Multi-task Learning: Theory and Practice

Massimiliano Pontil

Department of Computer Science
Centre for Computational Statistics and Machine Learning
University College London

Chinese Conference on Pattern Recognition
Beijing, 25/09/12

Outline

- Problem formulation
- Examples
- Different regularizers: quadratic, structured sparsity, spectral
- Statistical analysis of structure sparsity
- Optimization methods
- Numerical experiments
- OrthoMTL and sparse coding

Problem formulation

- Let $\mu_1, \ldots, \mu_T$ be prescribed probability distributions on $X \times Y$
- $(x_{t1}, y_{t1}), \ldots, (x_{tn}, y_{tn}) \sim \mu_t,\ t = 1, \ldots, T$
- Goal: find functions $f_t : X \to Y$ which minimize

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{(x,y) \sim \mu_t} \ell(f_t(x), y)$$

- Regularization approach:

$$\min \ \frac{1}{T} \sum_{t=1}^{T} \frac{1}{n} \sum_{i=1}^{n} \ell(f_t(x_{ti}), y_{ti}) + \lambda \ \Omega(f_1, \ldots, f_T)$$

- The penalty term "encourages" common structure among the tasks / uses prior knowledge that $f_1, \ldots, f_T$ are related

Problem formulation (cont.)

- Focus on linear regression and square loss: $X \subseteq R^d$, $Y \subseteq R$, $y_{ti} = w_t^\top x_{ti} + \epsilon_{ti}$

$$\min \ \frac{1}{T} \sum_{t=1}^{T} \underbrace{\frac{1}{n} \sum_{i=1}^{n} (y_{ti} - w_t^\top x_{ti})^2}_{\text{training error task t}} + \lambda \underbrace{\Omega(w_1, \ldots, w_T)}_{\text{joint regularizer}}$$

- Typical scenario: many tasks but only *few examples* per task
- If the tasks are "related", learning them **jointly** should perform better than learning each task *independently*

Example 1: user modeling

- Each task is to predict a user's ratings to products

CPU	**CD**	**RAM**	$\cdots$	**HD**	**Screen**	**Price**	**Rating**
1GHz	Y	1GB	$\cdots$	40G	15in	$1000	7
1GHz	N	1.5GB	$\cdots$	20G	13in	$1200	3
1.5GHz	Y	1.5GB	$\cdots$	40G	17in	$1700	5
2GHz	Y	2GB	$\cdots$	80G	15in	$2000	**?**
1.5GHz	N	2GB	$\cdots$	40G	13in	$1800	**?**

- The ways different people make decisions about products are related. *How do we exploit this?*

Example 2: object detection

- Multiple object detection in scenes: detection of each object corresponds to a binary classification task

- Learning common visual features enhances performance

Early work in ML used a hidden layer neural nets with hidden weights shared by all the tasks [Baxter 96, Caruana 97, Silver and Mercer 96, etc.]

Objective and questions

- High dimensional setting!
- What is the multi-task counterpart of smoothness / sparsity assumptions used in single-task learning?
- Statistical estimation
- Optimization techniques

Penalty function

$$\min \frac{1}{T}\sum_{t=1}^{T}\frac{1}{n}\sum_{i=1}^{n}(y_{ti} - w_t^\top x_{ti})^2 + \lambda\ \Omega(w_1, \ldots, w_T)$$

1. **Quadratic:** encourages closeness of task parameters, or other linear relationships

Quadratic regularizer [Evgeniou et al. 05]

$$\min \frac{1}{T}\sum_{t=1}^{T}\frac{1}{n}\sum_{i=1}^{n}(y_{ti}-w_t^\top x_{ti})^2+\lambda\,\Omega(w_1,\dots,w_T)$$

- Let $\Omega(w)=w^\top E w$, with $w\in R^{dT}$ the concatenation of $w_1,\dots,w_T$
- $E\in R^{dT\times dT}$, *symmetric positive definite*, models tasks relationships
- If E is block diagonal the tasks are learned *independently*
- Example [Evgeniou and P., 04]: stay close to the average

$$\Omega(w)=\sum_{t=1}^{T}\|w_t\|^2+\frac{1-\gamma}{\gamma}\sum_{t=1}^{T}\|w_t-\frac{1}{T}\sum_{s=1}^{T}w_s\|^2$$

$\gamma\in[0,1]$, $\gamma=1$: independent tasks, $\gamma=0$: identical tasks

Feature space point of view

- Equivalent to learn a single function on larger domain: $(x, t) \mapsto f_t(x)$
- Linear case: $f_t(x) = v^\top B_t x$, for some $v \in \mathbb{R}^p$ $(p \geq dT)$ and $B_t \in \mathbb{R}^{p \times d}$ matrices (task specific)
- The learning problem can be rewritten as:

$$S(w) = \sum_{t=1}^{T} \sum_{i=1}^{n} (y_{ti} - v^\top B_t x_{ti})^2 + \lambda v^\top v$$

- Linear multitask kernel: $K((x, t), (x', t')) = x B_t^\top B_{t'} x'$
- Can use kernel techniques (representer theorem, dual problem, etc.)

Equivalent prolems

$$R(w) = \sum_{t=1}^{T}\sum_{i=1}^{n}(y_{ti} - w_t^\top x_{ti})^2 + \lambda w^\top E w$$

$$S(v) = \sum_{t=1}^{T}\sum_{i=1}^{n}(y_{ti} - v^\top B_t x_{ti})^2 + \lambda v^\top v$$

Proposition. The problems are equivalent:

- Given $B := [B_1, \ldots, B_T]$ full rank (dT) then set $E = (B^\top B)^{-1}$
- Given E, let A be a square root of E and set $B = A^\top E^{-1}$

Example (revisited)

- We choose

$$B_t^\top = [(1-\gamma)^{\frac{1}{2}} \mathbf{I}_{d\times d}, \underbrace{\mathbf{0}_{d\times d}, \dots, \mathbf{0}_{d\times d}}_{t-1}, (\gamma T)^{\frac{1}{2}} \mathbf{I}_{d\times d}, \underbrace{\mathbf{0}_{d\times d}, \dots, \mathbf{0}_{d\times d}}_{T-t}]$$

- Interpretation

$$w_t = B_t^\top v = \sqrt{1-\gamma} v_0 + \sqrt{\gamma T} v_t = \text{``}\textit{common}\text{''} + \text{``}\textit{task specific}\text{''}$$

- $B_t^\top B_{t'} = (1-\gamma)\mathbf{I}_{d\times d} + \gamma T \delta_{tt'} \mathbf{I}_{d\times d}$. Computing $(B^\top B)^{-1}$ we confirm that

$$w^\top E w = \frac{1}{T}\left(\sum_{t=1}^T ||w_t||_2^2 + \frac{1-\gamma}{\gamma}\sum_{t=1}^T ||w_t - \frac{1}{T}\sum_{t'=1}^T w_{t'}||_2^2\right)$$

Penalty function

Define

$$W = \begin{pmatrix} | & & | \\ w_1 & \dots & w_T \\ | & & | \end{pmatrix} = \begin{pmatrix} - \, w^1 \, - \\ \vdots \\ - \, w^d \, - \end{pmatrix}$$

Consider

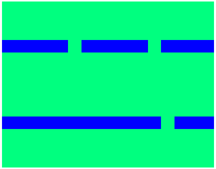
$$\min_W \frac{1}{T} \sum_{t=1}^{T} \frac{1}{n} \sum_{i=1}^{n} (y_{ti} - w_t^\top x_{ti})^2 + \lambda \, \Omega(W)$$

1. Quadratic: encodes closeness of task parameters
2. **Structured sparsity:** few common variables

2. Structured Sparsity

- Favour matrices with many zero rows (few variables shared by the tasks)

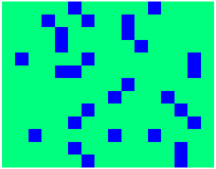
$$\Omega_{\rm s}(W) = \sum_{j=1}^{d} ||w^j||_2 = \sum_{j=1}^{d} \sqrt{\sum_{t=1}^{T} w_{tj}^2}$$

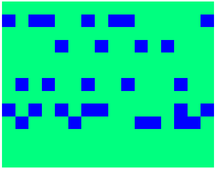


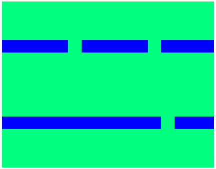
- Special case of group Lasso [Lounici et al. 09, Yuan and Lin, 06]

2. Structured Sparsity (cont.)

Compare matrices W favoured by different norms $(\mathrm{green} = 0,\ \mathrm{blue} = 1)$:







$\#\text{rows} =$	13	5	2
$\Omega_s =$	19	12	8
$\sum_{tj} \lvert w_{tj} \rvert =$	29	29	29

Statistical analysis of structured sparsity

- Linear regression model: $y_{ti} = w_t^\top x_{ti} + \epsilon_{ti}$, $i = 1, \dots, n$, $d \gg n$
- Noise: ϵ_{ti} are i.i.d. $N(0, \sigma^2)$
- Sparsity pattern $J(W) := \left\{ j : \sum_{t=1}^{T} w_{tj}^2 > 0 \right\}$. Assume $|J(W)| \leqslant s$
- Variable not too correlated: $\frac{1}{n} \left| \sum_{i=1}^{n} (x_{ti})_j (x_{ti})_k \right| \leq \frac{1-\rho}{7s}$, $\forall t$, $\forall j \neq k$

Q1 (estimation) $\frac{1}{T} \sum_{t=1}^{T} \|\hat{w}_t - w_t\|^2 \leqslant$?

Q2 (variable selection) $\mathrm{Prob}\left\{ J(\hat{W}) = J(W) \right\} \approx 1$?

Estimation error bound

Theorem [Lounici et al. 2011] If $\lambda = \frac{4\sigma}{\sqrt{nT}}\sqrt{1 + A\frac{\log d}{T}}$, $A \geq 4$ then w.h.p.

$$\frac{1}{T}\sum_{t=1}^{T} \|\hat{w}_t - w_t\|^2 \quad \leq \quad \left(\frac{c\sigma}{\rho}\right)^2 \frac{s}{n}\sqrt{1 + A\frac{\log d}{T}}$$

- Dependency on the dimension d is *negligible* for large T
- Compare to Lasso: $\frac{1}{T}\sum_{t=1}^{T} \|w_t^{(L)} - w_t\|^2 \geq c' \frac{s}{n} \log(d\ T)$
- Similar results for prediction error and variable selection

Penalty Function

$$\min_{W} \sum_{t=1}^{T} \sum_{i=1}^{n} (y_{ti} - w_t^\top x_{ti})^2 + \lambda \, \Omega(W)$$

1. Quadratic: encodes closeness of task parameters
2. Structured sparsity: few common variables
3. **Spectral:** few common features

Spectral regularization

- Favour matrices with low rank: $\Omega(W) = \mathrm{rank}(W)$ (task vectors w_t lie on a *low dimensional* subspace)
- Recall the SVD of a matrix

$$W = U \ \mathrm{Diag}(\sigma_1, \ldots, \sigma_r) \ V^\top$$

  where $U \in R^{d \times r}$ and $V \in R^{T \times r}$ are orthogonal, $r = \min(d, T)$
- Approximate the rank with the trace norm [Fazel et al. 01]

$$\Omega_{\mathrm{tr}}(W) = \sum_{i=1}^{r} \sigma_i(W)$$

- More general: $\Omega(W) = \omega(\sigma_1, \ldots, \sigma_r)$, e.g. Schatten norms

Optimization methods

- Proximal gradient methods – require solving subproblem

$$\min_{W} \frac{1}{2}\|W - W_0\|^2 + \lambda\Omega(W)$$

  OK for $\ell_{2,1}$-norm, trace norm

- Using variational form:

$$\Omega(W) = \frac{1}{2} \inf_{D \in \mathcal{D}} \mathrm{trace}(D^{-1}WW^\top + D)$$

  where $\mathcal{D}$ is a subset of set of psd matrices [Argyriou et al. 08]

- Diagonal case [Micchelli, Morales, P., 2010]:
  $\mathcal{D} = \{\mathrm{diag}(\lambda_1, \ldots, \lambda_d) : \lambda \in \Lambda\}$, with $\Lambda \subseteq R^d_{++}$ a convex cone

Variational form [Argyriou et al. 08]

Express Ω as

$$\Omega_{\mathrm{tr}}(W) = \frac{1}{2} \min_{D \succ 0} \left\{ \mathrm{tr}(W^\top D^{-1} W) + \mathrm{tr}(D) \right\}$$

$$\min_{W,\ D \succ 0} \sum_{t=1}^{T} \sum_{i=1}^{n} (y_{ti} - w_t^\top x_{ti})^2 + \frac{\lambda}{2} \Big[\underbrace{\mathrm{tr}(W^\top D^{-1} W)}_{\sum_{t=1}^{T} w_t^\top D^{-1} w_t = w^\top E w} + \mathrm{tr}(D) \Big]$$

$$E = \begin{pmatrix} D^{-1} & 0 & \cdots & 0 \\ 0 & D^{-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & D^{-1} \end{pmatrix}$$

Jointly convex in (W, D) – related to problem of **learning the kernel**.
[Bach et al. 04, Micchelli and Pontil, 2005]

Optimization algorithm

- W-minimization: solve T independent regularization problems (e.g. SVM, ridge regression, etc.)
- D-minimization: can be solved analytically (via an SVD)

$$D(W) = \frac{(WW^\top)^{\frac{1}{2}}}{\mathrm{tr}(WW^\top)^{\frac{1}{2}}}$$

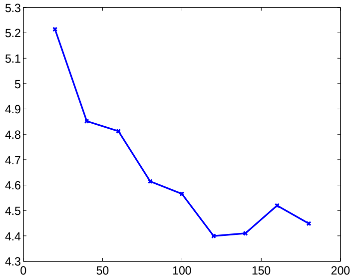
Theorem. By introducing a small perturbation

$$D(W) = \frac{(WW^\top + \varepsilon \mathbf{I})^{\frac{1}{2}}}{\mathrm{tr}(WW^\top + \varepsilon \mathbf{I})^{\frac{1}{2}}}$$

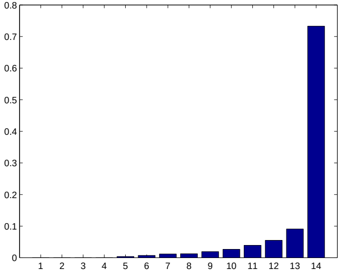
we can show that the algorithm converges to the optimal solution.

Experiment (Computer Survey) [Argyriou et al. 2008]

Test error vs. #tasks



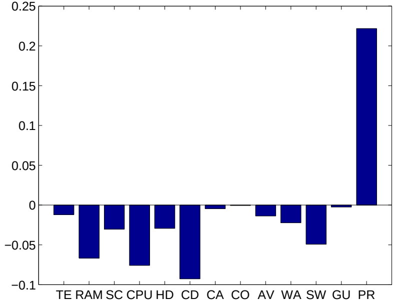
Eigenvalues of matrix D



- Performance improves with more tasks
- A single most important feature shared by everyone

Dataset: consumers' ratings of PC models: 180 persons (tasks), 8 training and 4 test examples. 13 binary inputs (RAM, CPU, price etc.). Integer output in $\{0, \ldots, 10\}$ (likelihood of purchase)

Experiment (Computer Survey)



Method	Test
Independent	15.05
Aggregate	5.52
Structured Sparsity	4.04
Trace norm	3.72
Quadratic + Trace	3.20

- The most important feature (eigenvector of D) weighs *technical characteristics* (RAM, CPU, CD-ROM) vs. *price*

Nonlinear MTL

Regularizers can be extended to nonlinear functions using reproducing kernel Hilbert spaces (RKHS)

- Quadratic: RKHS of vector-valued functions [Micchelli and P. 05, Evgeniou et al. 05, Caponnetto et al. 08]
- Sparsity: multiple kernel learning [Rakotomanonjy et al. 2011]
- Spectral: some technical issues of **function representation** arise [Argyriou, Micchelli, P, 2009]

More complex models and robustness

- Multitask clustering [Evgeniou et al. 2005, Jacob et al 2008]

- Composite regularizers: $\Omega(B \circ W)$, e.g. $\Omega([w_1 - \bar{w}, \ldots, w_T - \bar{w}])$.
  More challenging optimization problem [Argyriou et al. 2011]

- Robust regularizer $\Omega(W) = \min_{W=V+Z} \Omega(V) + \mathrm{sparse}(Z)$
  e.g. robustness against outlier tasks [Chen et al. 2011]

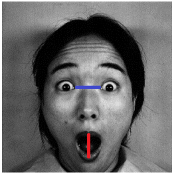
- Heterogeneous multitask feature learning [Argyriou et al. 2008b, Kang et al. 2011, Romera-Paredes et al., 2012]

- Extension of sparse coding [Olshausen and Field 1996] to MTL [Maurer et al. 2012] (see also [Kumar and Daumé III, 2012])

Diversification of features across groups

Example: recognizing identity and emotion on a set of faces

- emotion related feature
- identity related feature



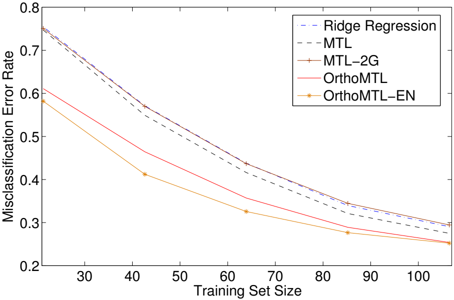
Assumptions:

- Tasks in the same group share a low dimensional representation
- Tasks from different groups tend to use different features

OrthoMTL [Romera-Paredes et al., 2012]

Encourage orthogonal features across different groups

$$\min\left\{\mathrm{Err}(W)+\mathrm{Err}(V)+\gamma\left[\|[W,V]\|_{\mathrm{tr}}+\rho\|W^\top V\|^2_{\mathrm{Fr}}\right]\right\}$$



- Related convex problem under some conditions (see paper)

Multi-task learning with dictionaries [Maurer et al. 2012]

- Method

$$\min_{U,A} \frac{1}{T} \sum_{t=1}^{T} \frac{1}{n} \sum_{i=1}^{n} \ell(\langle Ua_t, x_{ti} \rangle, y_{ti})$$

- $w_t = Ua_t$, where $a_t \in R^K$ and $U = [u_1, \ldots, u_K]$ (may be linearly dependent)
- Sparse coding constraint: $\|a_t\|_1 \leq \alpha$
- Scale constraint: $\|u_k\|_2 \leq 1$, $\{u_k\}_{k=1}^K$

Multi-task learning with dictionaries (II)

Theorem [Maurer, P., Romera-Paredes, 2012] Let X be the unit ball of a separable Hilbert space. Let $\delta > 0$ and $\mu_1, \ldots, \mu_T$ be probability measures on $X \times R$. With probability $\geq 1 - \delta$ in the draw of $\mathbf{z}_t \sim (\mu_t)^n$, $t = 1, ..., T$

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{z \sim \mu_t} \ell(\langle \hat{U} \hat{a}_t, x_{ti} \rangle, y_{ti}) - \inf_{U,A} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{z \sim \mu_t} \ell(\langle U a_t, x \rangle, y)$$

$$\leq L\alpha \sqrt{\frac{2K \mathrm{tr}(\hat{\Sigma})}{nT}} + L\alpha \sqrt{\frac{8 \|\hat{\Sigma}\| \log(2K)}{n}} + \sqrt{\frac{8 \log \frac{4}{\delta}}{nT}}$$

- Uniform distribution: $\mathrm{tr}(\hat{\Sigma}) \approx 1$, $\|\hat{\Sigma}\| \approx 1/n$
- $T < K$: tasks are learned independently
- $T > K$: term $\frac{\log K}{n}$ controls the bound (compare to $O(\sqrt{K/m})$ for independent task learning)

Conclusions

- Multi-task learning is ubiquitous – exploiting task relatedness can enhance learning performance
- Multi-task learning can be seen as a problem of matrix estimation
- Reviewed different types of regularization methods, which naturally extend complexity notions used in the single-task setting, addressing their statistical and computational properties
- Recent method to diversify features across heterogeneous groups of tasks
- MTL extension of sparse coding

Thanks

- Andreas Argyriou
- Nadia Berthouze
- Andrea Caponnetto
- Theodoros Evgeniou
- Karim Lounici
- Andreas Maurer
- Charles Micchelli
- Bernardino Romera-Paredes
- Alexandre Tsybakov
- Sara van de Geer
- Yiming Ying

Announcement: check out our 1-year master programme:
`http://www.csml.ucl.ac.uk/courses/msc_ml/`

References (I)

[Argyriou, Evgeniou, Pontil] **Multi-task feature learning.** NIPS 2006.

[Argyriou, Evgeniou, Pontil] **Convex multi-task feature learning.** Machine Learning 2008.

[Argyriou, Maurer, Pontil] **An algorithm for transfer learning in a heterogeneous environment**. ECML 2008b.

[Argyriou, Micchelli, Pontil] **When is there a representer theorem? Vector versus matrix regularizers.** JMLR 2009.

[Argyriou, Micchelli, Pontil, Shen, Xu] **Efficient first order methods for linear composite regularizers.** arXiv:1104.1436.

[Baxter] **A model for inductive bias learning.** JAIR 2000.

[Ben-David and Schuller] **Exploiting task relatedness for multiple task learning.** COLT 2003.

[Caponnetto, Micchelli, Pontil, Ying] **Universal mult-task kernels.** JMLR 2008.

[Carmeli, De Vito, Toigo] **Vector valued reproducing kernel Hilbert spaces, integrable functions and Mercer theorem.** Analysis and Applications, 2006.

[Caruana] **Multi–task learning.** Machine Learning 1998.

[Evgeniou and Pontil] **Regularized multi-task learning.** SIGKDD 2004.

[Evgeniou, Micchelli, Pontil] **Learning multiple tasks with kernel methods.** JMLR 2005.

[Fazel, Hindi and Boyd] **A rank minimization heuristic with application to minimum order system approximation.** American Control Conference, 2001.

[Lounici, Pontil, Tsybakov, van de Geer] **Taking advantage of sparsity in multi-task learning.** COLT 2009.

References (II)

[Lounici, Pontil, Tsybakov, van de Geer] **Oracle inequalities and optimal inference under group sparsity.** Annals of Statistics 2011.

[Izenman] **Reduced-rank regression for the multivariate linear model**, J. Multivariate Analysis, 1975.

[Jacob, Bach, Vert] **Clustered multi-task learning: a convex formulation.** NIPS 2008.

[Lenk, DeSarbo, Green, Young] **Hierarchical Bayes conjoint analysis: recovery of partworth heterogeneity from reduced experimental designs.** Marketing Science 1996.

[Maurer] **Bounds for linear multi-task learning.** JMLR 2006.

[Maurer and Pontil] **Structured sparsity and generalization.** JMLR 2012.

[Maurer, Pontil, Romera-Paredes] **Sparse coding for multitask and transfer learning.** arXiv:1209.0738.

[Micchelli, Morales, Pontil] **A family of penalty functions for structured sparsity.** NIPS 2010.

[Micchelli, Morales, Pontil] **Regularizers for structured sparsity.** Adv. Comp. Math. (to appear).

[Micchelli and Pontil] **On learning vector-valued functions.** Neural Computation 2005.

[Micchelli and Pontil] **On learning vector-valued functions.** Neural Computation, 2005.

[Romera-Paredes, Argyriou, Pontil, Berthouze] **Exploiting unrelated tasks in multi-task learning.** AISTATS 2012.

References (III)

[Salakhutdinov, Torralba, Tenenbaum] **Learning to share visual appearance for multiclass object detection**. CVPR 2011.

[Srivastava and Dwivedi] **Estimation of seemingly unrelated regression equations: A brief survey** J. Econometrics,1971.

[Silver & Mercer] **The parallel transfer of task knowledge using dynamic learning rates based on a measure of relatedness.** Connection Science 1996. [Yu, Tresp, Schwaighofer] **Learning Gaussian processes from multiple tasks.** ICML 2005.

[Thrun and Pratt] **Learning to learn**, Springer, 1998.

[Thrun and OSullivan]. **Clustering learning tasks and the selective crosstask transfer of knowledge.** 1998.

[Zellner] **An efficient method for estimating seemingly unrelated regression equations and tests for aggregation bias.** JASA, 1962.