The 14th China Workshop on Machine Learning and Applications
第十四届中国机器学习及其应用研讨会

2016年11月4~6日 南京大学 南京

Multi-Task Learning: Models, Optimization and Applications

Linli Xu
University of Science and Technology of China

中国科学技术大学
University of Science and Technology of China

Outline

- Introduction to multi-task learning (MTL): problem and models
- Multi-task learning with task-feature co-clusters
- Low-rank optimization in multi-task learning
- Multi-task learning applied to trajectory regression

Multiple Tasks

Examination Scores Prediction[1] (Argyriou *et. al.*'08)

School 1 - Alverno High School

Student id	Birth year	Previous score	…	School ranking	…
72981	1985	95	…	83%	…

student-dependent: Birth year, Previous score, …; school-dependent: School ranking, …

→

Exam score
?

⋮

School 138 - Jefferson Intermediate School

Student id	Birth year	Previous score	…	School ranking	…
31256	1986	87	…	72%	…

student-dependent: Birth year, Previous score, …; school-dependent: School ranking, …

→

Exam score
?

School 139 - Rosemead High School

Student id	Birth year	Previous score	…	School ranking	…
12381	1986	83	…	77%	…

student-dependent: Birth year, Previous score, …; school-dependent: School ranking, …

→

Exam score
?

[1]The Inner London Education Authority (ILEA)

Learning Multiple Tasks

Learning each task independently

School 1 - **Alverno High School**

Student id	Birth year	Previous score	School ranking	…
72981	1985	95	83%	…

→

Exam Score
?

⇨ 1st task

Excellent

⋮

School 138 - **Jefferson Intermediate School**

Student id	Birth year	Previous score	School ranking	…
31256	1986	87	72%	…

→

Exam Score
?

⇦ 138th task

Excellent

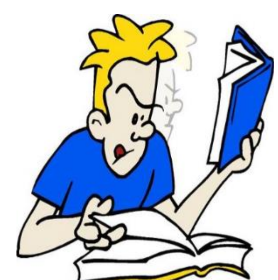
School 139 - **Rosemead High School**

Student id	Birth year	Previous score	School ranking	…
12381	1986	83	77%	…

→

Exam Score
?

⇨ 139th task



Excellent

Learning Multiple Tasks

Learning multiple tasks simultaneously

School 1 - Alverno High School

Student id	Birth year	Previous score	School ranking	…		Exam Score
72981	1985	95	83%	…	→	?

1st task

⋮

School 138 - Jefferson Intermediate School

Student id	Birth year	Previous score	School ranking	…		Exam Score
31256	1986	87	72%	…	→	?

138th task

School 139 - Rosemead High School

Student id	Birth year	Previous score	School ranking	…		Exam Score
12381	1986	83	77%	…	→	?

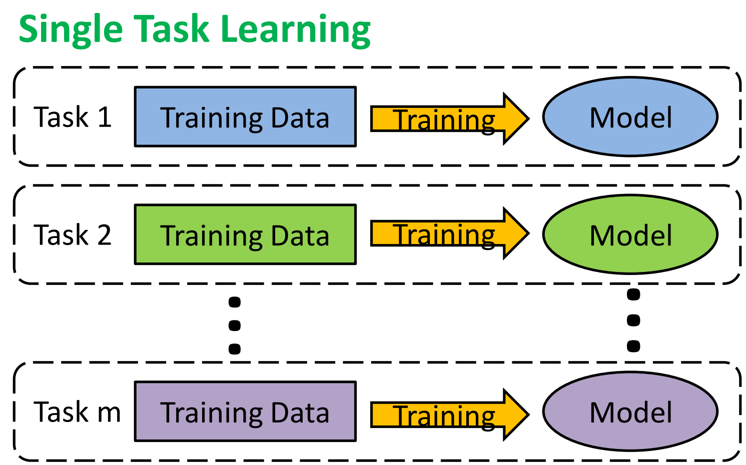
139th task

Learn tasks simultaneously
Model the task relationships

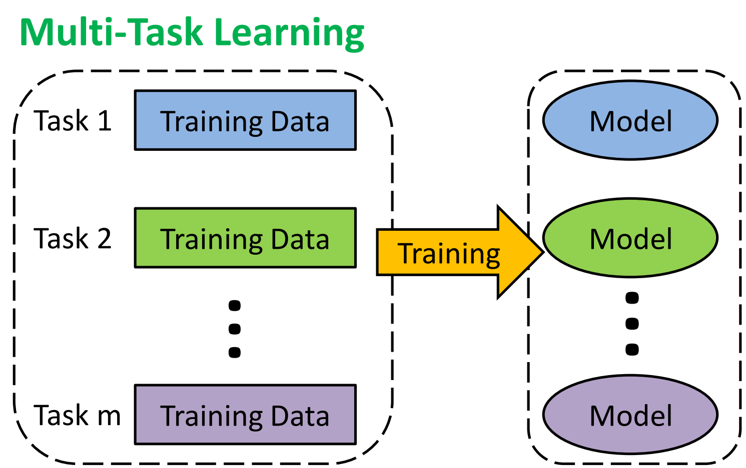
……

Multi-Task Learning

- Different from single task learning



- Training multiple tasks simultaneously to exploit task relationships



Exploiting Task Relationships

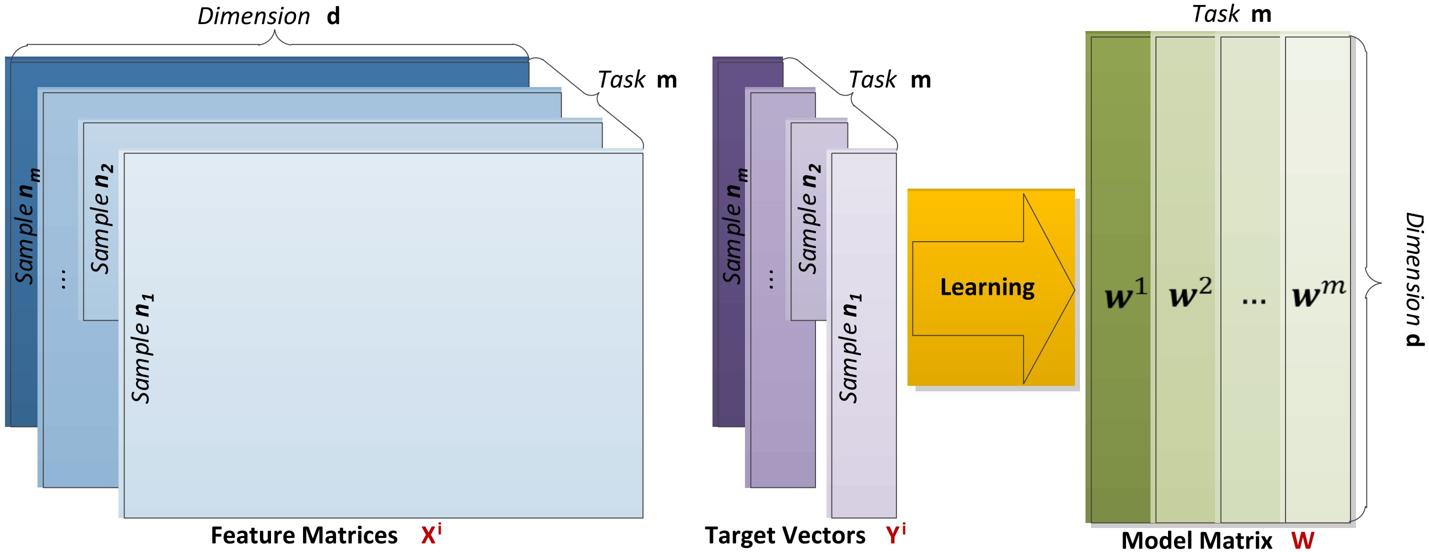
Key challenge in multi-task learning:

Exploiting (statistical) relationships between the tasks so as to improve individual and/or overall predictive accuracy (in comparison to training individual models)!

How Tasks Are Related?

- All tasks are related
  - Models of all tasks are close to each other;
  - Models of all tasks share a common set of features;
  - Models share the same low rank subspace
- Structure in tasks
  - clusters / graphs / trees
- Learning with outlier tasks

Regularization-based Multi-Task Learning



We focus on linear models: $Y^i \sim X^i \boldsymbol{w}^i$

$$X^i \in \mathbb{R}^{n_i \times d}, Y^i \in \mathbb{R}^{n_i \times 1}, W = [\boldsymbol{w}^1, \boldsymbol{w}^2, \dots, \boldsymbol{w}^m] \in \mathbb{R}^{d \times m}$$

Generic framework

$$\min_W \sum_i Loss(W, X^i, Y^i) + \lambda \boxed{Reg(W)}$$

Impose various types of relations on tasks with $Reg(W)$

How Tasks Are Related?

- All tasks are related
  - Models of all tasks are close to each other;
  - Models of all tasks share a common set of features;
  - Models share the same low rank subspace

- Structure in tasks
  - clusters / graphs / trees

- Learning with outlier tasks

MTL Methods: Mean-Regularized MTL

Evgeniou & Pontil, 2004 KDD

Assumption: model parameters of all tasks are close to each other.

- Advantage: simple, intuitive, easy to implement
- Disadvantage: too simple

Regularization

- Penalizes the deviation of each task from the mean

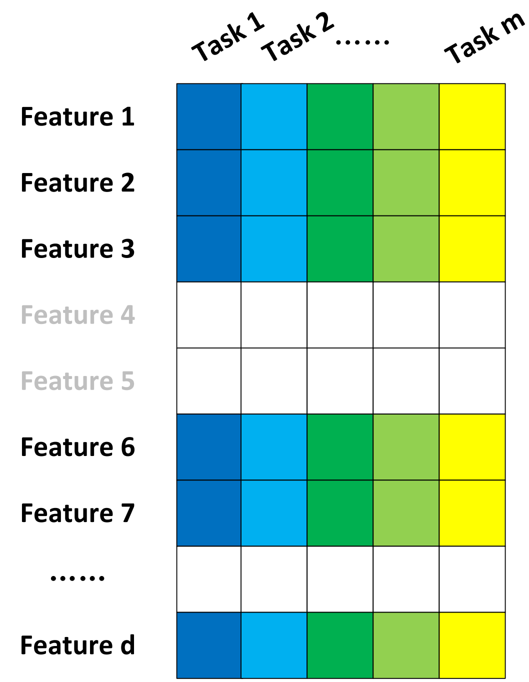
$$\min_{W} Loss(W) + \lambda \sum_{i=1}^{m} \left\| W^{i} - \frac{1}{m} \sum_{s=1}^{m} W^{s} \right\|_{2}^{2}$$

MTL Methods: Joint Feature Learning

Evgeniou *et al.* 2006 NIPS, Obozinski *et. al.* 2009 Stat Comput, Liu *et. al.* 2010 Technical Report

Assumption: models of all tasks share a common set of features

– Using group sparsity: $\ell_{1,q}$-norm regularization



Regularization

– $\|W\|_{1,q} = \sum_{i=1}^{d} \|\boldsymbol{w}_i\|_q$

– When $q > 1$ we have group sparsity

$$\min_{W} Loss(W) + \lambda \|W\|_{1,q}$$

MTL Methods: Low-Rank MTL

Ji *et. al.* 2009 ICML

<u>Assumption:</u> in high dimensional feature space, the linear models share the same low-rank subspace

Regularization - Rank minimization formulation

$$\min_{W} Loss(W) + \lambda \cdot \mathrm{rank}(W)$$

– Rank minimization is *NP-Hard* for general loss functions

- Convex relaxation: nuclear norm minimization

$$\min_{W} Loss(W) + \lambda \|W\|_*$$

($\|W\|_*$: sum of singular values of W)

How Tasks Are Related?

- All tasks are related
  - Models of all tasks are close to each other;
  - Models of all tasks share a common set of features;
  - Models share the same low rank subspace

- Structure in tasks
  - clusters / graphs / trees

- Learning with outlier tasks

MTL Methods: Clustered MTL

Zhou *et. al.* 2011 NIPS

Assumption: cluster structure in tasks - the models of tasks from the same group are closer to each other than those from a different group

Regularization - capture clustered structures

$$\min_{W,F:F^TF=I_k} Loss(\mathrm{W}) + \alpha\underbrace{[\mathrm{tr}(W^TW) - \mathrm{tr}(F^TW^TWF)]}_{\text{capture cluster structures}} + \beta\underbrace{\mathrm{tr}(W^TW)}_{\text{Improves generalization performance}}$$

Regularization-based MTL: Decomposition Framework

- In practice, it is too restrictive to constrain all tasks to share a single shared structure.
- Assumption: the model is the sum of two components $W = P + Q$
  - A shared low dimensional subspace and a task specific component (Ando and Zhang, 2005, JMLR)
  - A group sparse component and a task specific sparse component (Jalali et.al., 2010, NIPS)
  - A low rank structure among relevant tasks + outlier tasks (Gong et.al., 2011, KDD)

How Tasks Are Related?

- All tasks are related
  - Models of all tasks are close to each other;
  - Models of all tasks share a common set of features;
  - Models share the same low rank subspace

- Structure in tasks
  - clusters / graphs / trees

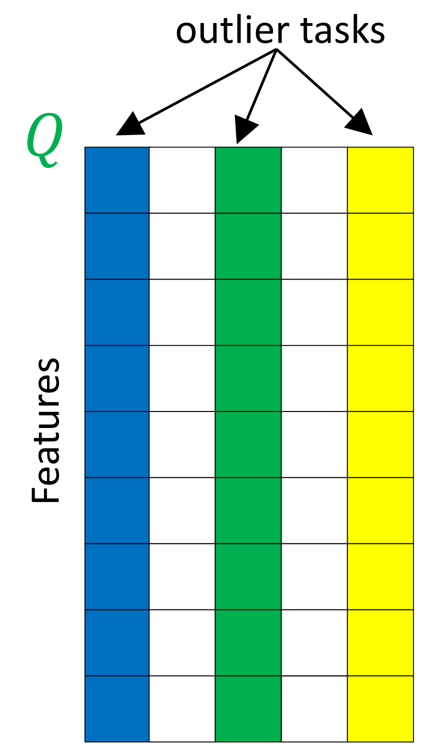
- Learning with outlier tasks

MTL Methods: Robust MTL

Chen *et. al.* 2011 KDD

<u>Assumption:</u> models share the same low-rank subspace + outlier tasks

$$W = P + Q$$



Regularization

- $\|P\|_*$: nuclear norm
- $\|Q\|_{2,1} = \sum_{j=1}^{m} \left\|\boldsymbol{q}_{:,j}\right\|_2$

$$\min_W Loss(W) + \underbrace{\alpha\|P\|_*}_{\text{low rank}} + \underbrace{\beta\|Q\|_{2,1}}_{\text{column-sparse}}$$

Summary So Far...

- All multi-task learning formulations discussed above can fit into the $W = P + Q$ schema.
  - Component P: shared structure
  - Component Q: information not captured by the shared structure

Outline

- Introduction to multi-task learning (MTL): problem and models
- Multi-task learning with task-feature co-clusters
- Low-rank optimization in multi-task learning
- Multi-task learning applied to trajectory regression

Recap: How Tasks Are Related?

- All tasks are related
  - Models of all tasks are close to each other;
  - Models of all tasks share a common set of features;
  - Models share the same low rank subspace

- Structure in tasks
  - clusters / graphs / trees

- Learning with outlier tasks

Task-level

How Tasks are Related

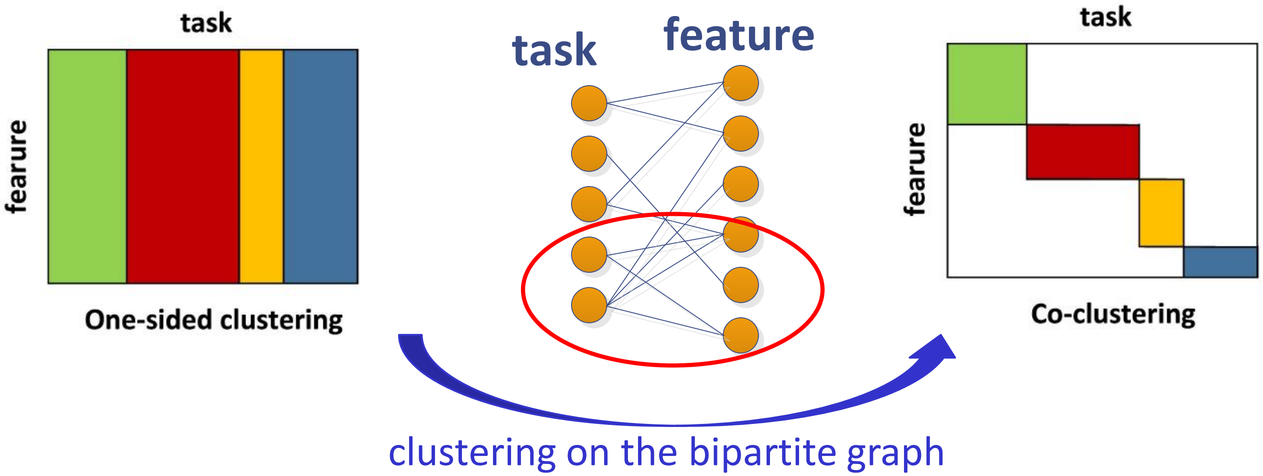
- Existing methods consider the structure at a general task-level

- Restrictive assumption in practice:
  - In document classification: different tasks may be relevant to different sets of words

  - In a recommender system: two users with similar tastes on one feature subset may have totally different preference on another subset

CoCMTL: MTL with Task-Feature Co-Clusters [Xu. et al, AAAI15]

- Motivation: feature-level groups



- Impose task-feature co-clustering structure with $Reg(W)$

CoCMTL: Model

- Decomposition model: $W = P + Q$

$$\min_{W} Loss(W) + \lambda_1\, \Omega_1(P) + \lambda_2\, \Omega_2(Q)$$

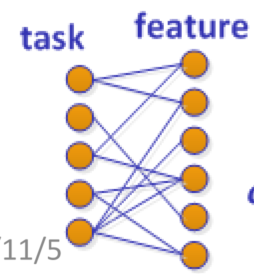
Global Similarities: $\Omega_1(P) = \sum_{i=1}^{m} \left\| p^i - \frac{1}{m}\sum_{j=1}^{m} p^j \right\|_2^2 = tr(PLP^T)$

Group Specific Similarities:

K-means Clustering with Spectral Relaxation:

$$\min_{H^T H = I} \left\{ tr(Z^T Z) - tr(H^T Z^T Z H) \right\}$$

- Z: data matrix.
- H: indicating matrix.



clustering on the bipartite graph

Co-clustering Scenario:

$$Z = \begin{pmatrix} 0 & Q \\ Q^T & 0 \end{pmatrix} \quad H = \begin{pmatrix} F \\ G \end{pmatrix} \quad \begin{matrix} F^T F = I \\ G^T G = I \end{matrix}$$

- F, G: indicating matrices for tasks and features.

$$\min_{F^T F = I,\, G^T G = I} \left\{ 2\|Q\|_F^2 - tr(F^T Q Q^T F) - tr(G^T Q^T Q G) \right\}$$

CoCMTL: Model

- Decomposition model: $W = P + Q$

$$\min_{W} Loss(W) + \lambda_1\, \Omega_1(P) + \lambda_2\, \Omega_2(Q)$$

Theorem 1. *For any given matrix $Q \in \mathbb{R}^{d\times m}$, any matrices $F \in \mathbb{R}^{d\times k}$, $G \in \mathbb{R}^{m\times k}$ and any nonnegative integer k, $k \le \min(d,m)$, Problem (4) reaches its minimum value at $F = (\mathbf{u}_1, \dots, \mathbf{u}_k)$, $G = (\mathbf{v}_1, \dots, \mathbf{v}_k)$, where $\mathbf{u}_i$ and $\mathbf{v}_i$ are the i-th left and right singular vectors of Q respectively. The minimum value is $2\sum_{i=k+1}^{\min(d,m)} \sigma_i^2(Q)$, where $\sigma_1(Q) \ge \sigma_2(Q) \ge \cdots \ge \sigma_{\min(d,m)}(Q) \ge 0$ are the singular values of Q.*

non-convex

$$\Omega_2(Q) = \sum_{i=k+1}^{\min(d,m)} \sigma_i^2(Q)$$

$$\min_{W} Loss(W) + \lambda_1\, \mathrm{tr}(PLP^T) + \lambda_2 \sum_{i=k+1}^{\min(d,m)} \sigma_i^2(Q)$$

CoCMTL: Optimization

- We follow the ***Proximal Alternative Linear Method (PALM)*** to solve the non-convex problem.

Formulation:

$$\min_{W,P,Q} \underbrace{\ell(W)+\lambda_1\Omega_1(P)}_{h(P,Q)}+\underbrace{\lambda_2\Omega_2(Q)}_{f(Q)}$$

General Form

$$\min_{P,Q} h(P,Q)+g(P)+f(Q)$$

$h(P,Q)$ *: convex*

$g(P), f(Q)$ *: lower semi continuous*

Specific Form

In the *r*-th iteration, we get two sub-problems.

$$\begin{cases} P_r = \arg\min_P \frac{\gamma_r}{2}\|P-C_P\|_F^2 \\ Q_r = \arg\min_Q \frac{\gamma_r}{2}\|Q-C_Q\|_F^2 + \lambda_2\Omega_2(Q) \end{cases}$$

$$\arg\min_Q \frac{\gamma_r}{2}\|Q-C_Q\|_F^2 + \lambda_2 tr(F_*^T QQ^T F_*)$$

Alternative Optimization

CoCMTL: Results

School data: #Tasks 139, #Features 27, #Samples 15k

	Training Ratio	Ridge	L21	Low Rank	rMTL	rMTFL	Dirty	Flex-Clus	CMTL	CoCMTL
nMSE	10%	1.1031	1.0931	0.9693	0.9603	1.3838	1.1421	0.8862	0.9914	**0.8114**
	20%	0.9178	0.9045	0.8435	0.8198	1.0310	0.9436	0.7891	0.8462	**0.7688**
	30%	0.8511	0.8401	0.8002	0.7833	0.9103	0.8517	**0.7634**	0.8064	**0.7515**
aMSE	10%	0.2891	0.2867	0.2541	0.2515	0.3618	0.2983	0.2315	0.2593	**0.2118**
	20%	0.2385	0.2368	0.2207	0.2147	0.2702	0.2470	0.2062	0.2214	**0.2009**
	30%	0.2212	0.2197	0.2091	0.2049	0.2378	0.2225	**0.1992**	0.2107	**0.1961**
rMSE	10%	11.5321	11.5141	11.2000	11.1984	12.1233	11.6401	10.9991	11.2680	**10.7430**
	20%	10.7318	10.7011	10.5427	10.4866	10.9928	10.8033	10.3986	10.5500	**10.3110**
	30%	10.1831	10.1704	10.0663	10.0291	10.3338	10.1956	**9.9767**	10.0865	**9.9221**

Outline

- Introduction to multi-task learning (MTL): problem and models
- Multi-task learning with task-feature co-clusters
- Low-rank optimization in multi-task learning
- Multi-task learning applied to trajectory regression

Recap: Low-Rank MTL

<u>Assumption:</u> in high dimensional feature space, the linear models share the same low-rank subspace

Regularization - Rank minimization formulation

$$\min_{W} Loss(W) + \lambda \cdot \mathrm{rank}(W)$$

– Rank minimization is *NP-Hard* for general loss functions

- Convex relaxation: nuclear norm minimization

$$\min_{W} Loss(W) + \lambda \, \|W\|_*$$

($\|W\|_*$: sum of singular values of W)

More on Nuclear Norm

Rank minimization formulation

$$\min_{W} Loss(W) + \lambda \cdot \mathrm{rank}(W)$$

- $\mathrm{rank}(W) =$ #non-zero singular values
- $\|W\|_* = \sum \sigma_i(W)$: sum of singular values

- Limitation of $\|W\|_*$
  - Large singular values are penalized more heavily
  - Large singular values are dominant in determining the properties of a matrix

Idea: Weighted Nuclear Norm

[Zhong et al, AAAI15; Xu et al, ICDM16]

Non-convex

$$\min_{W} Loss(W) + \lambda \boxed{\sum_i p_i \sigma_i(W)}$$

- Intuition: penalize large singular values less
  - Non-descending weights p_i

- Reweighting strategy:
  - Given current weights $\boldsymbol{p}^{k-1}$, solve for W^{k-1}
  - Reweighting of $\boldsymbol{p}$
    - $p_i^k = \frac{r}{\left(\sigma_i(W^k)+\epsilon\right)^{1-r}}$, where $0 < r < 1, \epsilon > 0$
    - Each weight inversely proportional to the corresponding singular value

Idea: Weighted Nuclear Norm

$$\min_{W} Loss(W) + \lambda \boxed{\sum_i p_i \sigma_i(W)}$$

$$p_i^k = \frac{r}{(\sigma_i(W^k) + \epsilon)^{1-r}}$$

$$\min_{W} Loss(W) + \lambda \boxed{\sum_i (\sigma_i(W) + \epsilon)^r}$$

Enhances low rank approximation

$\to \mathrm{rank}(W)$ when $\epsilon \to 0, r \to 0$

Optimization: Proximal Operator

First-order approximation of $Loss(W)$, regularized by a proximal term

$$P_{t^k}(W, W^k) = Loss(W^k) + \langle W - W^k, \nabla Loss(W^k) \rangle + \frac{t^k}{2} \|W - W^k\|^2$$

Generate the sequence

$$W^k = \arg\min_W \frac{t^k}{2\lambda} \left\| W - \left(W^k - \frac{1}{t^k} \nabla Loss(W^k) \right) \right\|_F^2 + (p^k)^T \sigma(W)$$

- ☹ Non-convex proximal operator problem
- ☺ Has closed form solution by exploiting structure of the weighted nuclear norm (unitarily invariant property)

Theorem. Suppose that $A = U\Sigma V^T$, then, $W^* = UD(x^*)V^T$ is a global solution of the problem

$$\min_X \frac{\mu}{2} \|W - A\|_F^2 + p^T \sigma(W)$$

where x^*can be denoted as $x^* = \max\left(\sigma(A) - \frac{1}{\mu} p, 0\right)$

Optimization: Algorithm

Algorithm 1 Iterative Shrinkage-Thresholding and Reweighted Algorithm (ISTRA)

Input: $0 < t_{\min} < t_{\max}$, $0 < \tau < 1$, $0 < r < 1$, $\lambda > 0$, $\delta > 0$, $\epsilon > 0$, $\rho > 1$

Output: X^*

Initialize: $k = -1$, $w^0 = \mathbf{1}^T$, X^{-1}, X^0

repeat

  $k = k + 1$ Barzilai Borwein (BB) rule

  update t^k

  make $t^k \in [t_{\min}, t_{\max}]$

  while true **do**

    update X^{k+1}

    if line search criterion is satisfied **then**

      Break;

    end if

    $t^k = \rho t^k$

  end while decrease the step size

  update the weights w_i^{k+1}, $i = 1 \cdots q$ reweighting strategy

until stop criterion $\boxed{\|X^{k+1} - X^k\|^2 \leq \delta}$ is satisfied

Convergence Analysis

- Critical points

Theorem. The sequence $\{W^k\}$ generated by the ISTRA algorithm makes the objective function monotonically decrease, and all accumulation points (i.e. the limit points of convergent subsequence in $\{W^k\}$) are critical points (i.e. 0 belongs to the subgradients)

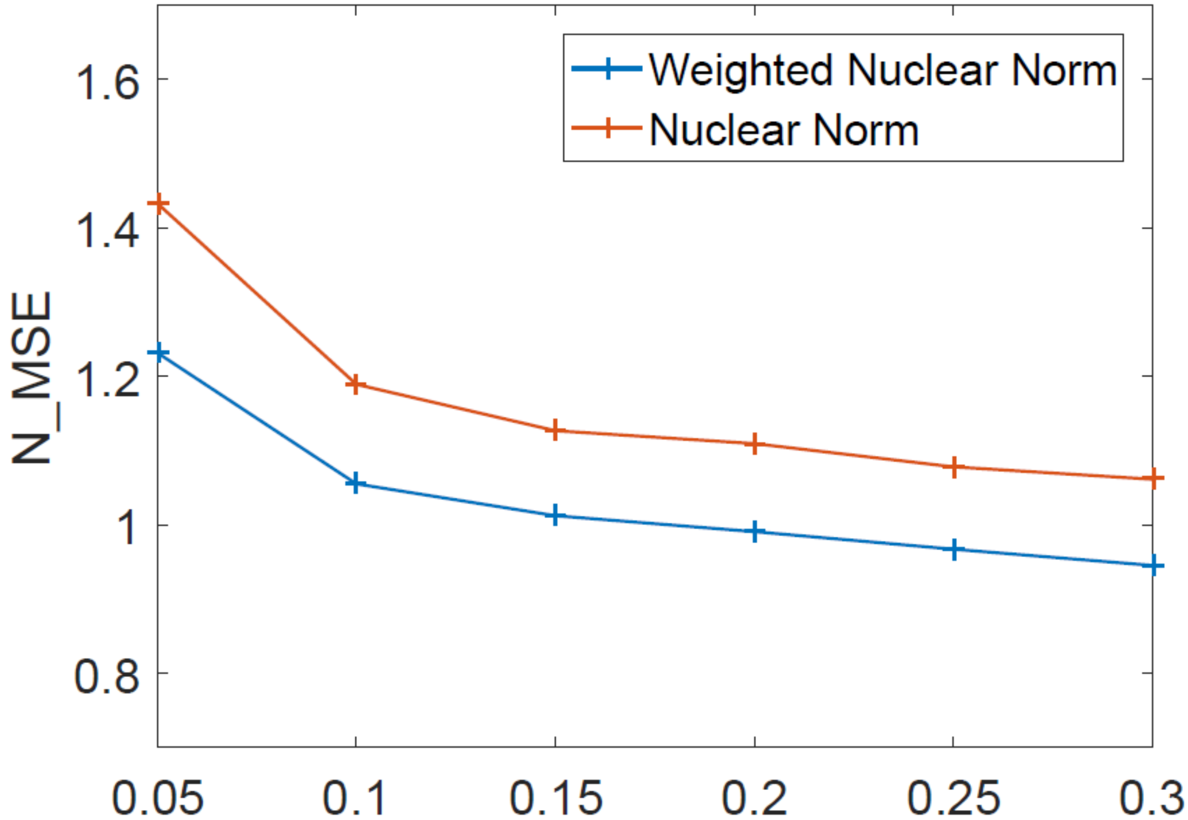
- Sublinear convergence rate

Theorem. Suppose that $\{W^k\}$ is the sequence generated by the ISTRA algorithm, and W^* is an accumulation point of $\{X^k\}$, then

$$\min_{0\le k\le n}\left\|W^{k+1}-W^k\right\|^2 \le 2\big(g(W^0)-g(W^*)\big)/n\tau t_{\min}$$

Results

School data



Outline

- Introduction to multi-task learning (MTL): problem and models
- Multi-task learning with task-feature co-clusters
- Low-rank optimization in multi-task learning
- Multi-task learning applied to trajectory regression

Trajectory Regression: Problem

Trajectory:

A sequence of link (road segments), where any two consecutive links share an intersection

Goal:

Estimate the total travel time of an arbitrary trajectory

Trajectory Regression: Problem

Given a set consisting of N trajectory-cost pairs:

$$D \equiv \{(\boldsymbol{x}_i, y_i) | i = 1,2, \dots, N\}, \boldsymbol{x}_i \in \mathbb{R}_d$$

- Each feature of $\boldsymbol{x}_i$ corresponds to a link — distance traveled along the link

Road Seg. 1	Road Seg. 2	Road Seg. 3	...	Road Seg. d
5km	2km	1km	...	0

$\rightarrow$

Cost
20min

Goal: Learn the weights $w \in \mathbb{R}_d$ that encode the cost per distance unit for each link

Single task learning: $\min_{\boldsymbol{w}} \|Y - X\boldsymbol{w}\|_2^2 + \beta \|\boldsymbol{w}\|_2^2$

Trajectory Regression: Key Challenges

- Dynamic: costs of road segments are not static over time
  - Cost of a road segment fluctuates smoothly most of the time
  - Costs can be abruptly different between peak periods and off-peak periods
- Trajectories are extremely sparse
  - A driving path spans just a small fraction of road segments
- Insufficient instances

Trajectory Regression: Idea

[Huang et al. ICDM14]

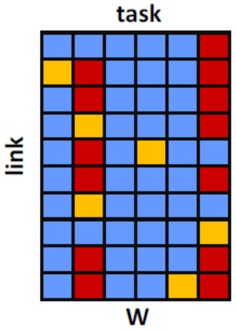
Dynamic trajectory regression in an MTL framework

- Divide D into m disjoint subsets ordered by time
- Multi-task learning framework: each time slot corresponds to a task
  - leverage the inherent relations of tasks to enhance the predictive performance, especially when the data samples are insufficient

Trajectory Regression

$$\min_{W} \sum_{i} Loss(W, X^i, Y^i) + \lambda\, Reg(W) = \min_{W} \sum_{i} \|Y^i - X^i \boldsymbol{w}^i\| + \lambda\, Reg(W)$$

W Structure in the trajectory regression problem

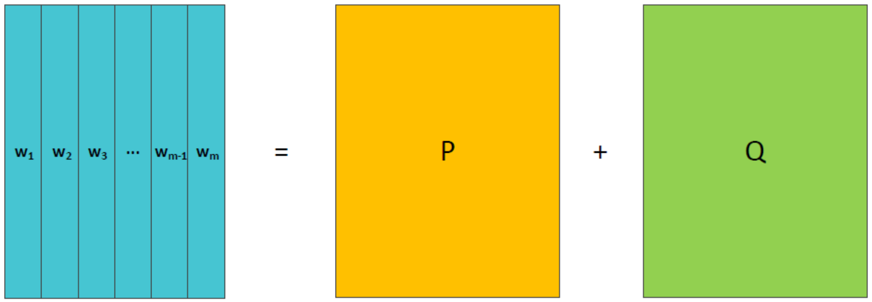


- Global temporal smoothness:
  - Link costs change smoothly most of the time
- Global spatial smoothness:
  - Costs are similar if the two corresponding links are close to each other
- Local temporal patterns:
  - Significant temporal changes in rush hours

Trajectory Regression - Additive Model

$$\min_{W} \sum_{i} Loss(W, X^i, Y^i) + \lambda \, Reg(W) = \min_{W} \sum_{i} \|Y^i - X^i \mathbf{w}^i\| + \lambda \, Reg(W)$$

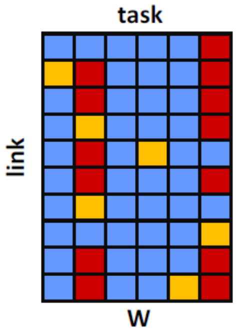
$$W = P + Q$$



$W = [\mathbf{w}^1, \mathbf{w}^2, \ldots, \mathbf{w}^m]$

- P: models the global smoothness over links and time
- Q: captures the local "outliers" including rush hours

Trajectory Regression - Regularization



$W = P + Q$

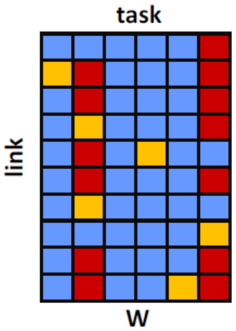
- P: models the global smoothness over links and time

- Global temporal smoothness

$$\Omega_1 = \sum_{t=1}^{m} \left\| P_{:,t} - \frac{1}{m} \sum_{r=1}^{m} P_{:,r} \right\|_2^2 = tr(P L_1 P^T) \qquad L_1 = I - \frac{1}{m} \mathbf{1}\mathbf{1}'$$

  - Enforces the columns of P or the tasks to be similar with some discrepancy

Trajectory Regression - Regularization



$W = P + Q$

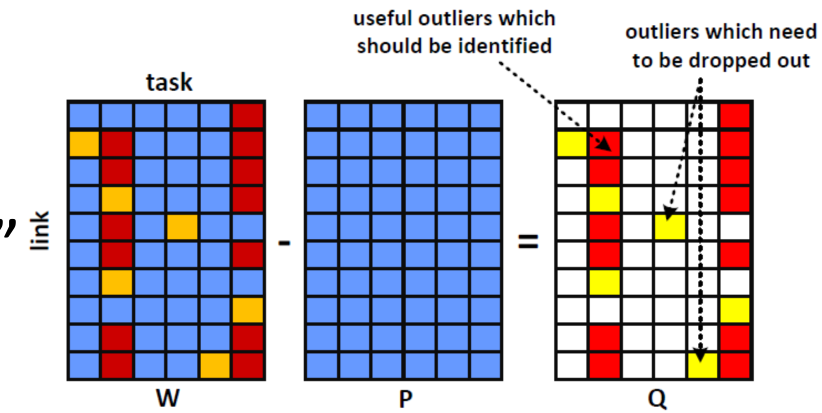
- P: models the global smoothness over links and time

- Global spatial smoothness

$$\Omega_2 = \sum_{i,j=1}^{d} S_{ij} \left\| P_{i,:} - P_{j,:} \right\|_2^2 = tr(P^T L_2 P)$$

  - S measures the spatial closeness of links
  - Costs are similar if the two corresponding links are close to each other

Trajectory Regression - Regularization



$W = P + Q$

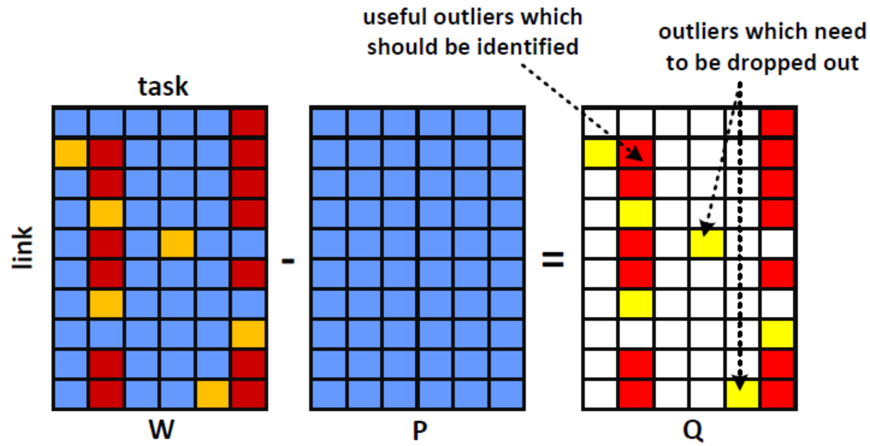
- Q: captures the local "outliers"

- Local significant temporal transitions

  $\Omega_3 = \|Q\|_{\infty,1}$

  - $\|Z\|_{\infty,1} = \sum_j \|Z_{:,j}\|_{\infty}, \quad \|Z_{:,j}\|_{\infty} = \max_i |Z_{ij}|$
  - Enforces column sparsity to identify peak traffic
  - The $\ell_{\infty,1}$ norm is only influenced by the maximum elements of the nonzero columns — the cost of a trajectory is mostly decided by the link with highest cost during traffic peaks
  - Leaves out the outliers — ROBUST

Trajectory Regression - Model



$$\min_W \sum_{i=1}^{m} \|Y^i - X^i \boldsymbol{w}^i\|_2^2 + \lambda_1 tr(PL_1P^T) + \lambda_2 tr(P^T L_2 P) + \lambda_3 \|Q\|_{\infty,1}$$

Trajectory Regression - Optimization

$$\min_{W} \sum_{i=1}^{m} \|Y^i - X^i \boldsymbol{w}^i\|_2^2 + \lambda_1 tr(PL_1P^T) + \lambda_2 tr(P^T L_2 P) + \lambda_3 \|Q\|_{\infty,1}$$

Convex problem, but non-trivial for optimization due to the $\ell_{\infty,1}$ term

Proximal Method:

$$\min_{W}\{F(W) + R(W)\} \begin{cases} F(W) = L(W) + \lambda_1 tr(PL_1P^T) + \lambda_2 tr(P^T L_2 P) \\ R(W) = \lambda_3 \|Q\|_{\infty,1} \end{cases}$$

$$P_r = arg\min_{P} \frac{\gamma_r}{2} \|P - C_P(P_{r-1})\|_F^2 \,,$$

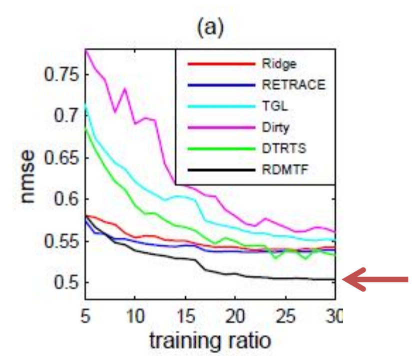
$$Q_r = arg\min_{Q} \frac{\gamma_r}{2} \left\|Q - C_Q(Q_{r-1})\right\|_F^2 + \lambda_3 \|Q\|_{\infty,1}$$

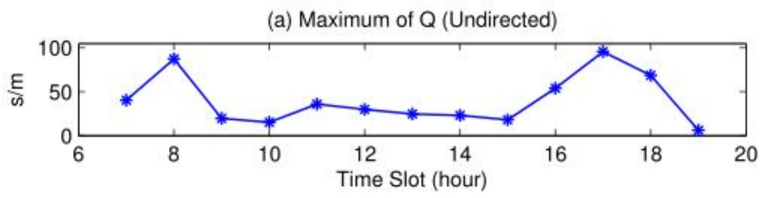
$$\boxed{\min_{\boldsymbol{q}^i} \frac{1}{2}\left\|\boldsymbol{q}^i - \boldsymbol{c}^i\right\|_2^2 + \lambda \|\boldsymbol{q}^i\|_\infty \xrightarrow[\boldsymbol{c} = prox_R(\boldsymbol{c}) + prox_{R*}(\boldsymbol{c})]{\text{Moreau Decomposition}} \min_{\boldsymbol{q}^i} \left\{\boldsymbol{c}^i - \left(\frac{1}{2}\left\|\boldsymbol{q}^i - \boldsymbol{c}^i\right\|_2^2 + \lambda \|\boldsymbol{q}^i\|_1\right)\right\}}$$

Trajectory Regression - Results

- Suzhou Traffic Data
  - Contains 59593 trajectory records of 4797 taxies from 7:00 to 19:59 in urban area of Suzhou during the first week in March, 2012

Training Ratio	20%(nmse)	30%(nmse)	40%(nmse)
Ridge	0.549 ± 0.013	0.547 ± 0.019	0.540 ± 0.025
STL-Ridge	0.617 ± 0.027	0.589 ± 0.040	0.560 ± 0.029
RETRACE	0.546 ± 0.013	0.545 ± 0.022	0.538 ± 0.027
STL-RETRACE	0.668 ± 0.035	0.628 ± 0.042	0.587 ± 0.036
TGL	0.570 ± 0.032	0.581 ± 0.063	0.538 ± 0.046
Dirty	0.626 ± 0.034	0.615 ± 0.056	0.590 ± 0.034
DTRTS	0.612 ± 0.051	0.596 ± 0.048	0.531 ± 0.022
L2,1	0.525 ± 0.024	0.562 ± 0.068	0.525 ± 0.049
RDMTF	**0.494 ± 0.014**	**0.498 ± 0.040**	**0.481 ± 0.023**





Summary

- Multi-task Learning (MTL)
  - MTL is preferred when dealing with multiple related tasks with small number of training samples
  - Key issue of MTL: Exploiting relationships among the tasks
- Optimization
  - General formulations, classical algorithms apply
  - Distributed optimization
- Applications
  - Task relationships are specific to the nature of the problem

References

- Xu, L., Huang, A., Chen, J., and Chen, E. **Exploiting Task-Feature Co-Clusters in Multi-Task Learning.** In *Proceedings of the 29th National Conference on Artificial Intelligence (AAAI-15).*
- Zhong, X., Xu, L., Li., Y, Liu, Z., and Chen, E. **A Nonconvex Relaxation Approach for Rank Minimization Problems.** In *Proceedings of the 29th National Conference on Artificial Intelligence (AAAI-15)*.
- Xu, L., Chen, Z., Zhou, Q., Chen, E., Yuan, N.J., Xie, X. **Aligned Matrix Completion: Integrating Consistency and Independency in Multiple Domains.** In *Proceedings of the 16th IEEE Conference on Data Mining (ICDM-16).*
- Huang, A., Xu, L., Li., Y, and Chen, E. **Robust Dynamic Trajectory Regression on Road Networks: A Multi-Task Learning Framework.** In *Proceedings of the 14th IEEE Conference on Data Mining (ICDM-14).*

- Zhou, J., Chen, J., and Ye, J. **Multi-Task Learning: Theory, Algorithms, and Applications.** *SDM 2012 Tutorial*.
- Evgeniou, T., and Pontil, M. **Regularized Multi-Task Learning**. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-04).*
- Argyriou, A., Evgeniou, T., and Pontil, M. **Multi-Task Feature Learning**. In *Advances in Neural Information Processing Systems 19 (NIPS-06).*
- Argyriou, A., Evgeniou, T., and Pontil, M. **Convex Multi-Task Learning**. Machine Learning, 73, 2008.
- Ji, S., and Ye, J. **An Accelerated Gradient Method for Trace Norm Minimization**. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML-09).*
- Zhou, J., Chen, J., and Ye, J. **Clustered Multi-Task Learning Via Alternating Structure Optimization**. In *Advances in Neural Information Processing Systems 24 (NIPS-11).*
- Chen, J., Zhou, J., and Ye, J. **Integrating Low-Rank and Group-Sparse Structures for Robust Multi-Task Learning**. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-11).*

Joint Work With

Students

- Aiqing Huang
- Xiaowei Zhong
- Zaiyi Chen
- Yitan Li

Collaborators

- Enhong Chen
- Jianhui Chen

Thanks!

Email: linlixu@ustc.edu.cn