Multilinear Multitask Learning

Rethinking Convex Relaxations for Tensor Completion

Bernardino Romera-Paredes

University College London

25/09/2013

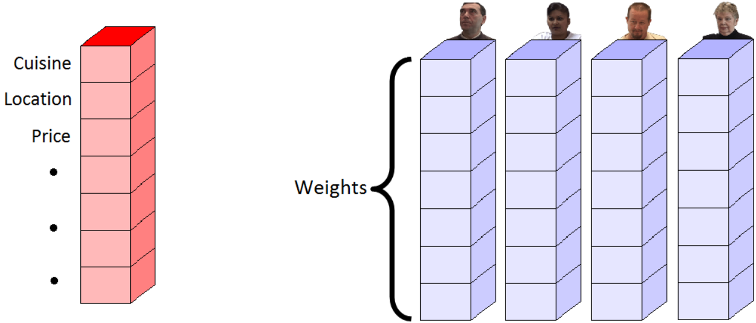
LSOLDM

Outline

- Problem and motivation
- Proposed solution
- Non-convex approach
- Convex approach
- Rethinking the convex approach
- Conclusion

Problem

We would like to learn how people value a product or service in terms of its features



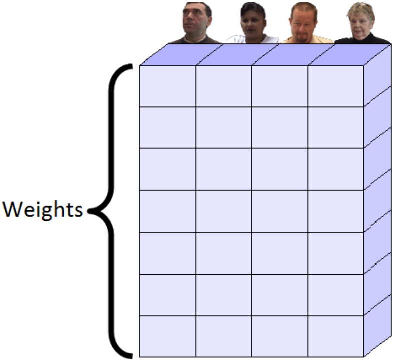
E.g: Value restaurants in terms of their features

Problem

Assumption: the way people rate restaurants is related $\longrightarrow$ Multitask learning

$$\underset{W}{\operatorname{argmin}} \sum_{t=1}^{T} \|X_t w_t - y_t\|_2^2 + \gamma \operatorname{rank}(W)$$

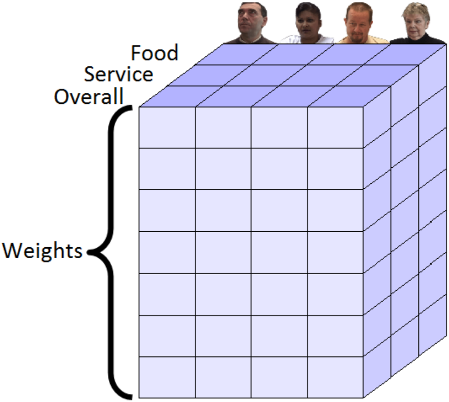
Generalization of matrix completion (collaborative filtering)



Problem

Multitask learning (MTL) scenario in which tasks can be referenced by multiple indices

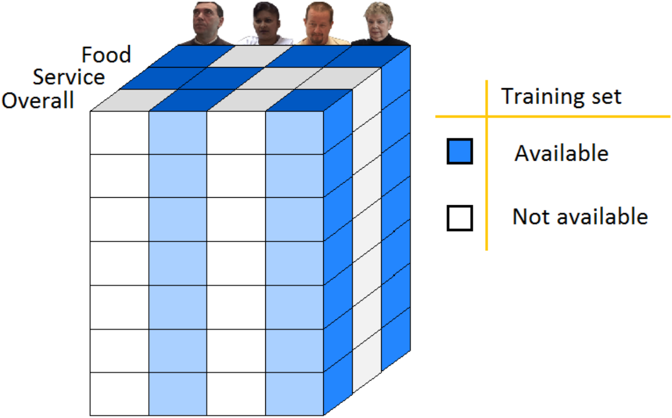
E.g: ( , Food)



Multi-dimensional indexing information would be lost in a traditional MTL approach

Transfer Learning

Advantage: It can learn tasks even in the absence of training data



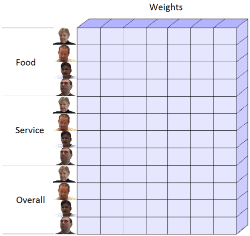
Proposed solution: Multilinear Multitask Learning (MLMTL)

Multilinear models are a natural underpinning to represent this structural information:

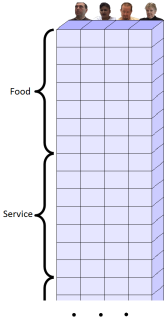
$$\underset{\mathcal{W}}{\operatorname{argmin}}\, F(\mathcal{W}) + \frac{\gamma}{N}\sum_{n=1}^{N} \operatorname{rank}\left(W_{(n)}\right)$$

$W_{(n)}$ is the n-th matricization of the tensor. E.g:

$W_{(1)}$

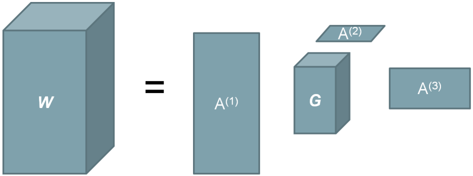


$W_{(3)}$



Non-convex approach: Tucker decomposition

We rely on the Tucker decomposition to look for low rank representations of the tensor



We attempt to solve this problem by alternating minimization

$$\underset{\mathcal{G}, A^{(1)} \ldots A^{(N)}}{\operatorname{argmin}} F\left(\mathcal{G} \times_1 A^{(1)} \ldots \times_N A^{(N)}\right) + \gamma \left(\|\mathcal{G}\|_{\mathrm{Fr}}^2 + \sum_{n=1}^{N} \|A^{(n)}\|_{\mathrm{Fr}}^2 \right)$$

Convex approach

The trace norm is a widely used convex surrogate for the rank. Therefore, we can consider the following convex relaxation:

$$\underset{\mathcal{W}}{\operatorname{argmin}}\, F(\mathcal{W}) + \frac{\gamma}{N} \sum_{n=1}^{N} \left\| W_{(n)} \right\|_{\mathrm{Tr}}$$

Regularizer previously employed for Tensor Completion (Liu et al, 2009), (Gandy et al, 2011), (Signoretto et al, 2012)

Alternating Direction Method of Multipliers (ADMM)

- We want to minimize

$$\min_{\mathcal{W}} \frac{N}{\gamma} F(\mathcal{W}) + \sum_{n=1}^{N} \left\| W_{(n)} \right\|_{\mathrm{Tr}}$$

- Decouple the regularization term

$$\min_{\mathcal{W},\mathcal{B}} \left\{ \frac{N}{\gamma} F(\mathcal{W}) + \sum_{n=1}^{N} \left\| B_{n(n)} \right\|_{\mathrm{Tr}} \ : \ \mathcal{B}_n = \mathcal{W},\ n = 1, \ldots, N \right\}$$

- Augmented Lagrangian:

$$\mathcal{L}(\mathcal{W}, \mathcal{B}, \mathcal{C}) =$$

$$\frac{N}{\gamma} F(\mathcal{W}) + \sum_{n=1}^{N} \left(\left\| B_{n(n)} \right\|_{\mathrm{Tr}} - \langle \mathcal{C}_n, \mathcal{W} - \mathcal{B}_n \rangle + \frac{\beta}{2} \left\| \mathcal{W} - \mathcal{B}_n \right\|_2^2 \right)$$

Alternating Direction Method of Multipliers (ADMM)

$$\mathcal{L}(\mathcal{W}, \mathcal{B}, \mathcal{C}) =$$
$$\frac{N}{\gamma} F(\mathcal{W}) + \sum_{n=1}^{N} \left(\left\| B_{n(n)} \right\|_{\mathrm{Tr}} - \langle \mathcal{C}_n, \mathcal{W} - \mathcal{B}_n \rangle + \frac{\beta}{2} \left\| \mathcal{W} - \mathcal{B}_n \right\|_2^2 \right)$$

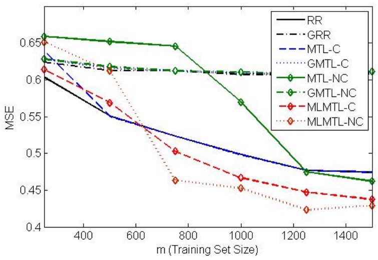
Updating equations:

- $\mathcal{W}^{[i+1]} \leftarrow \underset{\mathcal{W}}{\operatorname{argmin}} \, \mathcal{L}\left(\mathcal{W}, \mathcal{B}^{[i]}, \mathcal{C}^{[i]}\right)$
- $\mathcal{B}_n^{[i+1]} \leftarrow \underset{\mathcal{B}_n}{\operatorname{argmin}} \, \mathcal{L}\left(\mathcal{W}^{[i+1]}, \mathcal{B}, \mathcal{C}^{[i]}\right)$
- $\mathcal{C}_n^{[i+1]} \leftarrow \mathcal{C}_n^{[i+1]} - \left(\beta \mathcal{W}^{[i+1]} - \mathcal{B}_n^{[i+1]}\right)$

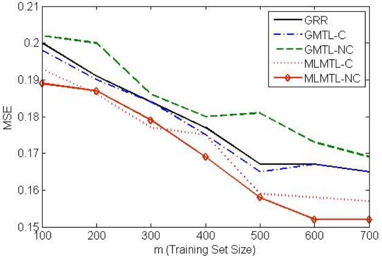
2nd step involves the computation of proximity operator of $\|\cdot\|_{\mathrm{Tr}}$.

Experiments

Restaurant dataset
(All tasks have training instances)



Shoulder Pain dataset
(Some tasks have no training instances)



Remarks

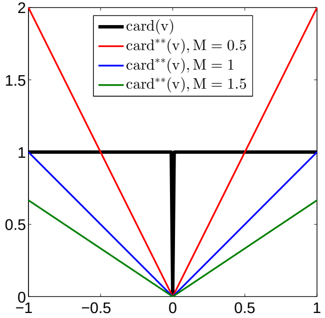
- When tasks are referenced by multiple indices, MLMTL methods outperform other approaches.
- The MLMTL non-convex approach obtains slightly better results than the convex counterpart.

Rethinking the convex approach

Convex envelope of a function f on a set S is the largest convex function f^{**} majorized by f for all points in S

E.g: cardinality of a vector:

- $f(v) = card(v)$
- $S = \{v : ||v||_\infty \leq M\}$
- $f^{**}(v) = ||v||_1 / M$



In practise M is unknown and tuned by cross validation.
Trade off: the smaller S, the tighter the convex envelope.

Rethinking the convex approach

- In the regular MTL case, W is a matrix and we want to use the regularizer $\mathrm{rank}(\mathrm{W})$
- (Fazel 2001) $\|W\|_{\mathrm{Tr}}/M$ is the convex envelope of $\mathrm{rank}(\mathrm{W})$ in the set

$$\left\{W : \|W\|_{\mathrm{Sp}} \leq M\right\}$$

- In the MLMTL case, by using the regularizer $\sum_{n=1}^{N} \left\|W_{(n)}\right\|_{\mathrm{Tr}}$ we implicitly assume the same M for the different matricizations.
- However:

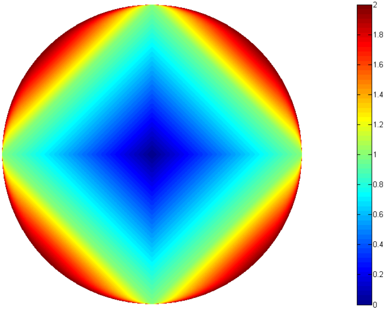
$$\left\|W_{(1)}\right\|_{\mathrm{Sp}} \neq \left\|W_{(2)}\right\|_{\mathrm{Sp}} \neq \ldots \neq \left\|W_{(N)}\right\|_{\mathrm{Sp}}$$

Rethinking the convex approach

- We are interested in convex functions over matrices invariant to matricizations of a tensor.
- The Frobenius norm is very appealing:
  - $\left\|W_{(1)}\right\|_{\mathrm{Fro}} = \left\|W_{(2)}\right\|_{\mathrm{Fro}} = \ldots = \left\|W_{(N)}\right\|_{\mathrm{Fro}}$
  - It is also a spectral function
- Therefore, we consider the set $S = \{W : \|W\|_{\mathrm{Fro}} \leq M\}$
- In that set, calculating the convex envelope of the rank can be reduced to calculate the convex envelope of $\mathrm{card}\,(\mathrm{v})$ on the set $\{v : \|v\|_2 \leq M\}$, where v is the vector of singular values of W.

Rethinking the convex approach

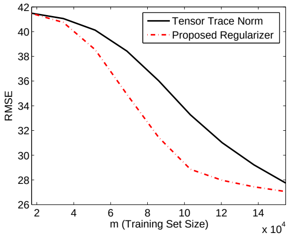
- Convex envelope of $\mathrm{card}\,(\mathrm{v})$ on the set $\{v \,:\, \|v\|_2 \leq M\}$
- Property:
  If $\|v\|_2 = M \rightarrow$ $\mathrm{card}(\mathrm{v}) = \mathrm{card}^{**}(\mathrm{v})$



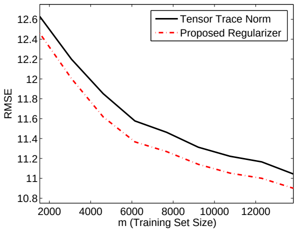
- The resultant function is difficult to compute explicitly.
- However, it is feasible to compute its proximal operator (Romera-Paredes & Pontil, 2013).
  - That is all we need to solve the problem via ADMM!

Experiments on tensor completion

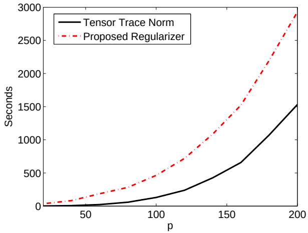
Video compression
($160 \times 112 \times 32 \times 3$ tensor)



Exam score prediction
($139 \times 11 \times 3 \times 3 \times 2$ tensor)



Time comparison:



Conclusions

- MLMTL approaches account for the scenario where tasks are described by multiple indices
  - They get a significant improvement over conventional approaches
- In the convex approach, we have found out that the trace norm is not the best option for tensor regularization
  - We have proposed an alternative based on the convex envelope of the rank on the Frobenius ball