MACHINE LEARNING AND APPLICATIONS GROUP

Regularization for Multi-task learning

Miloš Jovanović

some slides taken from:
Zhou,Chen,Ye (2012) Multi-Task Learning: Theory, Algorithms, and Applications
12th SIAM SDM, 2012

Learning objective

- Minimizing loss function:
  - squared error: $$\frac{1}{n}\sum_{i=1}^{n}(y_i - (w^T x_i + w_0))^2$$
  - logistic loss: $$\frac{1}{n}\sum_{i=1}^{n}\log(1 + e^{-y_i (w^T G_i + w_0)})$$
- Useful properties:
  - convexity
  - differentiability
  - smoothness

Regularization

$$\min_{w} \mathcal{L}(\boldsymbol{w}) + \Omega(\boldsymbol{w})$$

- Fighting ill-posed problems:
  - non-unique solutions
  - non-smoothness
- "Penalty", Lagrangian dual
- In learning:
  - Fighting sample variance / overfitting
    => limiting capacity of the model

Null Regularization

- Modified objective:

$$\min_{w} \mathcal{L}(\boldsymbol{w}) + \lambda \|\boldsymbol{w}\|$$

- "Resistance" for parameters to take large values (*Shrinking*)
  - Linear regression
  - Logistic regression
- Prior towards the **null** hypothesis: "no link between input and output" => statistical (scientific) caution (**unbiased**)

Prior Regularization

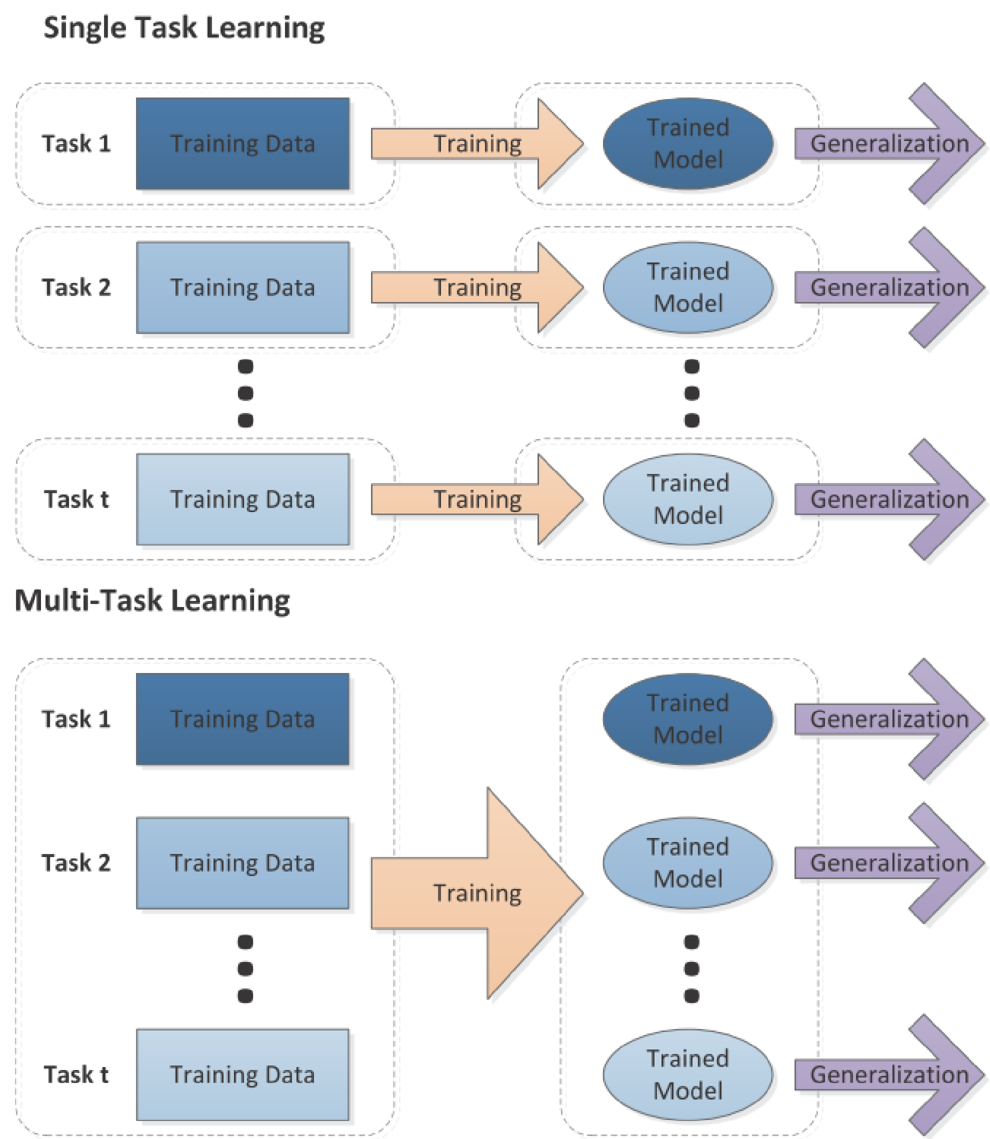
- Prior for parameter values:

$$\min_{w} \mathcal{L}(\boldsymbol{w}) + \lambda \|\boldsymbol{w} - \boldsymbol{w}^{0}\|$$

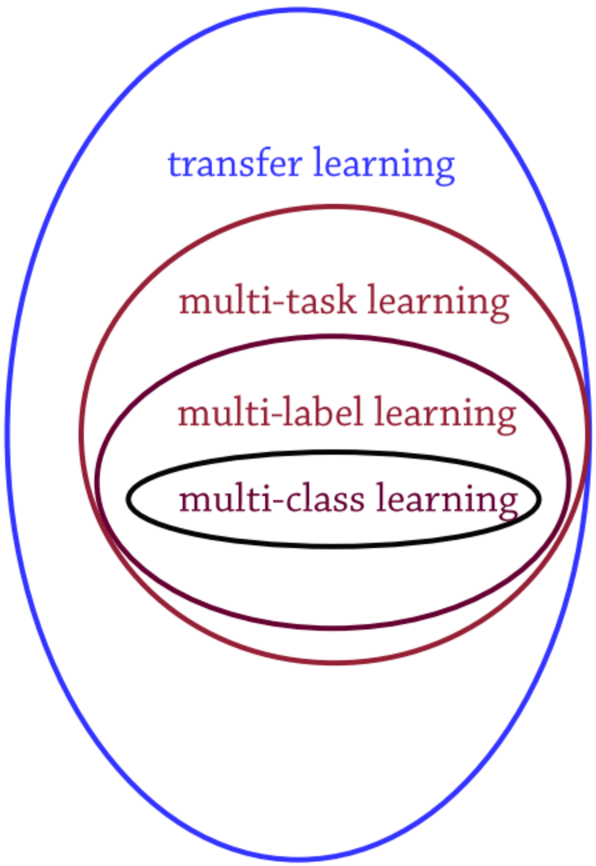
- Prior belief:
  - previous regression parameters!
  - prior assumptions
- Penalty for breaking our prior (Bayesian, Scientific)
  - Data vs Knowledge

Multi-task problem

- Global model
- Local model
- Local model + null-regularization
- Best regularization?
  - more data!
- Regularization for sharing data
  - penalty for being different



Learning Methods

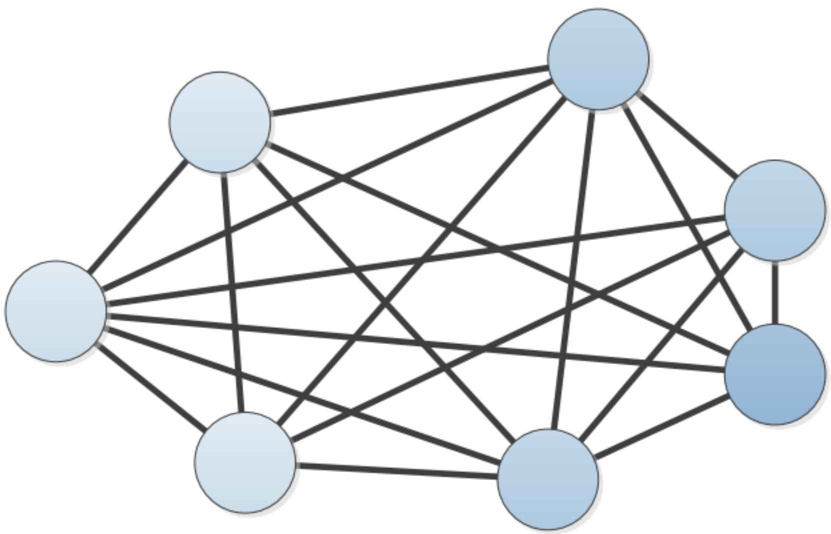


- Transfer Learning
  - Define source & target domains
  - Learn on the source domain
  - Generalize on the target domain
- Multi-task Learning
  - Model the task relatedness
  - Learn all tasks simultaneously
  - Tasks may have different data/features
- Multi-label Learning
  - Model the label relatedness
  - Learn all labels simultaneously
  - Labels share the same data/features
- Multi-class Learning
  - Learn the classes independently
  - All classes are exclusive

Zhou,Chen,Ye (2012)

MULTI-TASK MODELS

How Tasks Are Related



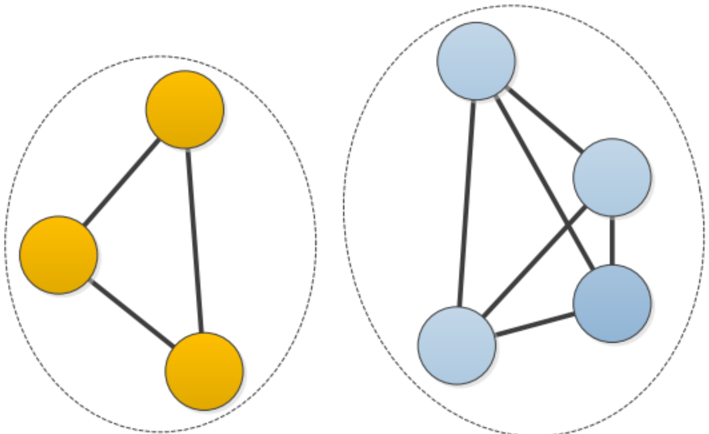
Assumption:
All tasks are related

Methods

- Mean-regularized MTL
- Joint feature learning
- Trace-Norm regularized MTL
- Alternating structural optimization (ASO)
- Shared Parameter Gaussian Process

Zhou,Chen,Ye (2012)

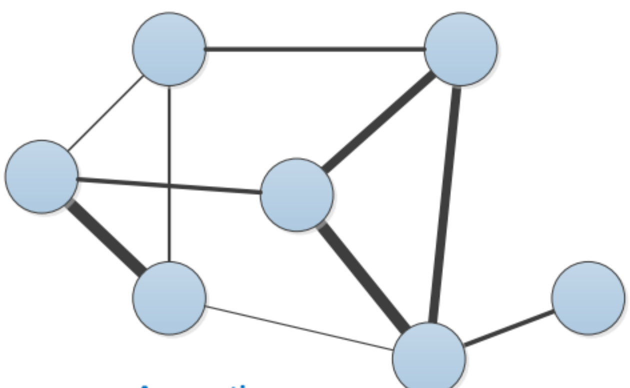
How Tasks Are Related



Assumption:
Tasks have group structures

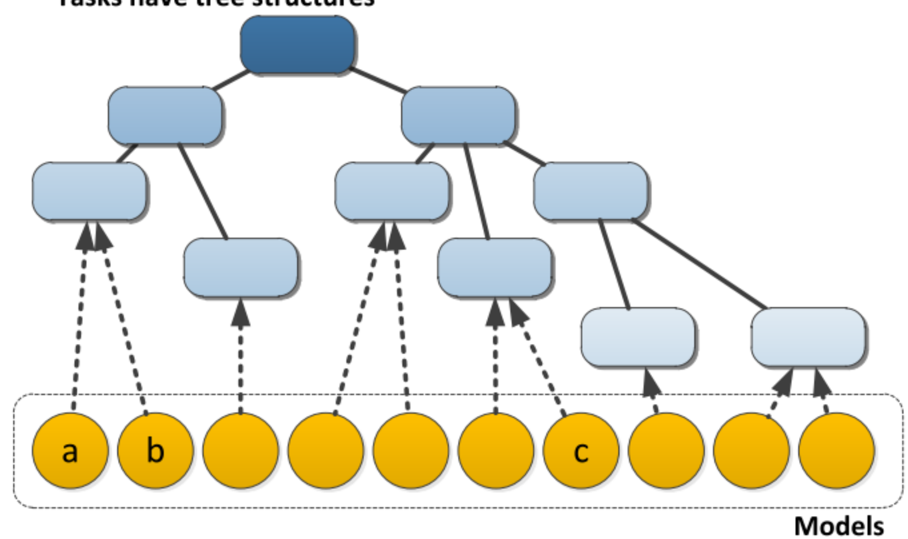
Methods

- Clustered MTL
- Tree MTL
- Network MTL



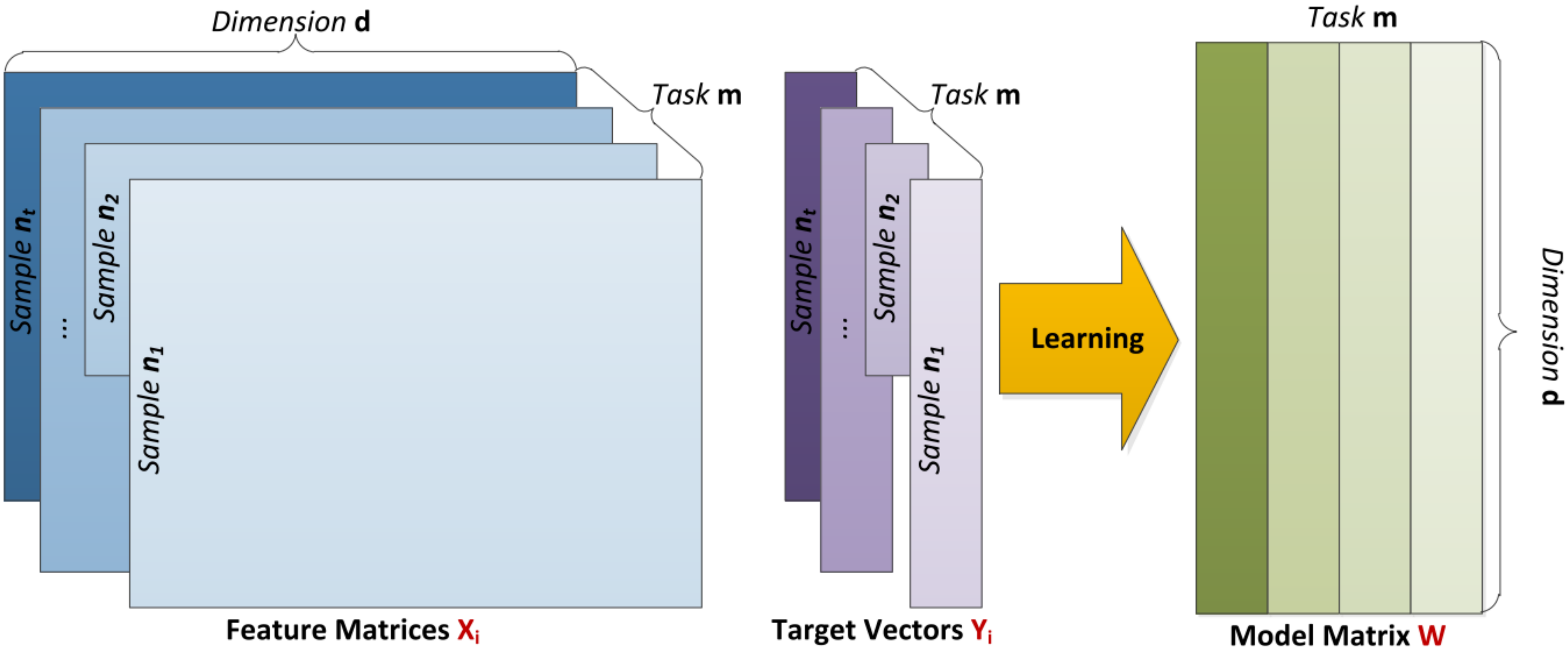
Assumption:
Tasks have graph/network structures

Assumption:
Tasks have tree structures



Zhou,Chen,Ye (2012)

Notation



- We focus on linear models: $Y_i = X_i \times W_i$

  $X_i \in \mathbb{R}^{n_i \times d}, Y_i \in \mathbb{R}^{n_i \times 1}, W = [W_1, W_2, \ldots, W_m]$

Zhou,Chen,Ye (2012)

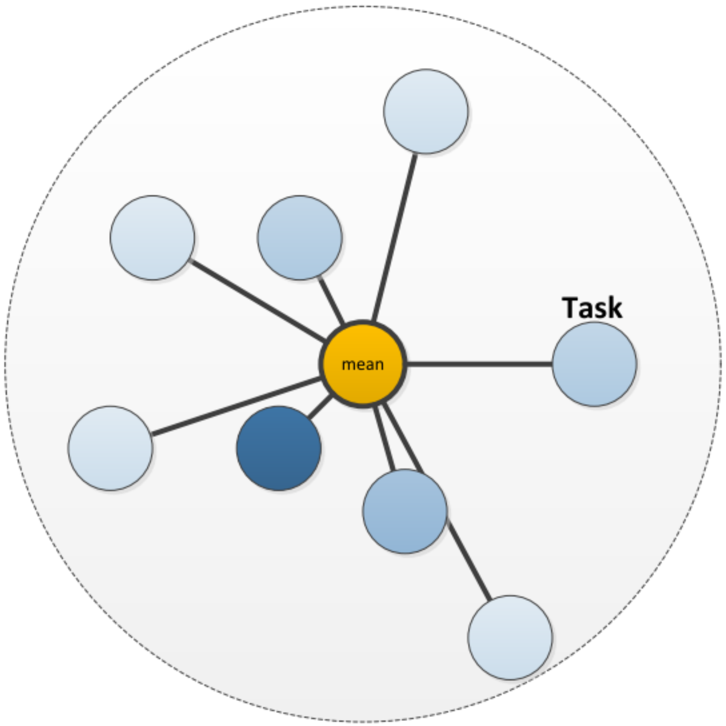
Mean-Regularized Multi-Task Learning

Evgeniou & Pontil, 2004 KDD

- Assumption: task parameter vectors of all tasks are close to each other.
  - Advantage: simple, intuitive, easy to implement
  - Disadvantage: may not hold in real applications.

Regularization
penalizes the deviation of each task from the mean

$$\min_{W} \frac{1}{2}\|XW - Y\|_F^2 + \lambda \sum_{i=1}^{m} \left\| W_i - \frac{1}{m}\sum_{s=1}^{m} W_s \right\|_2^2$$



Zhou,Chen,Ye (2012)

Multi-Task Learning with High Dimensional Data

- In practical applications, we may deal with high dimensional data.
  - Gene expression data, biomedical image data
- Curse of Dimensionality
- Dealing with high dimensional data in multi-task learning
  - Embedded feature selection: L_1/L_q - Group Lasso
  - Low-rank subspace learning: low-rank assumption – ASO, Trace-norm regularization

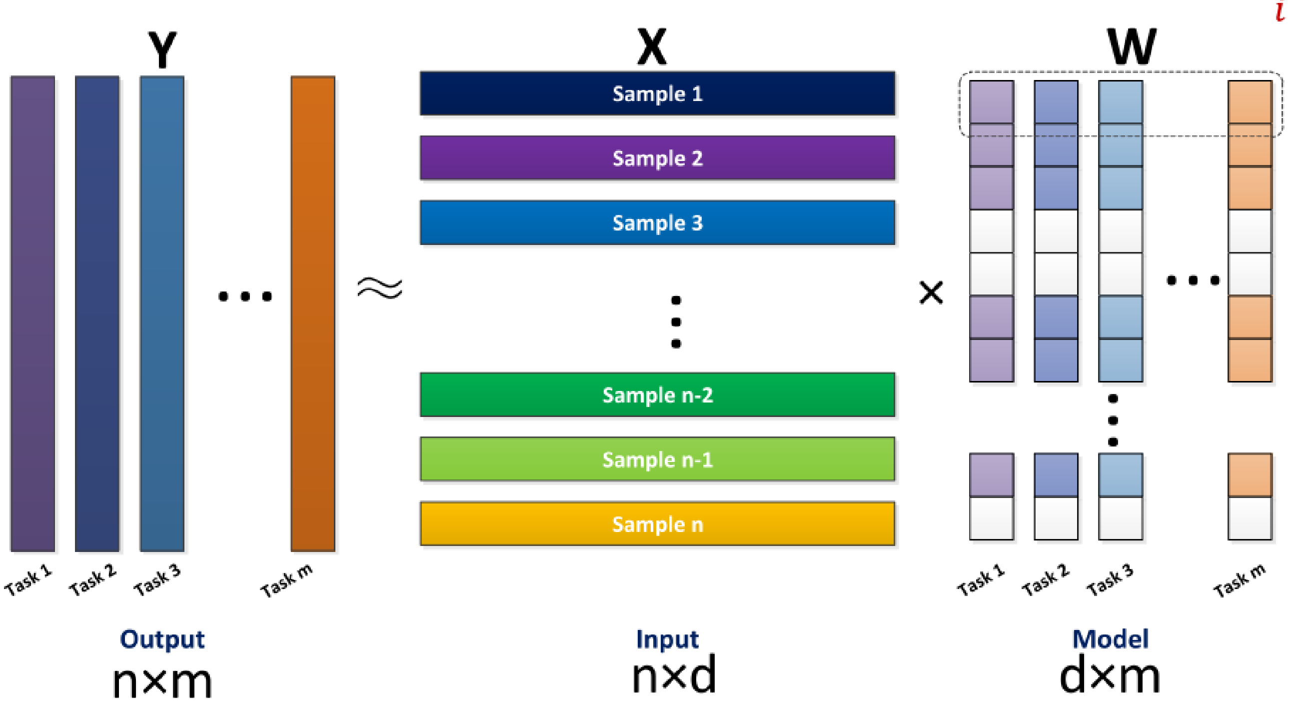
Zhou,Chen,Ye (2012)

Multi-Task Learning with Joint Feature Learning

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

- Using group sparsity: ℓ_1/ℓ_q-norm regularization
- When q>1 we have group sparsity.

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

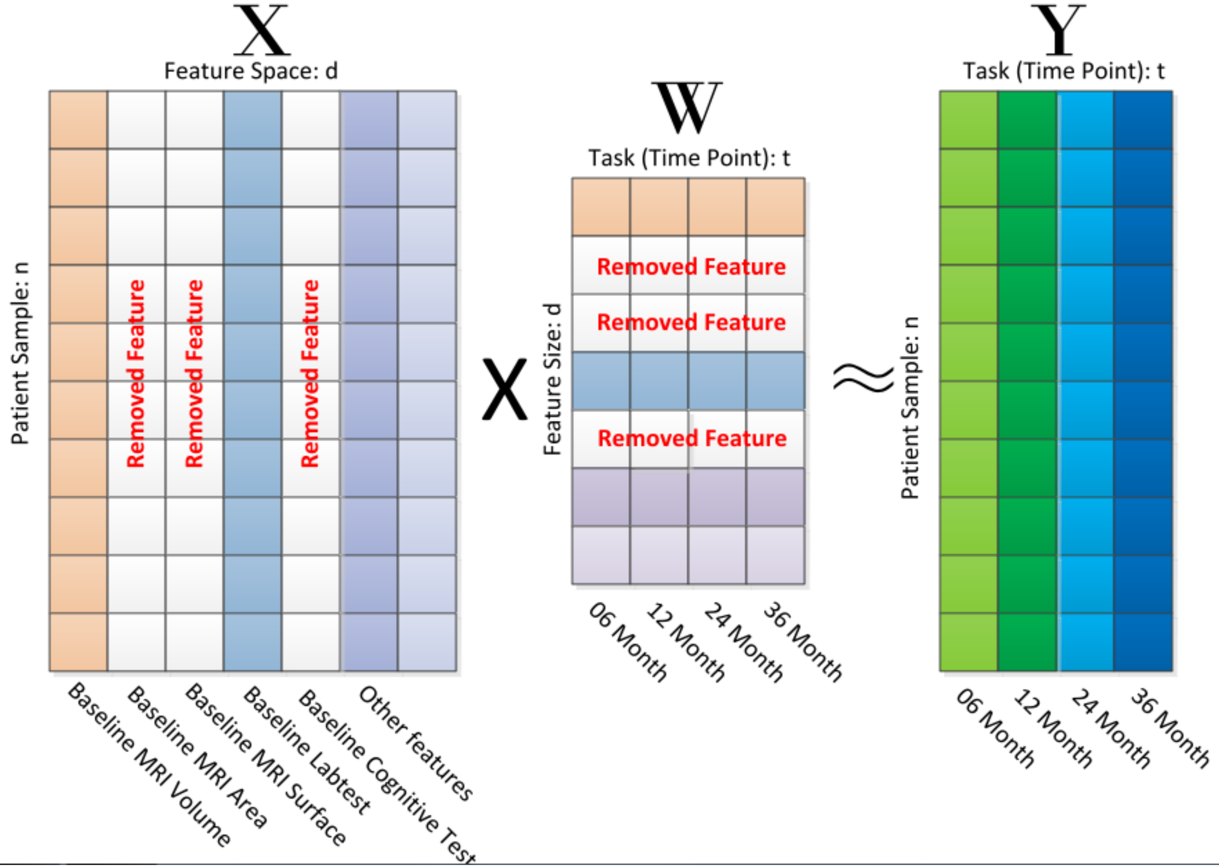


$$\min_{W} \frac{1}{2} \|XW - Y\|_F^2 + \lambda \|W\|_{1,q}$$

Zhou,Chen,Ye (2012)

Joint Feature Selection in Disease Progression

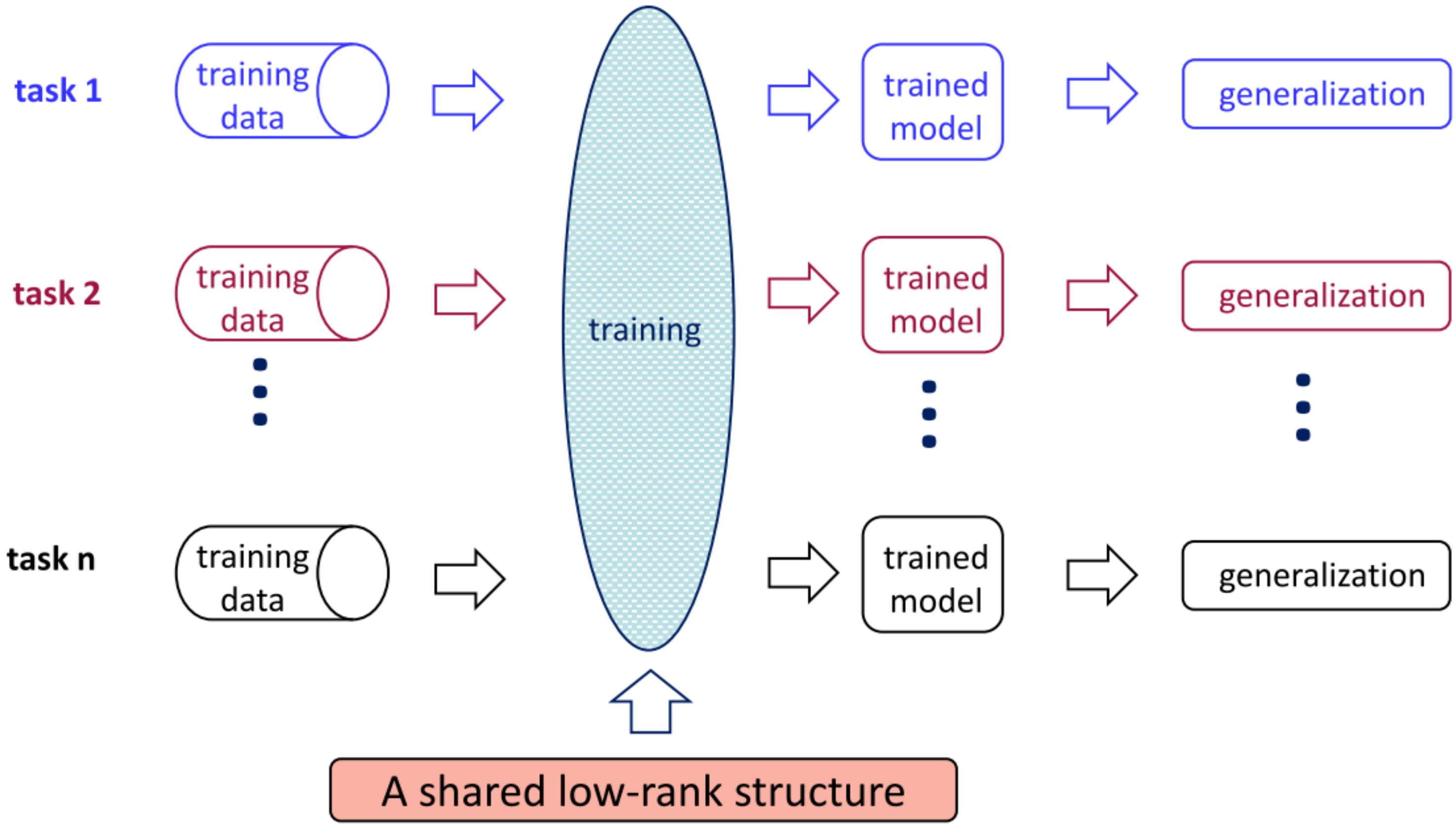
- The progression of disease is assumed to involve the same set of features at different time points [Zhou et.al. KDD 11].



Zhou,Chen,Ye (2012)

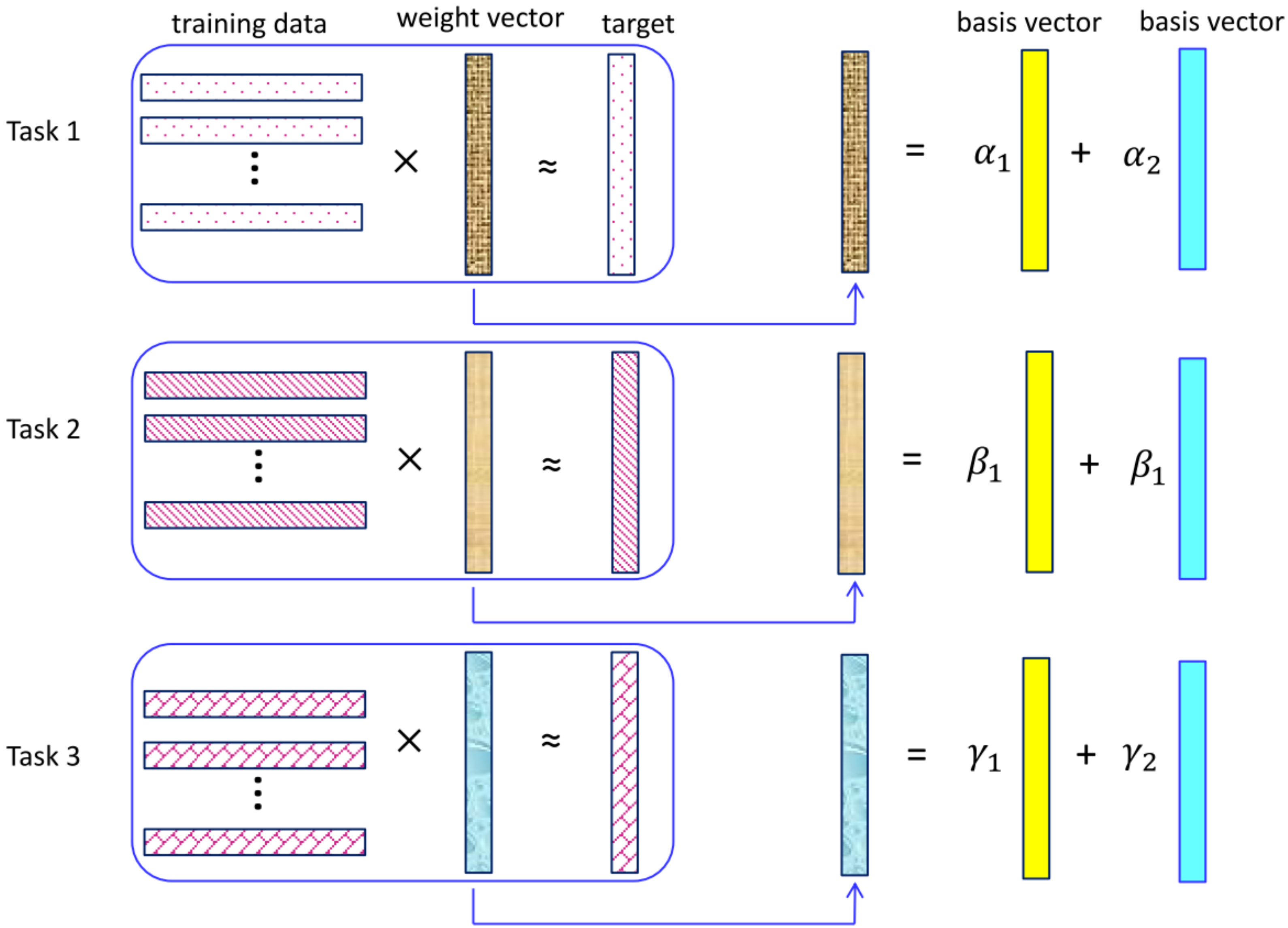
Trace-Norm Regularized MTL

- Capture Task Relatedness via a Shared Low-Rank Structure



Zhou,Chen,Ye (2012)

Low-Rank Structure for MTL



Zhou,Chen,Ye (2012)

Low-Rank Structure for MTL

Ji *et. al.* 2009 ICML

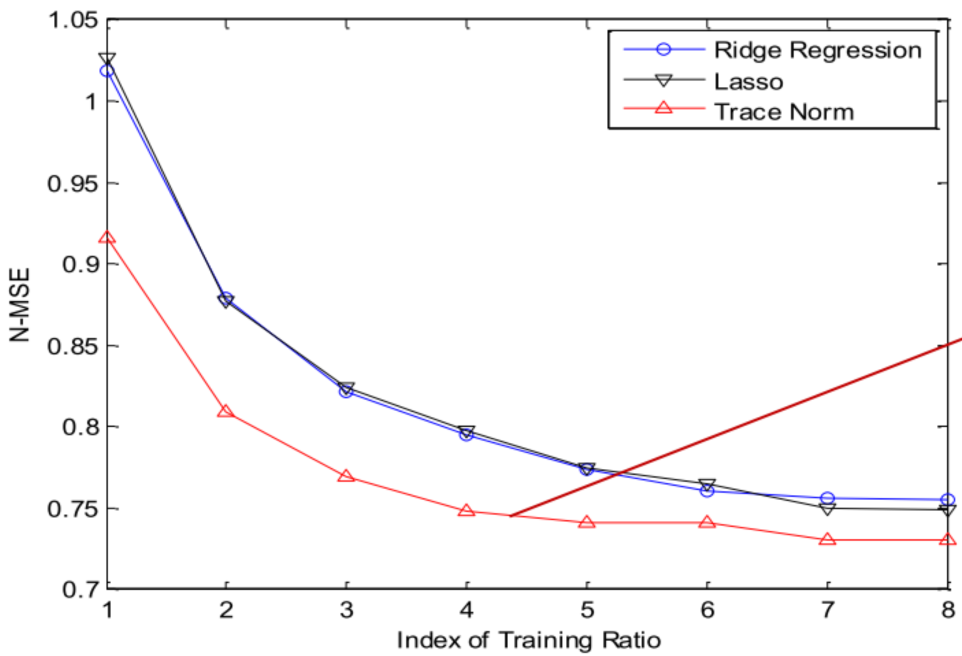
- Rank minimization formulation
  - $\min_{W} \mathrm{Loss}(W) + \lambda \times \mathrm{Rank}(W)$
  - Rank minimization is *NP-Hard* for general loss functions

- Convex relaxation: trace norm minimization
  - $\min_{W} \mathrm{Loss}(W) + \lambda \times \|W\|_*$ $\|W\|_*$: sum of singular values of W
  - The trace norm is theoretically shown to be a good approximation for rank function (Fazel et al., 2001).

Zhou,Chen,Ye (2012)

Low-Rank Structure for MTL

- Evaluation on the *School* data[1]:
  - Predict exam scores for 15362 students from 139 schools
  - Describe each student by 27 attributes
  - Compare Ridge Regression, Lasso, and Trace Norm (for inducing a low-rank structure)



Performance measure:

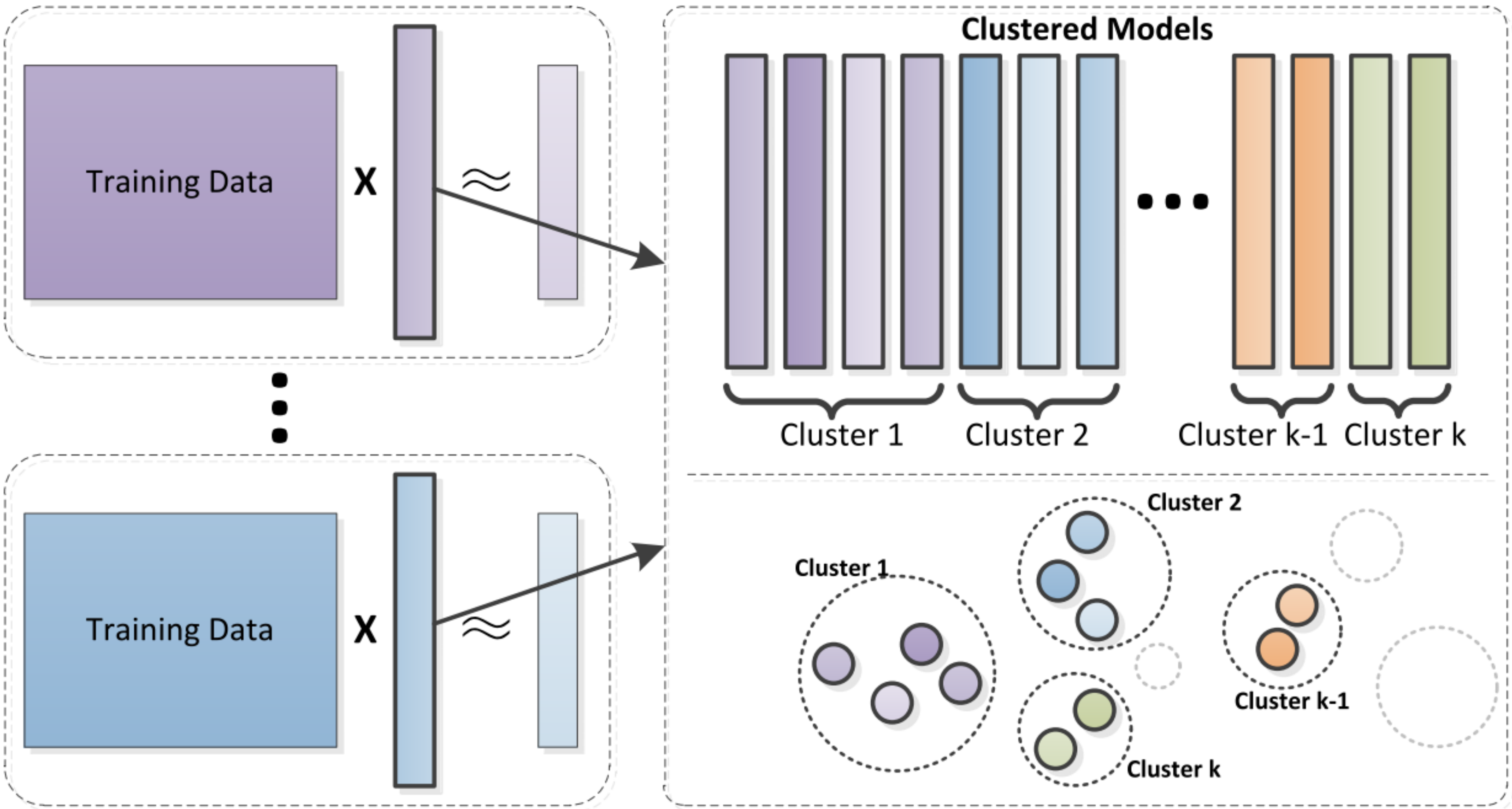
$$\text{N-MSE} = \frac{\text{mean squared error}}{\text{variance (target)}}$$

The Low-Rank Structure (induced via Trace Norm) leads to the smallest N-MSE.

Zhou, Chen, Ye (2012)

Clustered Multi-Task Learning

- Use regularization to capture clustered structures.



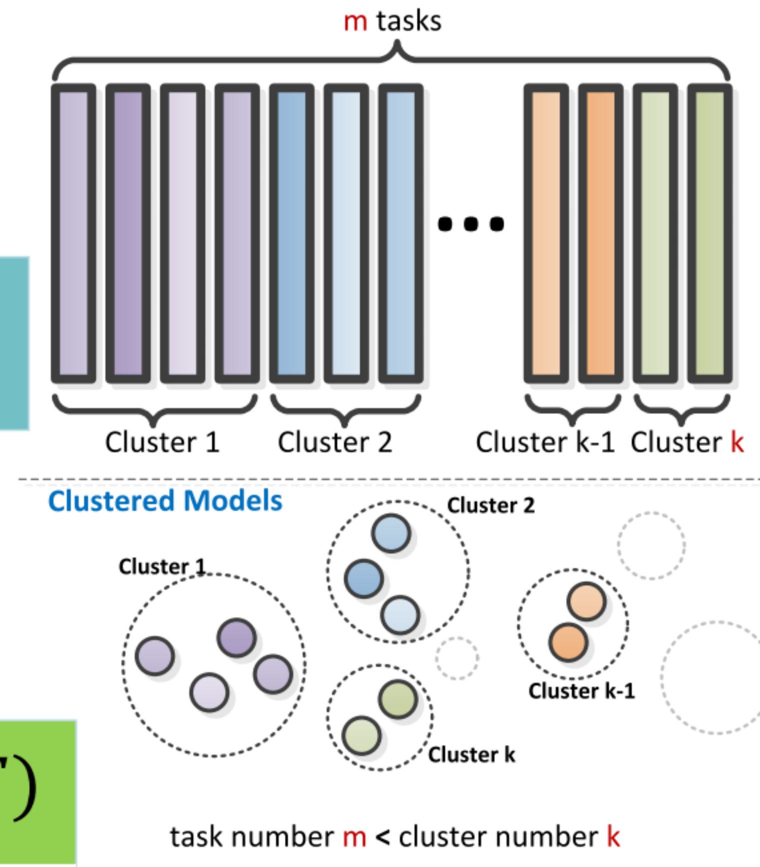
Zhou, Chen, Ye (2012)

Clustered Multi-Task Learning

- Capture structures by minimizing sum-of-square error (SSE) in K-means clustering:

$$\min_{I} \sum_{j=1}^{k} \sum_{v \in I_j} \left\| w_v - \bar{w}_j \right\|_2^2$$

I_j index set of j^{th} cluster



Equivalent

$$\min_{F} \operatorname{tr}(W^T W) - \operatorname{tr}(F^T W^T W F)$$

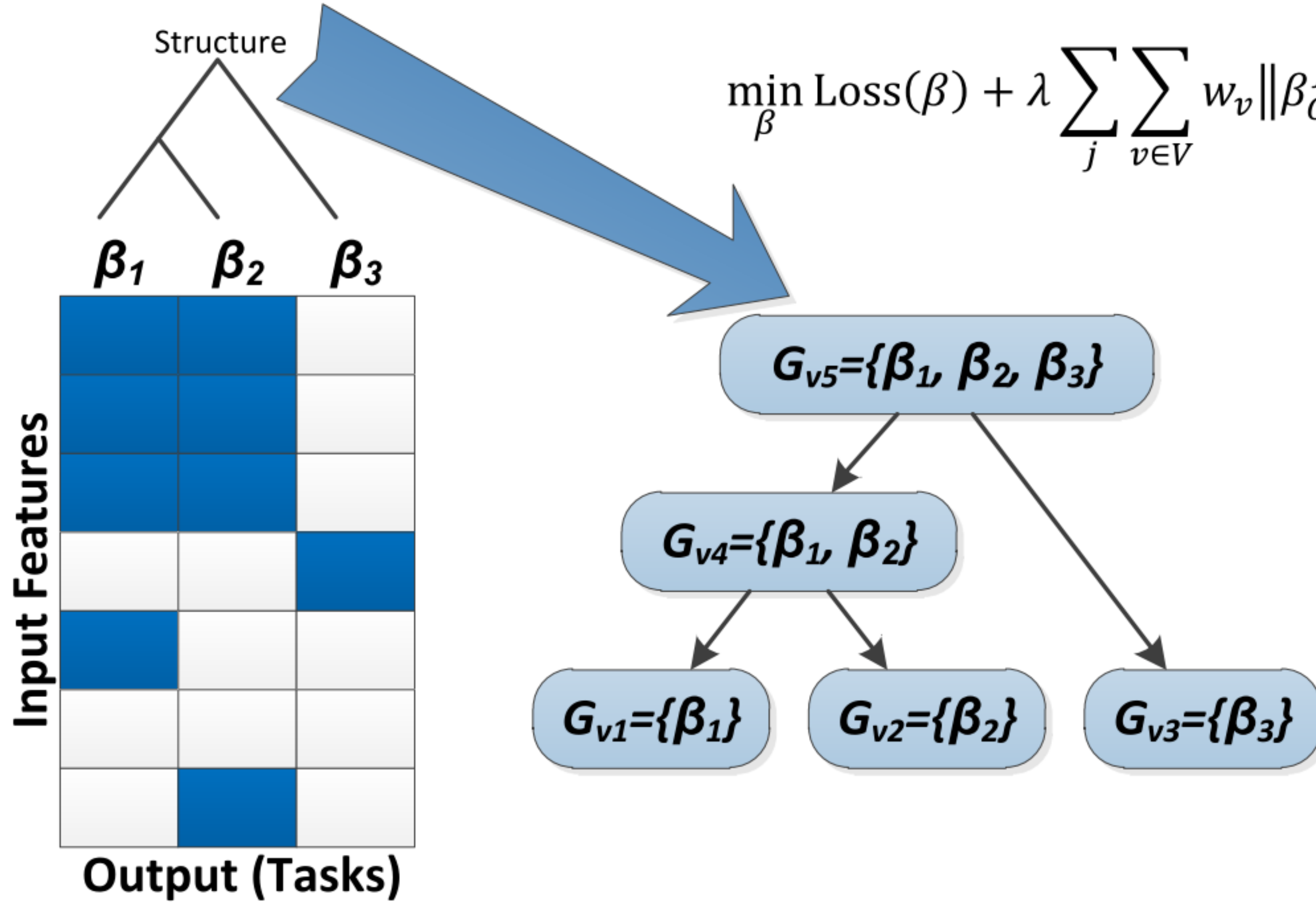
F : $m \times k$ orthogonal cluster indicator matrix
$F_{i,j} = 1/\sqrt{n_j}$ if $i \in I_j$ and 0 otherwise

Zhou, Chen, Ye (2012)

Multi-Task Learning with Tree Structures

- Tree-Guided Group Lasso (Kim and Xing 2010 ICML)

$$\min_{\beta} \mathrm{Loss}(\beta) + \lambda \sum_{j} \sum_{v \in V} w_v \left\| \beta^{j}_{G_v} \right\|_2$$



Zhou, Chen, Ye (2012)

Multi-Task Learning with Graph Structures

- A simple way to encode graph structure is to penalize the difference of two tasks that have an edge between them
- Given a set of edges E, we thus penalize:

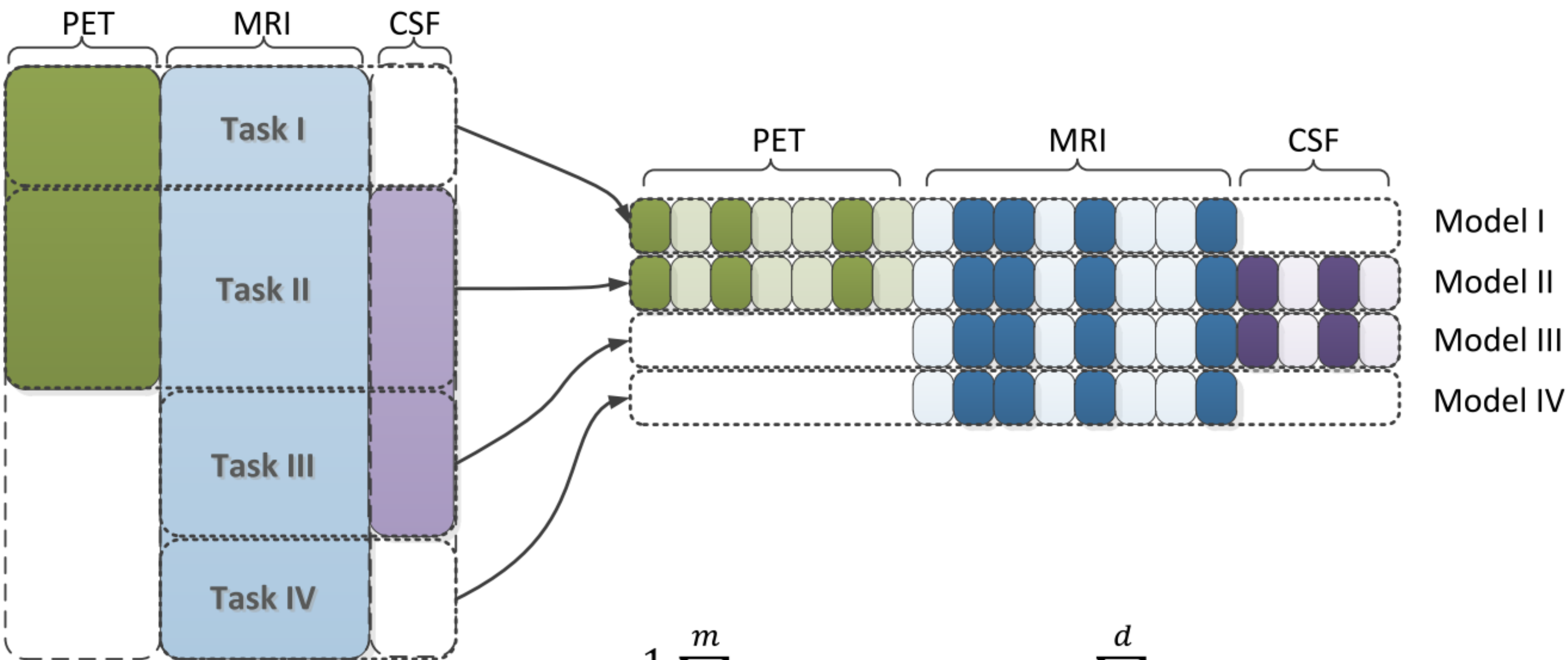
$$\sum_{i=1}^{|E|} \left\| W_{e_{\{i,1\}}} - W_{e_{\{i,2\}}} \right\|_2^2 = \|WR^T\|_F^2 \quad R \in \mathbb{R}^{|E| \times m}$$

- The graph regularization term can also be represented in the form of Laplacian term

$$\|WR^T\|_F^2 = tr\left((WR^T)^T WR^T\right) = tr(WR^T RW^T) = tr(W\mathcal{L}W^T)$$

Zhou, Chen, Ye (2012)

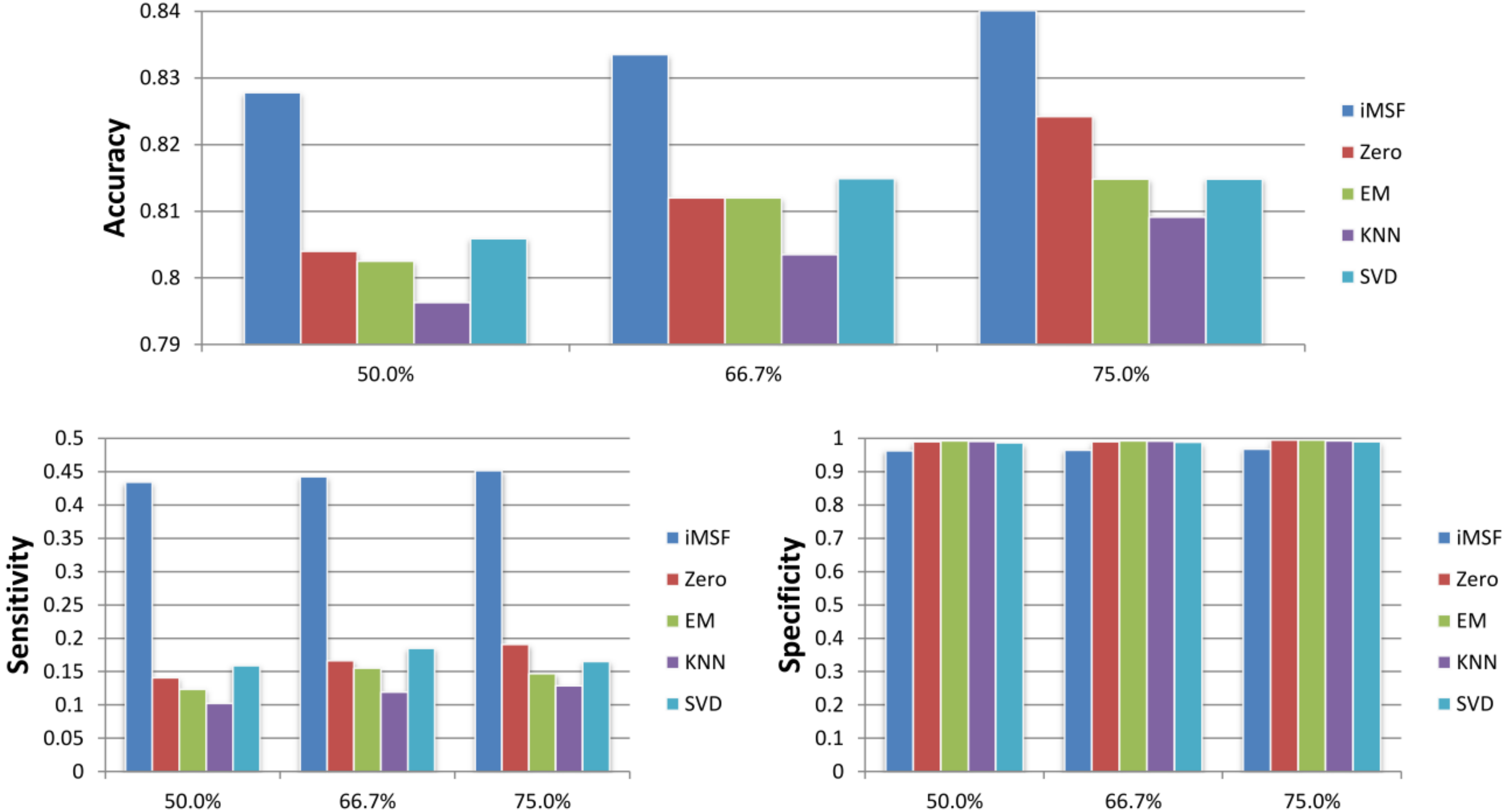
Yuan *et. al.* 2012 NeuroImage



$$\min_{W} \frac{1}{m} \sum_{i=1}^{m} Loss(X_i, Y_i, W_i) + \lambda \sum_{k=1}^{d} \|W_{G_k}\|_2$$

Zhou, Chen, Ye (2012)

Yuan *et. al.* 2012 NeuroImage



Zhou,Chen,Ye (2012)

MALSAR

MULTI-TASK LEARNING VIA STRUCTURAL REGULARIZATION

JIAYU ZHOU, JIANHUI CHEN, JIEPING YE

- A multi-task learning package
- Encode task relationship via structural regularization
- www.public.asu.edu/~jye02/Software/MALSAR/

Zhou,Chen,Ye (2012)

MTL Algorithms in MALSAR 1.0

- Mean-Regularized Multi-Task Learning
- MTL with Embedded Feature Selection
  - Joint Feature Learning
  - Dirty Multi-Task Learning
  - Robust Multi-Task Feature Learning
- MTL with Low-Rank Subspace Learning
  - Trace Norm Regularized Learning
  - Alternating Structure Optimization
  - Incoherent Sparse and Low Rank Learning
  - Robust Low-Rank Multi-Task Learning
- Clustered Multi-Task Learning
- Graph Regularized Multi-Task Learning

Zhou, Chen, Ye (2012)

Other Multi-task

- NN hidden layers
- Gaussian processes shared kernel parameters

Q&A

Thank you!