Regularized Multi-task Learning

by

T.E.M.P.

Setup and Notation

T learning tasks, m data points per task:

$$\{(\mathbf{x}_{1t}, y_{1t}), \mathbf{x}_{2t}, y_{2t}) \ldots, (\mathbf{x}_{mt}, y_{mt})\}$$

sampled from P_t on $X \times Y$. P_t's are related.

Goal: Learn T functions $f_1, f_2, \ldots, f_T$ such that $f_t(\mathbf{x}_{it}) \approx y_{it}$.

($T = 1$ is the standard (single–task) learning problem.)

First assume: $f_t(\mathbf{x}) = \mathbf{w}_t \cdot \mathbf{x}$

Some Examples

- Same “y’s”, different “$\mathbf{x}$’s”: integrating information from heterogeneous databases (Ben-David et al 2002)

- Same “$\mathbf{x}$’s”, different “y’s”: “function heterogeneity”, multi-class classification

- $(\mathbf{x}, y)$ belong to different $X_t \times Y_t$: learning–by–components, general multi–task learning

Hierarchical Bayesian Methods

Assume $\mathbf{w}_t$ are samples from a Gaussian with mean $\mathbf{w}_0$ and covariance Σ.

Use some (Gibbs sampling) iterative approach to estimate *simultaneously*:

$$\{\mathbf{w}_0, \quad \Sigma, \quad \mathbf{w}_t\}$$

A simple idea

Assume $\mathbf{w}_t = \mathbf{w}_0 + \mathbf{v}_t$, with $\mathbf{v}_t$ "small". Solve:

$$\min_{\mathbf{w}_0, \mathbf{v}_t, \xi_{it}} \sum_{t=1}^{T} \sum_{i=1}^{m} \xi_{it} + \frac{\lambda_1}{T} \sum_{t=1}^{T} \|\mathbf{v}_t\|^2 + \lambda_2 \|\mathbf{w}_0\|^2$$

$$y_{it}(\mathbf{w}_0 + \mathbf{v}_t) \cdot \mathbf{x}_{it} \geq 1 - \xi_{it}$$

$$\xi_{it} \geq 0$$

for $\forall i \in \{1, 2, \ldots, m\}$ and $t \in \{1, 2, \ldots, T\}$

Optimal Solution is an Average

The optimal solution of the multi–task optimization method satisfies the equation

$$\mathbf{w}_0^* = \frac{\lambda_1}{\lambda_2 + \lambda_1} \frac{1}{T} \sum_{t=1}^{T} \mathbf{w}_t^*$$

That is, $\mathbf{w}_0^*$ is the average of the individual task models $\mathbf{w}_t^*$.

Equivalent Optimization Problem

$$\min_{\mathbf{w}_t,\xi_{it}} \left\{ \sum_{t=1}^{T}\sum_{i=1}^{m} \xi_{it} + \rho_1 \sum_{t=1}^{T} \|\mathbf{w_t}\|^2 + \rho_2 \sum_{t=1}^{T} \|\mathbf{w}_t - \frac{1}{T}\sum_{s=1}^{T} \mathbf{w}_s\|^2 \right\}$$

$$y_{it}\mathbf{w}_t \cdot \mathbf{x}_{it} \geq 1 - \xi_{it}$$

$$\xi_{it} \geq 0$$

for $\forall i \in \{1, 2, \ldots, m\}, t \in \{1, 2, \ldots, T\}$

For appropriate ρ_1, ρ_2.

Also Equivalent

For $\mu = \frac{T\lambda_2}{\lambda_1}$, define the feature map:

$$\Phi((\mathbf{x}, t)) = (\frac{\mathbf{x}}{\sqrt{\mu}}, \underbrace{\mathbf{0}, \ldots, \mathbf{0}}_{t-1}, \mathbf{x}, \underbrace{\mathbf{0}, \ldots, \mathbf{0}}_{T-t})$$

Then we are solving a single–task problem of estimating:

$$\mathbf{w} = (\sqrt{\mu}\mathbf{w}_0, \mathbf{w}_1, \ldots, \mathbf{w}_T).$$

By construction we have that $\mathbf{w} \cdot \Phi((\mathbf{x}, t)) = (\mathbf{w}_0 + \mathbf{w}_t) \cdot \mathbf{x}$ and

$$\|\mathbf{w}\|^2 = \sum_{t=1}^{T} \|\mathbf{w}_t\|^2 + \mu \|\mathbf{w}_0\|^2.$$

Dual Formulation

Let $C := \frac{T}{2\lambda_1}$, $\mu = \frac{T\lambda_2}{\lambda_1}$. Define the kernel:

$$K_{st}(\mathbf{x}, \mathbf{z}) := \left(\frac{1}{\mu} + \delta_{st}\right) \mathbf{x} \cdot \mathbf{z}, \quad s, t = 1, \ldots, T.$$

The dual problem is:

$$\max_{\alpha_{it}} \left\{ \sum_{i=1}^{m} \sum_{t=1}^{T} \alpha_{it} - \frac{1}{2} \sum_{i=1}^{m} \sum_{s=1}^{T} \sum_{j=1}^{m} \sum_{t=1}^{T} \alpha_{is} y_{is} \alpha_{jt} y_{jt} K_{st}(\mathbf{x}_{is}, \mathbf{x}_{jt}) \right\}$$

$0 \leq \alpha_{it} \leq C$ for $\forall i \in \{1, 2, \ldots, m\}, t \in \{1, 2, \ldots, T\}$

$\rightarrow$ A single–task SVM with a kernel parameterized by μ (the "task-relatedness" parameter).

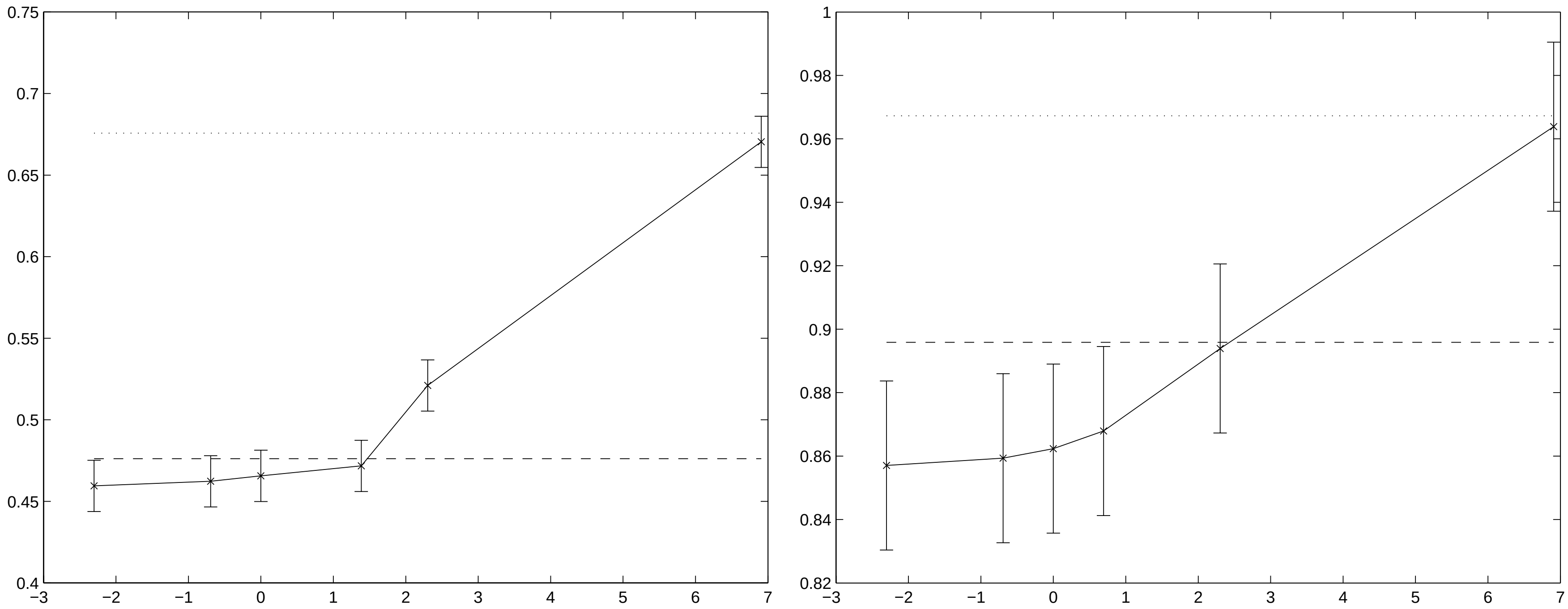
Modeling consumer heterogeneity: Few tasks

There are 30 tasks (individuals). RMSE and hit errors reported.

Noise	Similar	HB	$\mu = 0.1$	SVM
H	L	0.85 26.14	0.81* 25.86*	0.84 26.22
H	H	0.90 31.03	**0.86** **30.58**	0.97 31.60
L	L	0.60 14.34	0.58* 14.12*	0.65 16.00
L	H	0.48 13.42	0.46* 13.19*	0.68 17.11

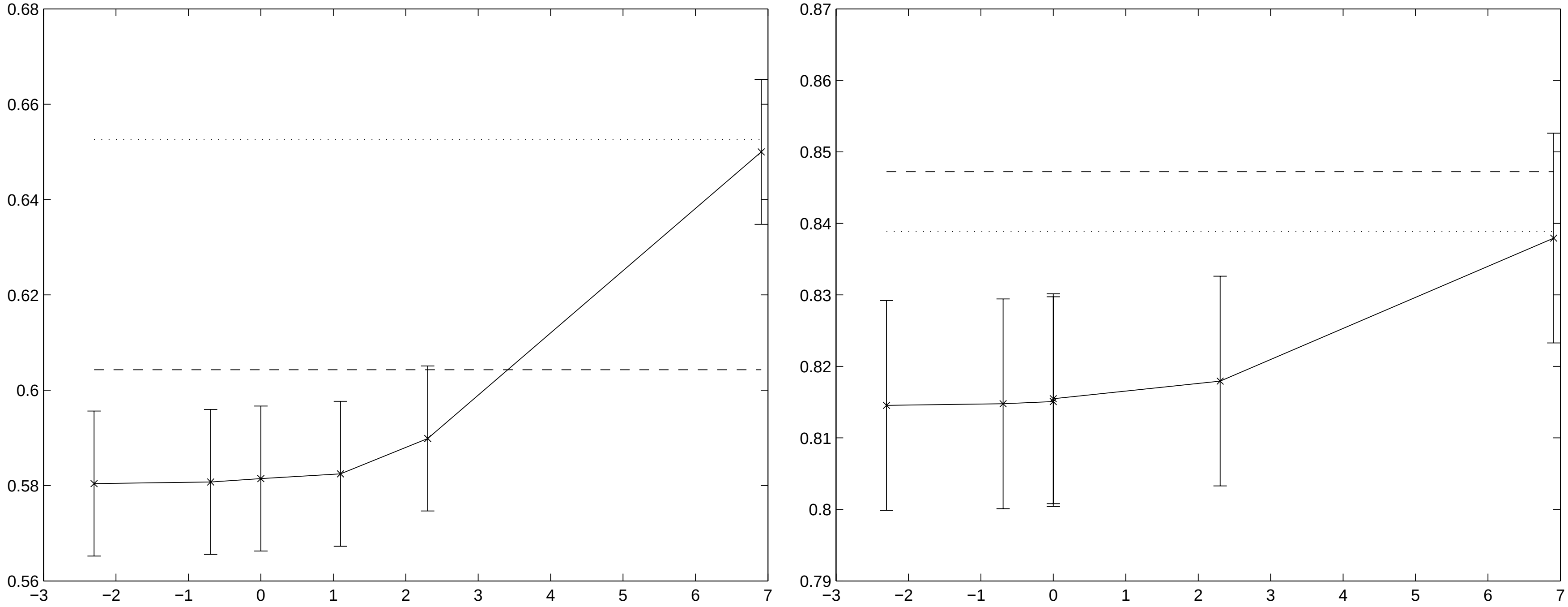
High simlarity tasks (individuals)

Left: Low nose; Right: High noise



Low similarity tasks (individuals)

Left: Low nose; Right: High noise



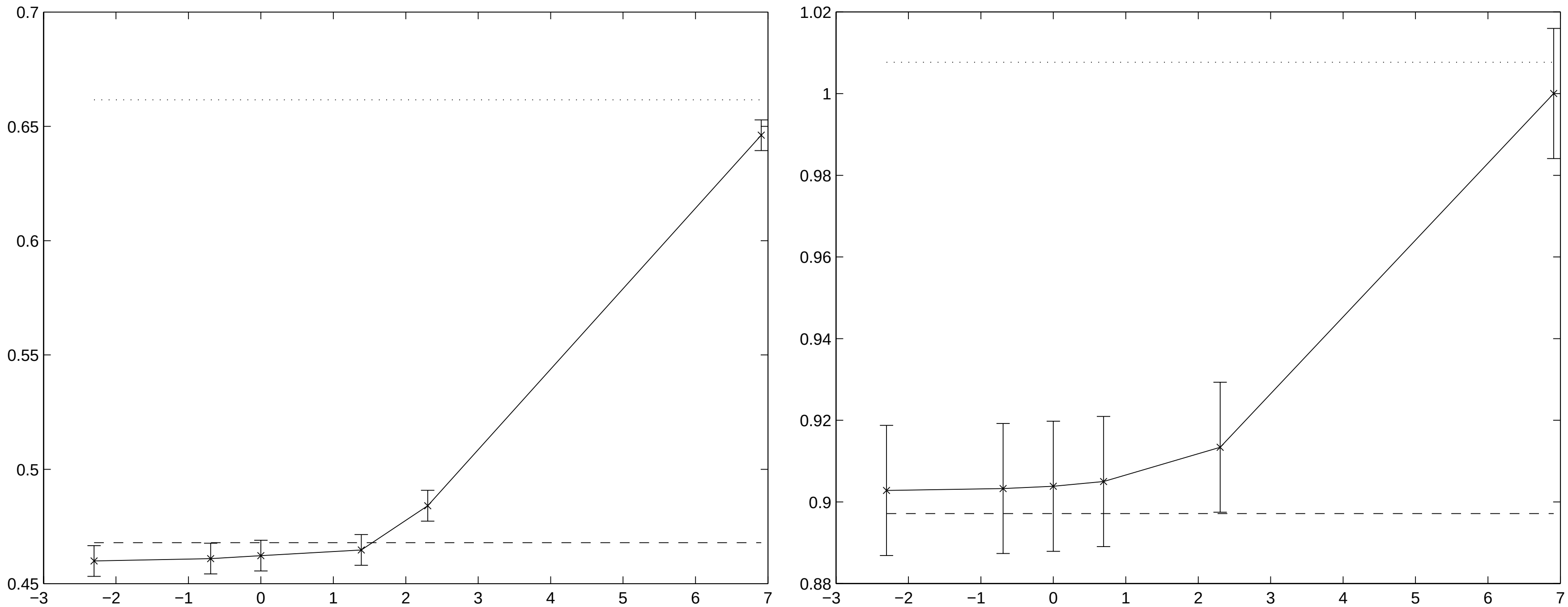
Modeling consumer heterogeneity: Many tasks

There are 100 tasks (individuals). RMSE and hit errors reported.

Noise	Similar	HB	$\mu = 0.1$	SVM
H	L	0.81 24.65	**0.79** **24.24**	0.82 24.98
H	H	**0.90** **31.49**	**0.90** **31.48**	1.01 33.13
L	L	**0.59** **13.97**	**0.58** **14.02**	0.66 15.57
L	H	**0.47** **13.05**	**0.46** **13.28**	0.66 16.98

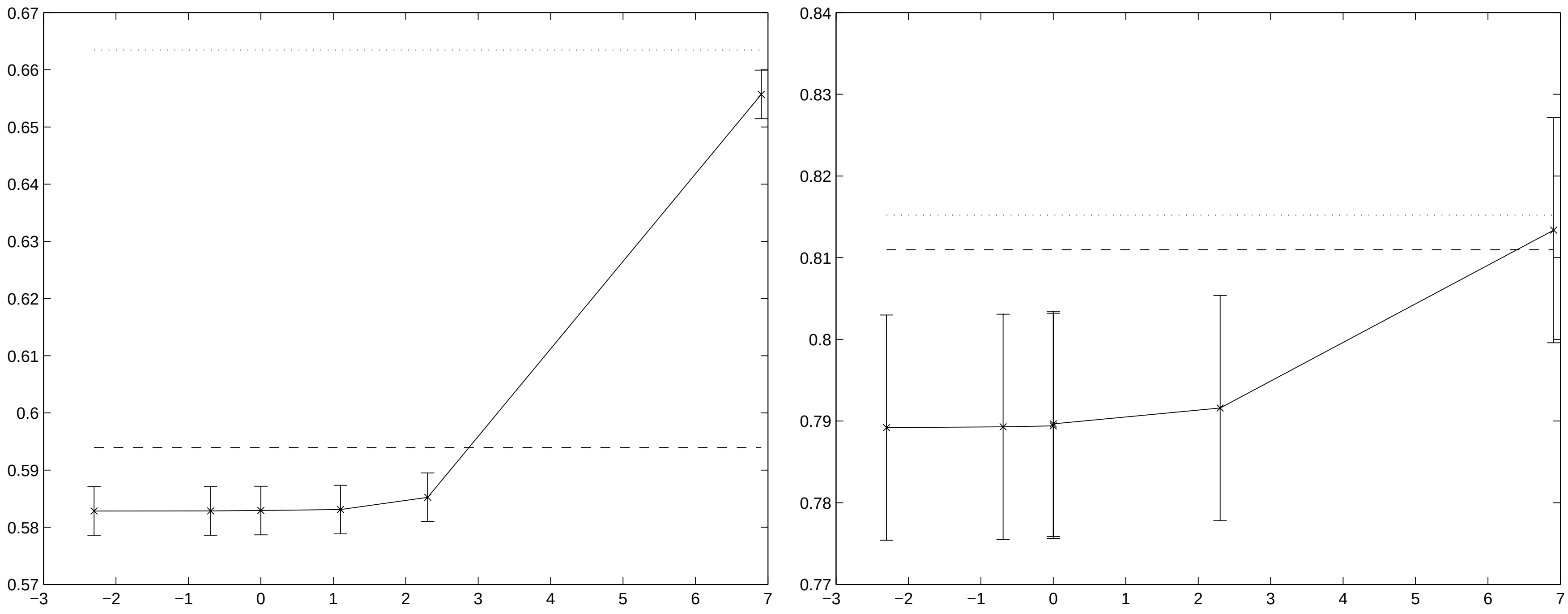
High similarity tasks (individuals)

Left: Low nose; Right: High noise



Low similarity tasks (individuals)

Left: Low nose; Right: High noise



Multi-company Information Integration

School Data: first column is with $C = 0.1$ and second with $C = 1$. Bayesian stands for the task clustering method of (Bakker and Heskes 2003)

$\mu = 0.5$	34.30 ± 0.3	34.37 ± 0.4
$\mu = 1$	34.28 ± 0.4	34.37 ± 0.3
$\mu = 2$	34.26 ± 0.4	34.11 ± 0.4
$\mu = 10$	34.32 ± 0.3	29.71 ± 0.4
$\mu = 1000$	11.92 ± 0.5	4.83 ± 0.4
Bayesian	29.5 ± 0.4	29.5 ± 0.4

A simple generalization

Assume

$$f_t = g + g_t$$

Then the kernel becomes:

$$K_{st}(\mathbf{x},\mathbf{z}) := \frac{1}{\mu} K_1(\mathbf{x},\mathbf{z}) + \delta_{st} K_2(\mathbf{x},\mathbf{z}), \quad s,t = 1,\ldots,T.$$

Other directions

- Kernels can be defined so that tasks are clustered (Bakker and Heskes 2003): use $\mathbf{w}_{01}, \mathbf{w}_{02}, \ldots \mathbf{w}_{0K}$ for K clusters.

- Consider many tasks that share similar features: learn common features among tasks by defining the kernel matrix (Baxter 2000).

- Assume

$$f_t = g^{(0)} + g_t^{(1)} + g_t^{(2)} + \ldots$$

where the higher index i of $g^{(i)}$ is, the higher the “resolution” we use to learn the tasks.

Concluding remarks

- Multi-task approach can lead to significant improvements (when tasks are related (?))

- Many possible directions for future theoretical research: rigorous definition of task relatedness, common features across tasks, multi–resolution multi–task learning, task clustering, single-task theory extensions, etc

- Many applications: integration of information sources, learning-by-components, multi–modal learning, etc