Efficient Multitask Feature and Relationship Learning

Han Zhao, Otilia Stretcu, Alex Smola, Geoff Gordon

Carnegie Mellon University, Amazon, Microsoft Research

Motivation

Multitask Learning:
- Input: W, (human, dog), (male, female)
- Joint Learning: W
- Target: (human, dog)
- Feature: W
- Task: W

Solvers for W when \(\Sigma_1, \Sigma_2\) are fixed:
minimize \(h(W) = ||Y - XW||_F^2 + \eta ||W||_F^2 + \rho ||\Sigma_1^{1/2}W\Sigma_2^{1/2}||_F^2\)

Three different solvers:
- A closed form solution with \(O(m^3d^3 + mnd^2)\) complexity:
  \[\text{vec}(W^*) = (I_m \otimes (X^TX) + \eta I_m + \rho \Sigma_2 \otimes \Sigma_1)^{-1} \text{vec}(X^TY)\]
- Gradient computation:
  \[\nabla h(W) = X^T(Y - XW) + \eta W + \rho \Sigma_1 W \Sigma_2\]

Conjugate gradient descent with \(O(\sqrt{\frac{1}{\varepsilon} (m^3d^3 + mnd^2)})\) complexity, \(\varepsilon\) is the condition number, \(\varepsilon\) is the approximation accuracy.
- Sylvester equation AX + XB = C using the Bartels-Stewart solver.
  The first-order optimality condition:
  \[\Sigma_1^{-1}(X^TX + \eta I)dW + W(\rho \Sigma_2) = \Sigma_1^{-1}X^TY\]

Exact solution for W computable in \(O(m^3 + d^3 + nd^2)\) time.

Optimization Algorithm

Solvers for \(\Sigma_1, \Sigma_2\) when W is fixed:
minimize \(\text{tr}(\Sigma_1 W \Sigma_2 W^T) - m \log(\Sigma_1)_\lambda\)
subject to \(U_d \leq \Sigma_1 \leq u I_d\)

minimize \(\text{tr}(\Sigma_1 W \Sigma_2 W^T) - d \log(\Sigma_2)_\lambda\)
subject to \(U_d \leq \Sigma_2 \leq u I_d\)

Exact solution by reduction to minimum-weight perfect matching:

Algorithms:

Input: \(W, \Sigma_1, \Sigma_2\) and \(\lambda, u, \eta\)
1. \([V, \nu] \leftarrow \text{SVD}(W \Sigma_1 W^T)\)
2. \(\Sigma_1 \leftarrow \nu \Sigma_1^\nu(d/\nu)\)
3. \(\Sigma_2 \leftarrow \nu \text{diag}(\lambda) V^T\)

- Exact solution only requires one SVD
- Time complexity: \(O(\max\{dn^2, md^2\})\)

Formulation

Empirical Bayes with prior:
\[W \mid \xi, \Omega_1, \Omega_2 \sim \left(\prod_{i=1}^N \mathcal{N}(w_i | 0, \xi I_d)\right) \cdot M \mathcal{N}_{d \times m}(W | 0_{d \times m}, \Omega_1, \Omega_2)\]

- \(\mathcal{N}_{d \times m}(W | 0_{d \times m}, \Omega_1, \Omega_2)\) is matrix-variate normal distribution
- \(\Omega_1 \in \mathbb{S}_{d \times d}^+, \Omega_2 \in \mathbb{S}_{m \times m}^+, \Omega_2 \) covariance matrix over features
- \(\Omega_2 \) covariance matrix over tasks
- \(W \in \mathbb{R}^{d \times m}\), weight matrix

Maximum marginal likelihood with empirical estimators:
iminimize \(\min_{W : \Omega_1, \Omega_2} \|Y - XW\|_F^2 + \eta \|W\|_F^2 + \rho \|\Sigma_1^{1/2}W\Sigma_2^{1/2}\|_F^2\)
subject to \(U_d \leq \Sigma_1 \leq u I_d, U_m \leq \Sigma_2 \leq u I_m\)

where \(\Omega_1 := \Omega_1^{-1}, \Omega_2 := \Omega_2^{-1}\)
- Multi-convex in \(W, \Sigma_1, \Sigma_2\)

Nonlinear extension:
- Replace the feature matrix \(X\) with the output of a neural network \(g(x, \theta)\) with learnable parameters \(\theta\).
- Estimate \(W\) and \(\theta\) using backpropagation.
- Optimize the two covariance matrices using our proposed approach.

Datasets:

- Synthetic data:
  - Randomly sample \(10^4\) instances, shared among all the tasks.
  - Gradually increase the dimension of features, \(d\), and the number of tasks, \(m\), to test scalability.
- Robot data (SARCOS):
  - \(d = 21, 7\) joint positions, \(7\) joint velocities, \(7\) joint accelerations, \(m = 7\) joint torques.
  - \(4,484\) train instances, \(4,449\) test instances.
- School data:
  - \(d = 27, m = 139, n = 15, 362\) instances.
  - Goal: predict student scores.

Convergence analysis:
- The closed form solution does not scale when \(md \geq 10^4\).

Experiments

Results (mean squared error):

<table>
<thead>
<tr>
<th>Method</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
<th>5th</th>
<th>6th</th>
<th>7th</th>
</tr>
</thead>
<tbody>
<tr>
<td>STL</td>
<td>31.40</td>
<td>22.90</td>
<td>9.13</td>
<td>10.30</td>
<td>0.14</td>
<td>0.84</td>
<td>0.46</td>
</tr>
<tr>
<td>MTFI</td>
<td>31.41</td>
<td>22.91</td>
<td>9.13</td>
<td>10.33</td>
<td>0.14</td>
<td>0.83</td>
<td>0.45</td>
</tr>
<tr>
<td>MTRL</td>
<td>31.09</td>
<td>22.69</td>
<td>9.08</td>
<td>9.74</td>
<td>0.13</td>
<td>0.83</td>
<td>0.45</td>
</tr>
<tr>
<td>STFRL</td>
<td>31.39</td>
<td>22.68</td>
<td>9.10</td>
<td>9.74</td>
<td>0.13</td>
<td>0.83</td>
<td>0.45</td>
</tr>
<tr>
<td>FE FZR</td>
<td>31.08</td>
<td>22.68</td>
<td>9.08</td>
<td>9.73</td>
<td>0.13</td>
<td>0.83</td>
<td>0.43</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
<th>5th</th>
<th>6th</th>
<th>7th</th>
</tr>
</thead>
<tbody>
<tr>
<td>STL-N</td>
<td>24.81</td>
<td>17.20</td>
<td>8.97</td>
<td>8.36</td>
<td>0.13</td>
<td>0.72</td>
<td>0.34</td>
</tr>
<tr>
<td>MTFI-N</td>
<td>12.01</td>
<td>10.54</td>
<td>5.02</td>
<td>7.15</td>
<td>0.09</td>
<td>0.70</td>
<td>0.27</td>
</tr>
<tr>
<td>MTRL-N</td>
<td>11.02</td>
<td>9.51</td>
<td>4.99</td>
<td>7.11</td>
<td>0.08</td>
<td>0.62</td>
<td>0.27</td>
</tr>
<tr>
<td>FETR-N</td>
<td>10.77</td>
<td>9.34</td>
<td>4.95</td>
<td>7.01</td>
<td>0.08</td>
<td>0.59</td>
<td>0.24</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>School</th>
</tr>
</thead>
<tbody>
<tr>
<td>STL</td>
<td>0.9882 ± 0.0196</td>
</tr>
<tr>
<td>MTFI</td>
<td>0.8891 ± 0.0380</td>
</tr>
<tr>
<td>MTRL</td>
<td>0.9007 ± 0.0407</td>
</tr>
<tr>
<td>STFRL</td>
<td>0.8451 ± 0.0197</td>
</tr>
<tr>
<td>FE FZR</td>
<td>0.8134 ± 0.0253</td>
</tr>
</tbody>
</table>

Feature covariance matrix and task covariance matrix:

(a) Covariance matrix over features. (b) Covariance matrix over tasks.