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 representations 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.





Inputs

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

Results

 $\mathbf{G}_{1}, \mathbf{G}_{2}, \dots, \mathbf{G}_{p}$ approximations **H** combined feature matrix **µ** regression line **β** regression coefficients

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





mklaren pseudocode

Compute standard ICD for each $\boldsymbol{G}_{\!\!\boldsymbol{\alpha}}$ for δ lookahead columns

kernel matrices approximated rather than computed explicitly

Cholesky factors represent implicit features

more relevant kernels wrt. targets get more columns





while dim(**H**) < K: Select k_{a} and pivot *i* using LAR Compute column with Cholesky step **g**_{ai} $\mathbf{G}_{\mathbf{q}} \leftarrow [\mathbf{G}_{\mathbf{q}} \ \mathbf{g}_{\mathrm{qi}}]$ $\mathbf{h}_{i} \leftarrow \text{standardize}(\mathbf{g}_{di})$ $\mathbf{H} \leftarrow [\mathbf{H} \mathbf{h}]$ Compute bisector ${\boldsymbol{u}}$ $\mathbf{A}(\mathbf{h}_1, \mathbf{u}) = \mathbf{A}(\mathbf{h}_2, \mathbf{u}) = \mathbf{A}(\mathbf{h}_{1-1}, \mathbf{u})$ Compute γ s.t.

$$\mathbf{r'} = \mathbf{r} - \gamma \mathbf{u}$$

$$\mathbf{\Delta}(\mathbf{h}_1, \mathbf{r'}) = \mathbf{\Delta}(\mathbf{h}_2, \mathbf{r'}) = \mathbf{\Delta}(\mathbf{h}_j, \mathbf{r'})$$
pdate $\mathbf{\mu}$ and $\mathbf{r} = \mathbf{r'}$

$$\mathbf{\mu} = \mathbf{\mu} + \gamma \mathbf{u}$$

Solve $H\beta = \mu$ for regression coefficients β

Incomplete Cholesky Decomposition

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

... with look-ahead columns ^[1] Evaluate gain with respect to μ , r

Use δ look-ahead columns

 $\mathbf{K}_{\mathbf{q}} \approx \mathbf{L}_{\mathbf{q}} = \mathbf{G}_{\mathbf{q}}^{\delta} \mathbf{G}_{\mathbf{q}}^{\delta \top}$

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

 $\operatorname{span}(\mathbf{H})$



 \mathbf{G}^{δ}

Combined kernel **HH**^T



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.

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.



No kernels

 $\mathbf{h}^{(2)}$

 $\mathbf{h}^{(1)}$

u

 $\gamma \mathbf{u}$

Least-angle regression

Alternative to step-, stage- wise feature selection in combined feature space spanned by **H**.

Select a column and update along the bisector **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 **G**_a and **H**.

		Explained variance (%)	
alignfc alignf align	Mklaren	0 5 10 15 20 25 30 35	Increase in
▲ very ▲ very ▼ the ▲	great		avalained variance
∡an ∡an ∡and ∡	excellent		
▼ Or ▼ Or ▼ İ ▲	need		upon incrementally
∗all ∗all ⊾a •	poor		including features
▲use ▼ofthe▼to ▲	satisfied		to the model with
▼ of the ▲ use ▼ it ▲	well		
• when • when • this	not	-	
have ∗ have ▲ is ▲	perfect	-	Selected
v be v be ▲ for v	money	-	kernels (features)
• the • the • of •	waste		by each method.
▲are ▲are ▲with ▲	sood		
≖num ≖not ⊾in ⊾	best		
• not • num • that •	return		
▲ you ▲ you ▲ my ▼	bad		
. in vas vbut v	don't_buy		
vwas ⊾in vnot ⊾	amazing		
vif ₄at von v	shouldn't		
▲at vif ▲have v	after		
▲ great ▲ is ▼ num ▼	junk		
v wouldv would⊾ you v item			
▲is ▲great was ▲	easy		
v but ▲and ▲be v	terrible		
⊾i ⊽but ⊾as ⊾	the_best		
▲as ⊾i ▲so ▼	worst		
	alignfc align align very very the an an an and or or or vi all all a use of the to of the use vit when when this have vhave is vhe vhe vher this vhe vhe vhe vit vhe vhe vhe vhe vit vhe vhe vhe vit vhe vhe vhe vhe vhe vhe vit vhe vhe vhe vhe vhe vhe vit vhe vhe vhe vhe vhe vhe vhe vit vhe vhe vhe vhe vhe vhe vhe vhe vhe vhe	alignf:alignMklaren• very• very• the• great• an• and• and• excellent• or• or• i• need• all• all• a• poor• use• of the• to• satisfied• of the• use• it• well• of the• use• it• perfect• be• have• is• perfect• be• be• for• money• the• of• waste• are• with• good• num• that• return• you• num• that• not• in• best• not• in• best• num• that• return• you• you• my• kas• in• anazing• in• not• anazing• if• have• after• great• is• num• at• num• junk• is• great• num• is• great• easy• but• and• be• but• and• terrible• is• but• as• but• as• terrible• as• is• terrible• as• is• terrible• as• is• so• as• is• terrible• as• is• so• but• as• terrible• as• is• so• but<	Explained variance (%) alignic align align Mklaren very very the very the verset an An And excellent or vor vi Aneed vall vall Aa vpoor vuse vofthe to Asatisfied vof the vuse vit Awell when vhen vthis vnot have vhave Ais Aperfect vbe vbe Afor vmoney vthe vthe vof vwaste Aare Aare Awith Agood rnum vnot Ain Abest vnot vnum vthat vreturn Ayou Ayou Amy vbad Ain vwas vbut vdon't_buy vwas Ain vnot Aamazing vif Aat von vshouldn't Aat vif Ahave vafter Agreat Ais vnum vjunk vwould vould you vittem Ais Agreat vwas Aeasy vbut Aand Abe vterrible Ai vbut Aas Athe_best Aas Ai Aso vworst



