

A two-step method for large output

Michiel Stock twitter: @michielstock

Motivation Introductor example Relational learning Other

Pairwise learning

Kronecker kerneridge regression
Two-step kerne

Computationa aspects

Cross-validation

Take home messages

# A two-step method to incorporate task features for large output spaces

Michiel Stock<sup>1</sup>, Tapio Pahikkala<sup>2</sup>, Antti Airola<sup>2</sup>, Bernard De Baets<sup>1</sup> & Willem Waegeman<sup>1</sup>

<sup>1</sup>KERMIT

Department of Mathematical Modelling, Statistics and Bioinformatics  ${\sf Ghent\ University}$ 

<sup>2</sup>Department of Computer Science University of Turku

NIPS: extreme classification workshop
December 12, 2015



A two-step method for large output spaces

Michiel Stock twitter: @michielstock

#### Motivation

#### Introductory

#### example

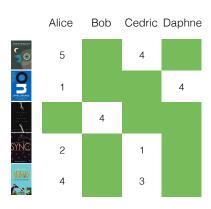
learning Other

Pairwise learning

Kronecker kerne ridge regression Two-step kerne

Computationa aspects

Cross-validation Exact online learning





A two-step method for large output spaces

Michiel Stock twitter: @michielstock

Matirotion

Introductory example

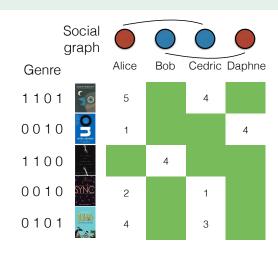
Relational learning Other

Pairwise learning

Kronecker kerne ridge regression Two-step kernel ridge regression

Computational aspects

Cross-validation Exact online learning





A two-step method for large output spaces

Michiel Stock twitter: @michielstock

@michielsto

Introductory example

Relational learning

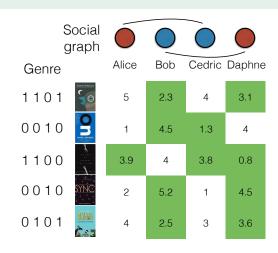
Other applications

learning methods

Kronecker kerne ridge regression Two-step kerne ridge regression

Computational aspects

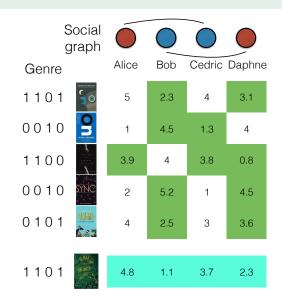
Cross-validation Exact online learning





@michielstock

#### example





A two-step method for large output spaces

Michiel Stock twitter: @michielstock

emichiesto

Introductory example

example Relational

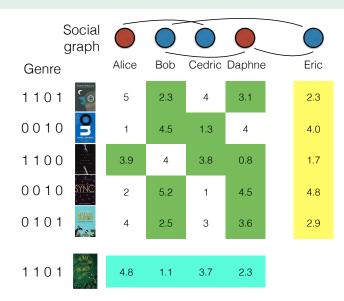
Other application

learning methods

ridge regression Two-step kernel ridge regression

Computationa aspects

Cross-validation Exact online learning





A two-step method for large output spaces

Michiel Stock twitter: @michielstock

@michielsto

Introductory

example Relational

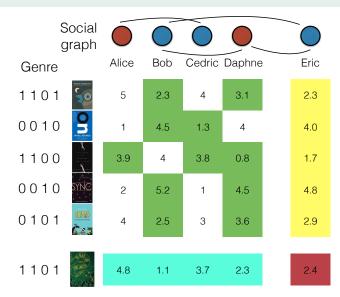
Other application:

learning methods

ridge regression Two-step kernel ridge regression

Computationa aspects

Cross-validation Exact online learning





## Learning relations

method fo

Michiel Stock twitter: @michielstock

Motivatio Introductor

example Relational

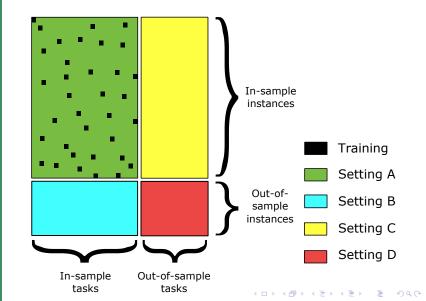
learning Other applications

Pairwise learning methods

ridge regressio

Computationa aspects

Cross-validation Exact online learning





## Other cool applications: drug design

A two-step method for large output

Michiel Stock twitter:

@michielstock

Introducto example Relational

Other applications

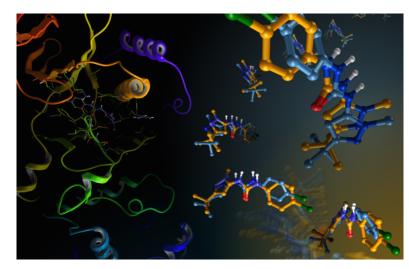
Pairwise learning methods

ridge regression
Two-step kerne

Computationa aspects

Cross-validation Exact online learning

Take home messages



Predicting interaction between proteins and small compounds



## Other cool applications: social network analysis

A two-step method for large output spaces

Michiel Stock twitter: @michielstock

Motivation Introductory example Relational learning

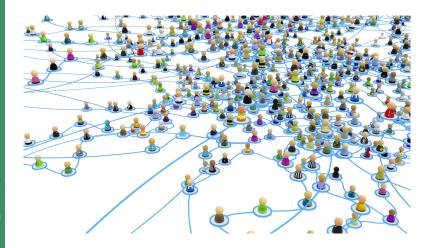
#### Other applicatio

Pairwise learning methods

Kronecker kern ridge regressior Two-step kerne ridge regressior

aspects
Cross-validation
Exact online

Take home



#### Predicting links between people



## Other cool applications: food pairing

A two-step method for large output

Michiel Stock twitter: @michielstock

Motivatio Introductor example Relational

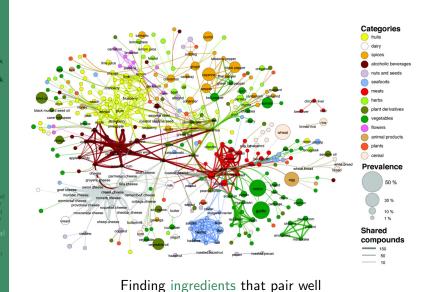
Other applicatio

Pairwise learning method

ridge regressior
Two-step kerne
ridge regressior

Computation: aspects

Cross-validatio Exact online learning





#### Learning with pairwise feature representations

A two-step method for large output spaces

Michiel Stock twitter: @michielstock

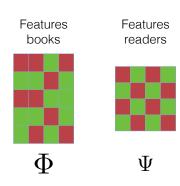
Motivation Introductor example Relational learning Other applications

Pairwise learning methods

Kronecker kernel ridge regression
Two-step kernel ridge regression

Computationa aspects

Cross-validation Exact online learning



- d: instance (e.g. book)
- $\phi(d)$ : instance features (e.g. genre)

- t : task (e.g. reader)
- $\psi(t)$ : task features (e.g. social network)



#### Learning with pairwise feature representations

A two-step method for large output spaces

Michiel Stock twitter: @michielstock

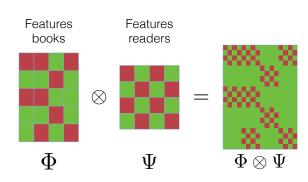
Introductory example Relational learning Other applications

Pairwise learning methods

Kronecker kernel ridge regression
Two-step kernel ridge regression

Computationa aspects

Cross-validation Exact online learning



- d: instance (e.g. book)
- $\phi(d)$ : instance features (e.g. genre)

- t : task (e.g. reader)
- $\psi(t)$ : task features (e.g. social network)



## Learning with pairwise feature representations

A two-step method for large output spaces

Michiel Stoc twitter: @michielstoc

Introductory example Relational learning Other applications

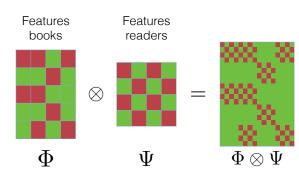
Pairwise learning methods

Kronecker kernel ridge regression Two-step kernel ridge regression

Computationa aspects

Cross-validation Exact online learning

Take home nessages



- d: instance (e.g. book)
- $\phi(d)$ : instance features (e.g. genre)

- t : task (e.g. reader)
- $\psi(t)$ : task features (e.g. social network)

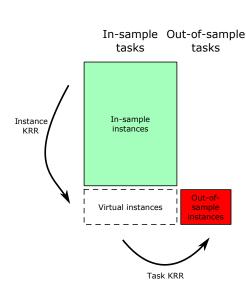
Pairwise prediction function:  $f(d,t) = \mathbf{w}^{\mathsf{T}}(\phi(d) \otimes \psi(t))$ 



## Learning relations in two steps

@michielstock

Two-step kernel ridge regression



- Build a ridge regression model to generalize to new instances
- Build a ridge regression model to generalize to new tasks





#### The two-step ridge regression

Prediction function:

$$f(d,t) = \phi(d)^{\mathsf{T}} \mathbf{W} \psi(t)$$

Parameters can be found by solving:

$$\mathbf{\Phi}^{\mathsf{T}}\mathbf{Y}\mathbf{\Psi} = (\mathbf{\Phi}^{\mathsf{T}}\mathbf{\Phi} + \lambda_{d}\mathbf{I})\mathbf{W}(\mathbf{\Psi}^{\mathsf{T}}\mathbf{\Psi} + \lambda_{t}\mathbf{I})$$

@michielstock

Two-step kernel ridge regression



## The two-step ridge regression

method for

Michiel Stock twitter: @michielstock

Prediction function:

$$f(d,t) = \phi(d)^{\mathsf{T}} \mathbf{W} \psi(t)$$

Parameters can be found by solving:

$$\mathbf{\Phi}^{\mathsf{T}}\mathbf{Y}\mathbf{\Psi} = (\mathbf{\Phi}^{\mathsf{T}}\mathbf{\Phi} + \lambda_{d}\mathbf{I})\mathbf{W}(\mathbf{\Psi}^{\mathsf{T}}\mathbf{\Psi} + \lambda_{t}\mathbf{I})$$

Two hyperparameters:  $\lambda_d$  and  $\lambda_t$ !

Motivation Introductory example

learning
Other
applications

Pairwise learning methods

ridge regression Two-step kernel ridge regression

Computationa aspects

Cross-validatio Exact online learning



## Four ways of cross validation

method for large outpu

Michiel Stock twitter: @michielstock

Introductor example Relational learning Other

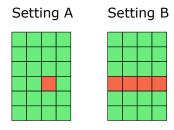
Pairwise learning methods

Kronecker kerne ridge regression Two-step kerne ridge regression

Computational aspects

Cross-validation Exact online learning

Take home messages





Train



#### Four ways of cross validation

method for large output

Michiel Stock twitter: @michielstock

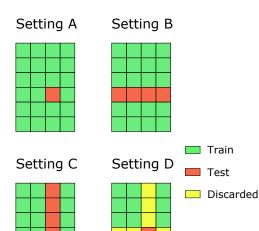
Introductory example Relational learning Other

Pairwise learning methods

ridge regression
Two-step kerne
ridge regression

Computational aspects

Cross-validation Exact online learning



- Analytic shortcuts can be derived to perform LOOCV for each setting!
- Tuning  $\lambda_d$  and  $\lambda_t$  essentially free!



## Effect of regularization for the four settings

A two-step method for large output

Michiel Stoc twitter: @michielstoc

Introductory example Relational learning Other

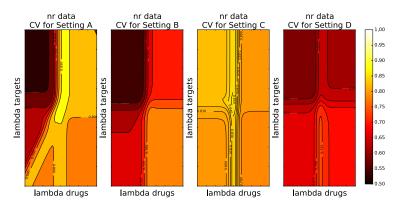
Pairwise learning methods

ridge regression
Two-step kerne
ridge regression

Computationa aspects

Cross-validation Exact online learning

Take home messages Data: protein-ligand interactions. Evaluation by AUC (lighter = better performance)



Clear difference between four settings and  $\lambda_d$  and  $\lambda_t$ !





#### Learning with mini-batches

@michielstock

Exact online learning

Tasks New training tasks Instances Initial training data New training instances Even more training

instances



## Learning with mini-batches

method fo large outpu

Michiel Stock twitter: @michielstock

Motivation Introductor example Relational learning Other

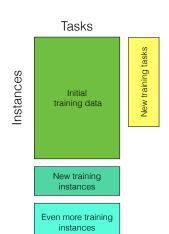
Pairwise learning methods

Kronecker kerne ridge regression Two-step kerne

Computationa aspects

Cross-validation
Exact online learning

Take home



Exact updating of the parameters when new training instances and/or taks become available

- scalable for "Big Data" applications
- updating model in dynamic environment



# Exact online learning for hierarchical text classification

A two-step method for large output spaces

Michiel Stock twitter: @michielstock

Motivation Introductory example Relational learning Other applications

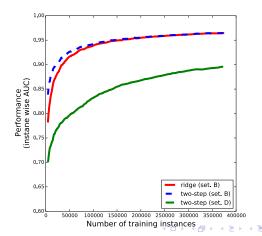
Pairwise learning methods

ridge regressior Two-step kerne ridge regressior

Computational aspects

Exact online learning

Take home messages Hierarchical text classification (> 12,000 labels): from 5,000 to 350,000 instances in steps of 1,000 instances.





## Why two-step ridge regression?

@michielstock

Take home messages

• Zero-shot learning, transfer learning, multi-task learning... in one line of code



## Why two-step ridge regression?

method for large outpu spaces

Michiel Stock twitter: @michielstock

Introductory example Relational learning Other

Pairwise learning

Kronecker kern ridge regression Two-step kerne ridge regression

Computationa aspects

Cross-validation

- Zero-shot learning, transfer learning, multi-task learning...
   in one line of code
- Theoretically well founded



## Why two-step ridge regression?

method for large output spaces

Michiel Stock twitter: @michielstock

Motivation Introductory example Relational learning Other applications

Pairwise learning methods

Kronecker kerne ridge regression Two-step kerne ridge regression

Computational aspects

Cross-validation Exact online learning

- Zero-shot learning, transfer learning, multi-task learning...
   in one line of code
- Theoretically well founded
- Allows for nifty computational tricks
  - 'free' tuning for the hyperparameters
  - 'free' LOOCV for all four settings!
  - closed-form solution for updating with mini-batches



A two-step method for large output

Michiel Stock twitter: @michielstock

Motivation Introductory example Relational learning Other

Pairwise learning

Kronecker kerne ridge regression Two-step kernel

Computationa

Cross-validation
Exact online

Take home messages

# A two-step method to incorporate task features for large output spaces

Michiel Stock<sup>1</sup>, Tapio Pahikkala<sup>2</sup>, Antti Airola<sup>2</sup>, Bernard De Baets<sup>1</sup> & Willem Waegeman<sup>1</sup>

<sup>1</sup>KERMIT

Department of Mathematical Modelling, Statistics and Bioinformatics  ${\sf Ghent\ University}$ 

<sup>2</sup>Department of Computer Science University of Turku

NIPS: extreme classification workshop

December 12, 2015