Learning with Whom to Share in Multi-task Feature Learning

Zhuoliang Kang
University of Southern California
zkang@usc.edu

Kristen Grauman
University of Texas Austin
grauman@cs.utexas.edu

Fei Sha
University of Southern California
feisha@usc.edu

Motivation

Multi-task learning (MTL)

- Given multiple related tasks
  - Can we do better than learning them ***independently***?
- Main idea
  - Learn multiple tasks ***jointly.***
  - Take advantage of ***relatedness*** between tasks.
- Benefits
  - Improve ***generalization*** performance.
  - Require ***less*** amount of data.

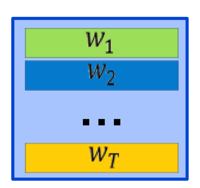
Regularization based approach

Solve a joint optimization problem for ***all tasks***.

Balance between ***total empirical risk*** and ***relatedness***.

$$\min_{W} \sum_{t=1}^{T} loss(w_t, X_t, Y_t) + R(W)$$

matrix W

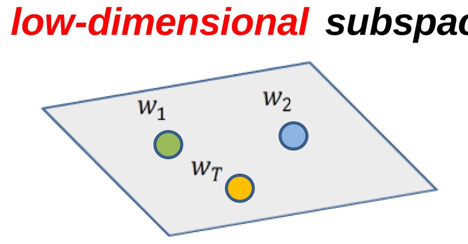


matrix **W** contains parameters of all tasks.

Multi-task feature learning (MTFL)

- Parameters share ***a common low-dimensional subspace.***
- Or equivalently, models share a common feature subspace.
- Structural constraint on matrix **W**: ***low rank***.

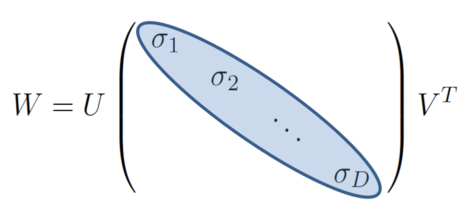
***low-dimensional* subspace**



Low-rank Regularization

- Use ***trace-norm (convex)***: L_1-norm of singular values

$$\|W\|_{tr} = \sum_d |\sigma_d|$$

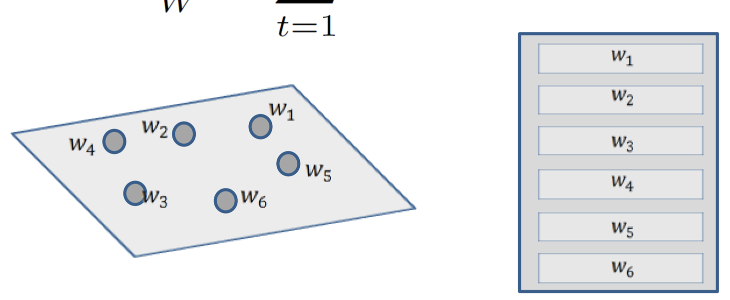


Singular Value Decomposition

Existing Multi-task feature learning

- ***single*** regularization term.
- ***All tasks*** are related.

$$\min_{W} \sum_{t=1}^{T} loss(w_t, X_t, Y_t) + \lambda \|W\|_{tr}^2$$



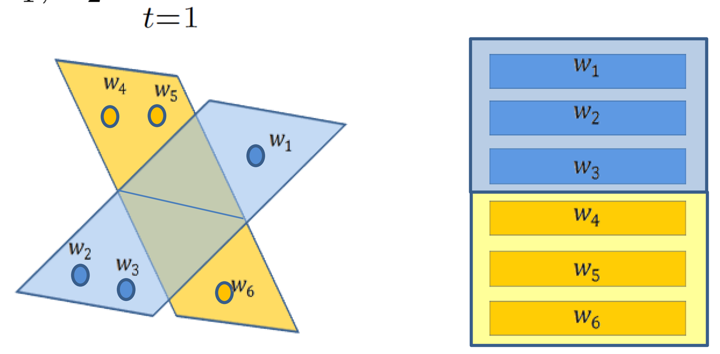
When models are in mixture of subspaces:

- Suboptimal to force with one regularizer.
- Ex: synthetic data in experimental part.

When tasks groups are given:

Regularize each group ***separately.***

$$\min_{W_1, W_2} \sum_{t=1}^{T} loss(w_t, X_t, Y_t) + \lambda \|W_1\|_{tr}^2 + \lambda \|W_2\|_{tr}^2$$

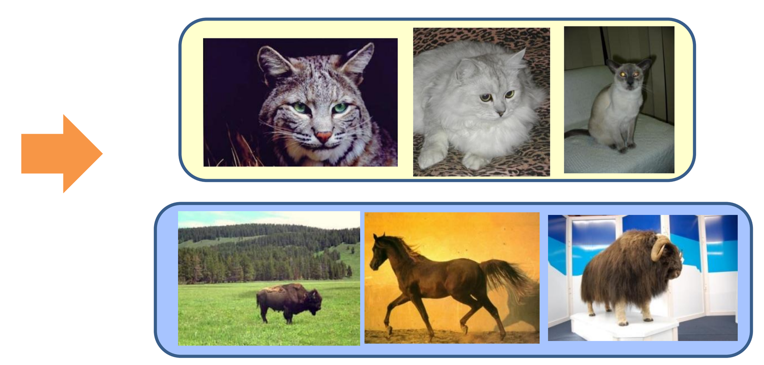


Standard MTFL

Treat all tasks as a single group.

Desiderata

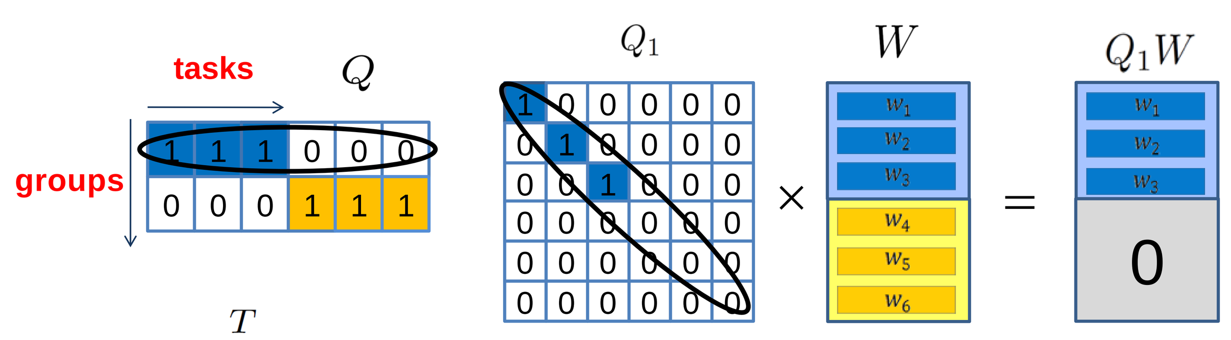
Jointly learn ***tasks grouping*** and ***model parameters***



Algorithm

Step 1: use indicator matrix

Reformulate with task group assignment matrix **Q**.



$$\min_{W,Q} \sum_{t=1}^{T} loss(w_t, X_t, Y_t)$$

$$+ \lambda \|Q_1 W\|_{tr}^2 + \lambda \|Q_2 W\|_{tr}^2 \quad \Leftarrow \quad \|Q_1 W\|_{tr}^2 = \|W_1\|_{tr}^2$$

$$\text{s.t} \quad q_{gt} \in \{0, 1\} \quad \Leftarrow$$

- ***Integer*** constraint

$$Q_1 + Q_2 = I \quad \Leftarrow$$

- ***Hard*** group assignment

[**Cf. Other approaches:** Argyriou et al, ECML, 2008; Rai et al, NIPS Workshop, 2010; Saha et al, AISTATS, 2011]

Step 2: relax the constraint

$$\min_{W,Q} \sum_{t=1}^{T} loss(w_t, X_t, Y_t)$$

$$+ \lambda \|\sqrt{Q_1} W\|_{tr}^2 + \lambda \|\sqrt{Q_2} W\|_{tr}^2$$

$$\text{s.t} \quad 0 \le q_{gt} \le 1$$

$$Q_1 + Q_2 = I$$

- **Approach 1:**
  convex relaxation
  - ***Continuous*** constraint
  - Convex but ***fractional*** solutions
- **Approach 2:**
  (used in experiments)
  non-convex relaxation
  - Use ***square root of Q***
  non-convex but ***integer*** solutions

Numerical optimization

Optimize W and Q *iteratively*

- **Fix Q, update W**
  - *For each group, we solve*

$$\min \sum_{t: q_{gt}=1} \ell(\mathcal{D}_t; \boldsymbol{w}_t) + \gamma \|\boldsymbol{W}_g\|_*^2$$

  - *Use existing algorithm*
    cf. [1]

Convex Multi-Task Feature Learning

Andreas Argyriou[1], Theodoros Evgeniou[2], and Massimiliano Pontil[1]

- **Fix W, update Q**
  - *Use gradient descent*

$$\min_{Q} \sum_{g} \|\sqrt{Q_g} W\|_{tr}^2$$

$$\text{s.t} \quad \sum_g Q_g = I \text{ with } 0 \le q_{gt} \le 1$$

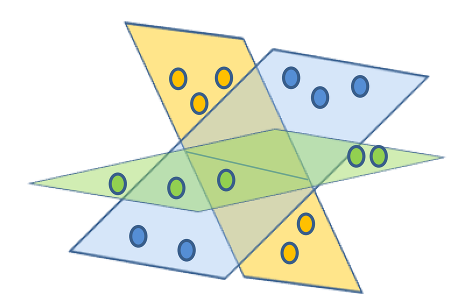
  - *Remove constraints*
    by re-parameterization: ***α is unconstrained***

$$q_{gt} = \frac{e^{\alpha_{gt}}}{\sum_{g=1}^{G} e^{\alpha_{gt}}}$$

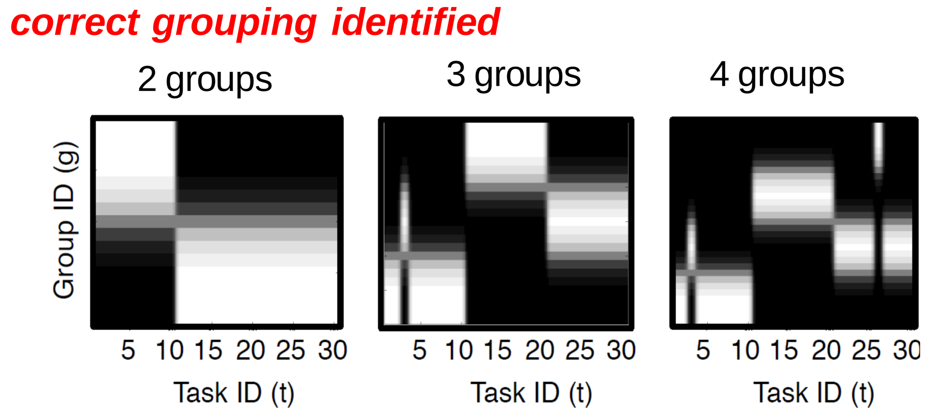
Experiments

Synthetic data

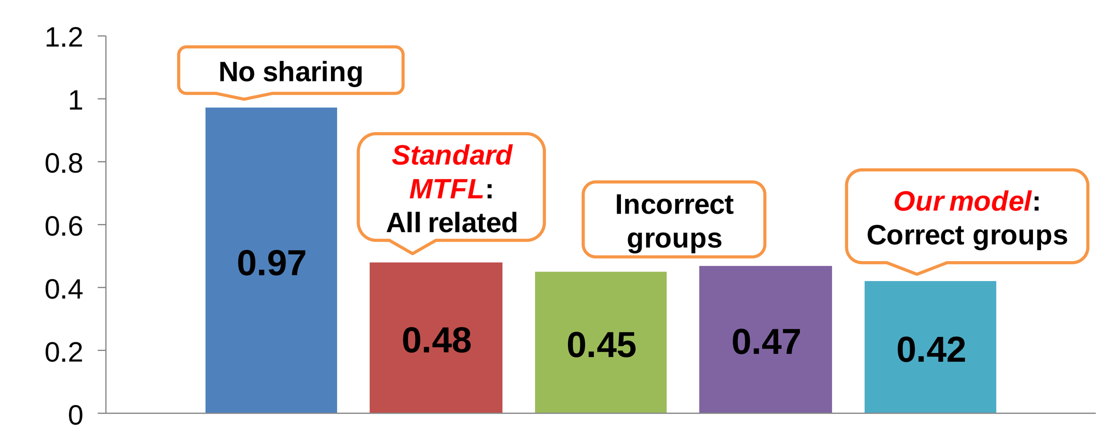
- 30 tasks within 3 groups (10 tasks per group).
- Each task is a regression problem.
- Tasks in the same group ***use the same feature.***



Results: *correct grouping identified*



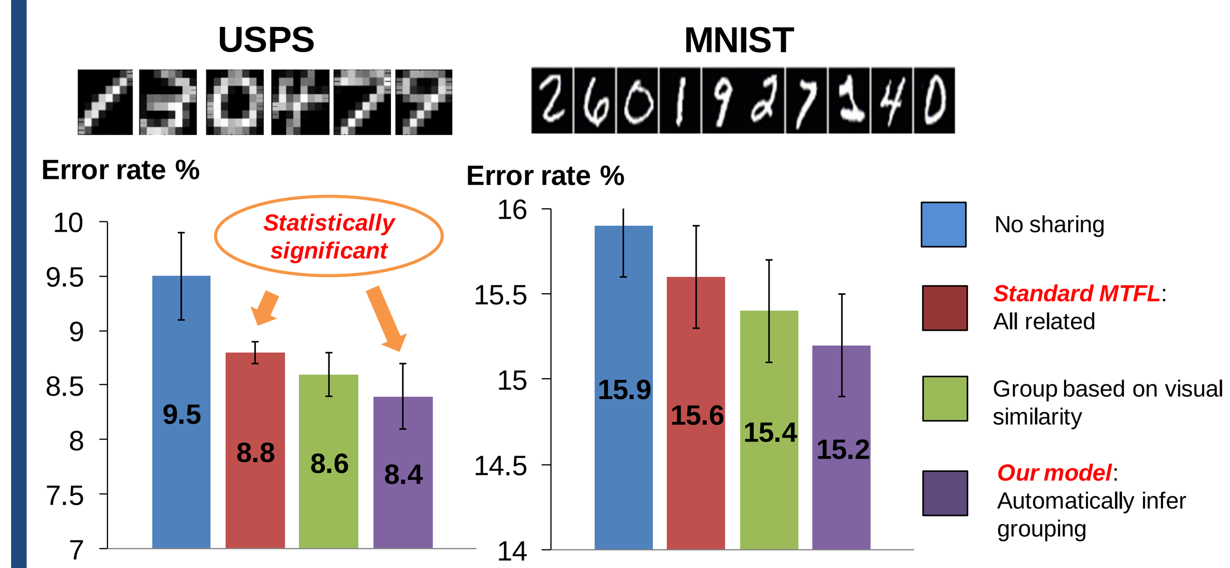
Average root mean square error



Digits data

Setup

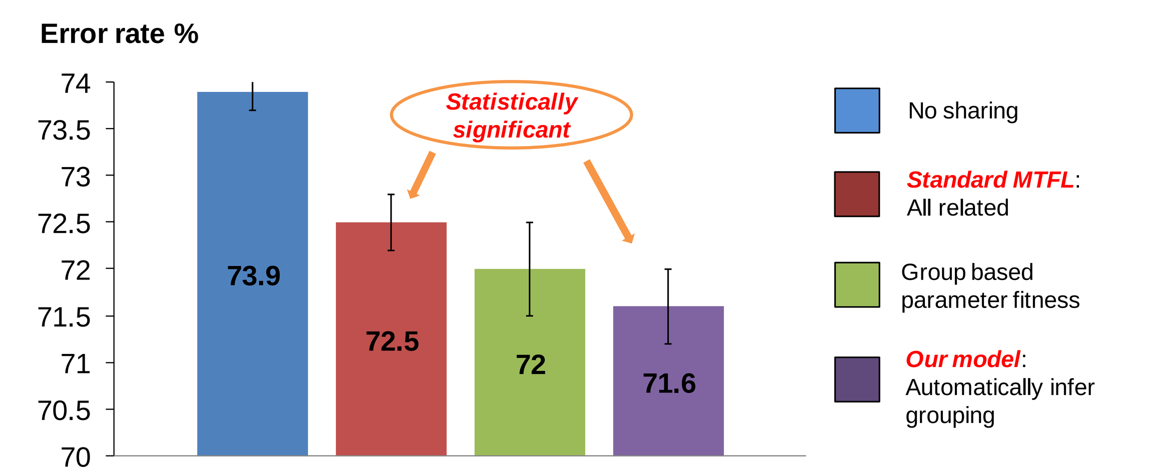
- 10-way classification on images of 10 handwritten digits
- Classifier: binary logistic regression



Animals data

Setup

- Data set: Animal with Attributes (images of 20 classes)
- Classifier: binary logistic regression
- Features: SIFT



Conclusions and Future work

- In many cases, forcing ***all tasks*** to be related is suboptimal.
- Jointly learning ***model parameters*** and ***tasks grouping*** is beneficial.
- In the future, consider more complicated structures.

References

[1]. Argyriou, Andreas, Evgeniou, Theodoros, and Pontil, Massimiliano. ***Convex multi-task feature learning***. Machine Learning, 73:243–272, 2008a.
[2]. Caruana, Rich. ***Multitask learning***. MLJ, 28:41–75, 1997.
[3]. Daumé, III, Hal. ***Bayesian multitask learning with latent hierarchies***. UAI 2009
[4]. Evgeniou, Theodoros and Pontil, Massimiliano. ***Regularized multi–task learning***. KDD 2004.
[5]. Lee, S.I., Chatalbashev, V., Vickrey, D., and Koller, D. ***Learning a meta-level prior for feature relevance from multiple related tasks***. ICML 2007
[6]. Parameswaran, Shibin and Weinberger, Kilian. ***Large margin multi-task metric learning***. NIPS 2010.
[7]. Yu, Kai, Tresp, Volker, and Schwaighofer, Anton. ***Learning gaussian processes from multiple tasks***. ICML 2005.
[8]. Zhang, Y. and Yeung, D.Y. ***A Convex Formulation for Learning Task Relationships in Multi-Task Learning***. UAI, 2010.