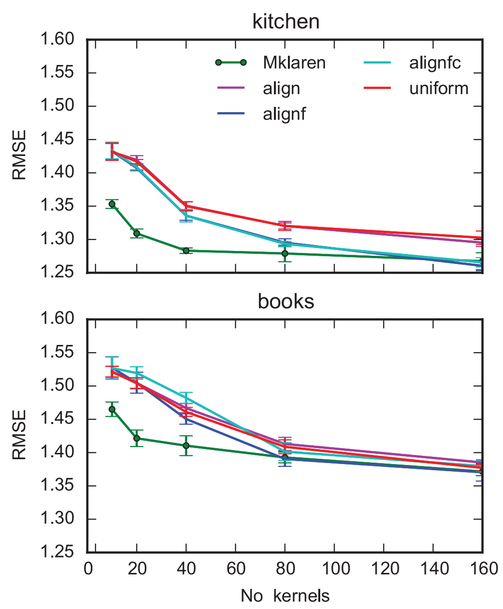
Improved low-rank approximations

Dataset	n	mklaren	csi	icd	Nyström
boston	506	**42**	63	> 140	119
kin	1000	**63**	> 140	> 140	> 140
pumadyn	1000	**49**	> 140	56	98
abalone	1000	**21**	28	35	49
comp	1000	**49**	63	> 140	> 140
ionosphere	351	**14**	**14**	42	35
bank	1000	**21**	42	42	112
diabetes	442	**14**	**14**	**14**	21

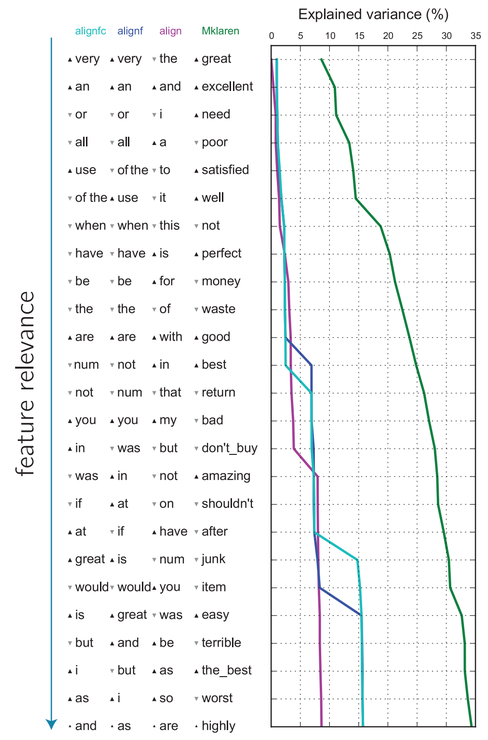
Comparison of minimal rank for which the RMSE differs by at most one standard deviation to RMSE obtained with the full kernel matrices using uniform kernel combination.

Exploiting correlations between kernels induces feature spaces with significantly lower rank.

Predictive performance & model interpretation



MKL with 4000 kernels on the Blitzer product review dataset [2]. RMSE on the test set for MKL methods, with rank equal to the number of kernels. Comparison with MKL based on centered alignment [3].



Increase in explained variance upon incrementally including features to the model with respect to order of selected kernels (features) by each method.

References

[1] F. R. Bach and M. I. Jordan, "Predictive low-rank decomposition for kernel methods," Proceedings of the 22nd international conference on Machine learning - ICML '05. ACM Press, New York, New York, USA, pp. 33–40, 2005.
[2] J. Blitzer, M. Dredze, F. Pereira, "Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification," in ACL, 2007, vol. 7, pp. 440–447.
[3] C. Cortes, M. Mohri, and A. Rostamizadeh, "Algorithms for Learning Kernels Based on Centered Alignment," J. Mach. Learn. Res., vol. 13, pp. 795–828, Mar. 2012.