Learning Nonparametric Kernel Matrices from Pairwise Constraints

--- ICML'07

Presenter: Lei Tang

Nonparametric Kernel Learning

- Existing methods all assume certain parametric form of the kernel;
- Or a linear combination of provided kernels.
- This work focus on non-parametric kernel learning from both labeled and unlabeled data.
- Actually, here kernel learning is more appropriate considered as a similarity matrix which must be semi-positive definite.

Problem Formulation

- Given: unlabeled data and some side information (i.e. must-link (S), cannot-link (D))

$$T_{i,j} = \begin{cases} +1 & (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{S} \\ -1 & (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{D} \\ 0 & \text{otherwise} \end{cases}$$

- Goal: Identify a kernel matrix that is consistent with all the pairwise constraints.

$$\underset{Z=V^\top V}{\arg\min} \; \|V\|_2^2 + c \sum_{(i,j)\in(\mathcal{S}\cup\mathcal{D})} \max\left(0, 1 - T_{i,j} Z_{i,j}\right)$$

Use Graph Laplacian as Regularizer

- The previous formulation dose not take into consideration the input pattern of data instances.
- Use Laplacian regularizer :

$$l(V,S) = \sum_{i,j=1}^{n} \frac{S_{i,j}}{\sqrt{d_i d_j}} \|\mathbf{v}_i - \mathbf{v}_j\|_2^2 = \mathrm{tr}(VLV^\top)$$

- Here, different from spectral clustering, v_i and v_j are vectors.
- So the objective is:

$$\arg\min_{Z=V^\top V} \mathrm{tr}(VLV^\top) + c \sum_{(i,j)\in(\mathcal{S}\cup\mathcal{D})} \max\left(0, 1 - T_{i,j} Z_{i,j}\right)$$

Primal Formulation

$$\arg\min_{Z=V^\top V} \ \mathrm{tr}(VLV^\top) + c \sum_{(i,j)\in(\mathcal{S}\cup\mathcal{D})} \max\left(0, 1 - T_{i,j} Z_{i,j}\right)$$

As $\mathrm{tr}(VLV^\top) = \mathrm{tr}(LV^\top V) = \mathrm{tr}(LZ).$

SDP

$$\begin{aligned} \arg\min_{Z,\epsilon} \quad & \sum_{i,j=1}^{N} L_{i,j} Z_{i,j} + c \sum_{(i,j)\in(\mathcal{S}\cup\mathcal{D})} \epsilon_{i,j} \qquad (3) \\ \text{s. t.} \quad & \forall (i,j) \in (\mathcal{S}\cup\mathcal{D}),\ T_{i,j} Z_{i,j} \ge 1 - \epsilon_{i,j},\ \epsilon_{i,j} \ge 0 \\ & Z \succeq 0 \end{aligned}$$

#var = N*N+|S|+|D|

Dual Formulation

$$
\begin{aligned}
\mathcal{L} &= \sum_{i,j=1}^{N} L_{i,j} Z_{i,j} + c \sum_{(i,j)\in(\mathcal{S}\cup\mathcal{D})} \epsilon_{i,j} \\
&\quad - \sum_{(i,j)\in(\mathcal{S}\cup\mathcal{D})} Q_{i,j}\left(T_{i,j} Z_{i,j} - 1 + \epsilon_{i,j}\right) \\
&\quad - \sum_{(i,j)\in(\mathcal{S}\cup\mathcal{D})} \xi_{i,j}\epsilon_{i,j} - \mathrm{tr}(MZ)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \epsilon_{i,j}} &= c - Q_{i,j} - \xi_{i,j} = 0 \rightarrow Q_{i,j} \le c \\
\frac{\partial \mathcal{L}}{\partial Z_{i,j}} &= L_{i,j} - Q_{i,j} T_{i,j} - M_{i,j} = 0 \rightarrow L \succeq Q \otimes T
\end{aligned}
$$

$$
\begin{aligned}
\arg\max_{Q} \quad & \sum_{(i,j)\in\mathcal{S}} Q_{i,j} + \sum_{(i,j)\in\mathcal{D}} Q_{i,j} \\
\text{s. t.} \quad & 0 \le Q_{i,j} \le c,\ \forall (i,j) \in (\mathcal{S}\cup\mathcal{D}) \\
& L \succeq Q \otimes T
\end{aligned}
$$

SDP

#var is equivalent to number of pairwise constraints

Efficient Dual Algorithm

- Using SDP solver to solve dual formulation first.
- Recover the primal kernel matrix Z efficiently based on KKT conditions:

$$M = L - Q \otimes T, \qquad MZ = 0$$

- Z can be factorized as $Z = UBU^\top$ where U is the eigenvector of M corresponding eig-value 0.
- |B|<=|S|+|D|+1

❖ Rank(L)=n-1, Rank($Q \otimes T$) <=|S|+|D|,

❖ Rank(M)>=n-1-|S|-|D|. Hence, the number of eigenvectors for zero eigenvalue <= |S|+|D|+1

Reformulated Primal

- Plug in the factorization of Z, we have

$$\begin{aligned} \arg\min_{B \succeq 0} \quad & \sum_{i,j=1}^{N} L_{i,j} Z_{i,j} + c \sum_{(i,j)\in(\mathcal{S}\cup\mathcal{D})} \epsilon_{i,j} \\ \text{s. t.} \quad & \forall (i,j) \in (\mathcal{S}\cup\mathcal{D}),\ T_{i,j} Z_{i,j} \geq 1 - \epsilon_{i,j}, \epsilon_{i,j} \geq 0 \\ & Z = UBU^\top \end{aligned}$$

$$\begin{aligned} \arg\min \quad & \mathrm{tr}(BU^\top LU) + c \sum_{(i,j)\in(\mathcal{S}\cup\mathcal{D})} \epsilon_{i,j} \qquad (6) \\ \text{s. t.} \quad & \forall (i,j) \in (\mathcal{S}\cup\mathcal{D}),\ T_{i,j} \mathbf{u}_i^\top B \mathbf{u}_j \geq 1 - \epsilon_{i,j} \\ & \forall (i,j) \in (\mathcal{S}\cup\mathcal{D}), \epsilon_{i,j} \geq 0 \\ & B \succeq 0 \end{aligned}$$

SDP with smaller #var

Algorithm Overview

- Solve Dual problem first obtain the dual matrix.
- Get the Langrange multiplier M

$$M = L - Q \otimes T, \qquad MZ = 0$$

- Calculate its eigen vector of zero eigen value.
- Solve the simplified primal problem:

$$\begin{aligned} \arg\min \quad & \mathrm{tr}(BU^\top LU) + c \sum_{(i,j)\in(\mathcal{S}\cup\mathcal{D})} \epsilon_{i,j} \qquad (6) \\ \text{s. t.} \quad & \forall (i,j) \in (\mathcal{S}\cup\mathcal{D}),\ T_{i,j}\mathbf{u}_i^\top B \mathbf{u}_j \geq 1 - \epsilon_{i,j} \\ & \forall (i,j) \in (\mathcal{S}\cup\mathcal{D}), \epsilon_{i,j} \geq 0 \\ & B \succeq 0 \end{aligned}$$

- Recover kernel matrix $Z = UBU^\top$

SMO-like algorithm for Dual

- Principle of SMO: Each iteration
  - with respect to small number of vars.
  - Closed-form solution.
- Here, for the dual, we optimize in terms of just one entry in the matrix.

$$
\begin{aligned}
\arg\max_{Q} \quad & \sum_{(i,j)\in\mathcal{S}} Q_{i,j} + \sum_{(i,j)\in\mathcal{D}} Q_{i,j} \\
\text{s. t.} \quad & 0 \le Q_{i,j} \le c,\ \forall (i,j) \in (\mathcal{S} \cup \mathcal{D}) \\
& L \succeq Q \otimes T
\end{aligned}
$$

Dual

$$
\begin{aligned}
\arg\max_{Q_{k,l}} \quad & Q_{k,l} \qquad (7) \\
\text{s. t.} \quad & 0 \le Q_{i,j} \le c,\ A^{k,l} - T_{k,l} Q_{k,l} I^{k,l} \succeq 0
\end{aligned}
$$

SMO-like Dual

$$\arg\max_{Q_{k,l}} \quad Q_{k,l} \tag{7}$$

$$\text{s. t.} \quad 0 \le Q_{i,j} \le c, \; A^{k,l} - T_{k,l} Q_{k,l} I^{k,l} \succeq 0$$

where matrix $A^{k,l}$ is defined as

$$A^{k,l} = L - (\tilde{Q} - \tilde{Q}_{k,l} I^{k,l}) \otimes T. \tag{8}$$

Note $I^{k,l}$ is a $n \times n$ matrix and is defined as

$$[I^{k,l}]_{i,j} = \begin{cases} 1 & (k = i \text{ and } l = j) \\ 1 & (k = j \text{ and } l = i) \\ 0 & \text{otherwise} \end{cases}$$

Closed-form Solution (1)

- Rewrite the constraint as follows:

$$\begin{pmatrix} A_1 & W \\ W^\top & A_2 \end{pmatrix}$$

where $W \in \mathbb{R}^{(n-2)\times 2}$, $A_2 \in \mathbb{R}^{(n-2)\times(n-2)}$, and A_1 is

$$A_1 = \begin{pmatrix} L_{k,k} & L_{k,l} - Q_{k,l}T_{k,l} \\ L_{k,l} - Q_{k,l}T_{k,l} & L_{l,l} \end{pmatrix}$$

- So $A_1 \succeq W^\top A_2^{-1} W$
- Let

$$W^\top A_2^{-1} W \equiv G = \begin{pmatrix} G_{1,1} & G_{1,2} \\ G_{2,1} & G_{2,2} \end{pmatrix}$$

$$\Rightarrow \begin{pmatrix} L_{k,k} - G_{1,1} & L_{k,l} - Q_{k,l}T_{k,l} - G_{1,2} \\ L_{k,l} - Q_{k,l}T_{k,l} - G_{1,2} & L_{l,l} - G_{2,2} \end{pmatrix} \succeq 0$$

Closed-form Solution (2)

1. $L_{k,k} - G_{1,1} \geq 0$,
2. $L_{l,l} - G_{2,2} \geq 0$, and
3. the determinant of the above matrix is non-negative, i.e., $(G_{1,2} + T_{k,l}Q_{k,l} - L_{k,l})^2 \leq (L_{k,k} - G_{1,1})(L_{l,l} - G_{2,2})$.

$$\begin{aligned} \max_{Q_{k,l}} \quad & Q_{k,l} \\ \text{s.t.} \quad & 0 \leq Q_{k,l} \leq c \\ & |G_{1,2} + T_{k,l}Q_{k,l} - L_{k,l}| \leq \mu_{k,l} \end{aligned}$$

$$Q_{k,l} = \min\left(c, \mu_{k,l} - T_{k,l}G_{1,2} + T_{k,l}L_{k,l}\right)$$

Avoid the matrix inverse

- W=($W_{a\,;}$ W_b)

$$W^\top A_2^{-1} W \equiv G = \begin{pmatrix} G_{1,1} & G_{1,2} \\ G_{2,1} & G_{2,2} \end{pmatrix}$$

$$G_{a,b} = \mathbf{w}_a^\top A_2^{-1} \mathbf{w}_b$$

$$\max_{\mathbf{x}} \quad -\mathbf{x}^\top A_2 \mathbf{x} + 2\mathbf{w}_a^\top \mathbf{x} \longrightarrow \mathbf{w}_a^\top A_2^{-1} \mathbf{w}_a$$

(Can be solved efficiently using conjugate gradient methods without matrix inverse)

$$\mathbf{w}_a^\top A_2^{-1} \mathbf{w}_b = \frac{1}{2}\Big((\mathbf{w}_a + \mathbf{w}_b)^\top A_2^{-1} (\mathbf{w}_a + \mathbf{w}_b) - \mathbf{w}_a^\top A_2^{-1} \mathbf{w}_a - \mathbf{w}_b^\top A_2^{-1} \mathbf{w}_b \Big)$$

SMO-like Overview

- Optimize with respect to only one var.
- Get closed-form solution.

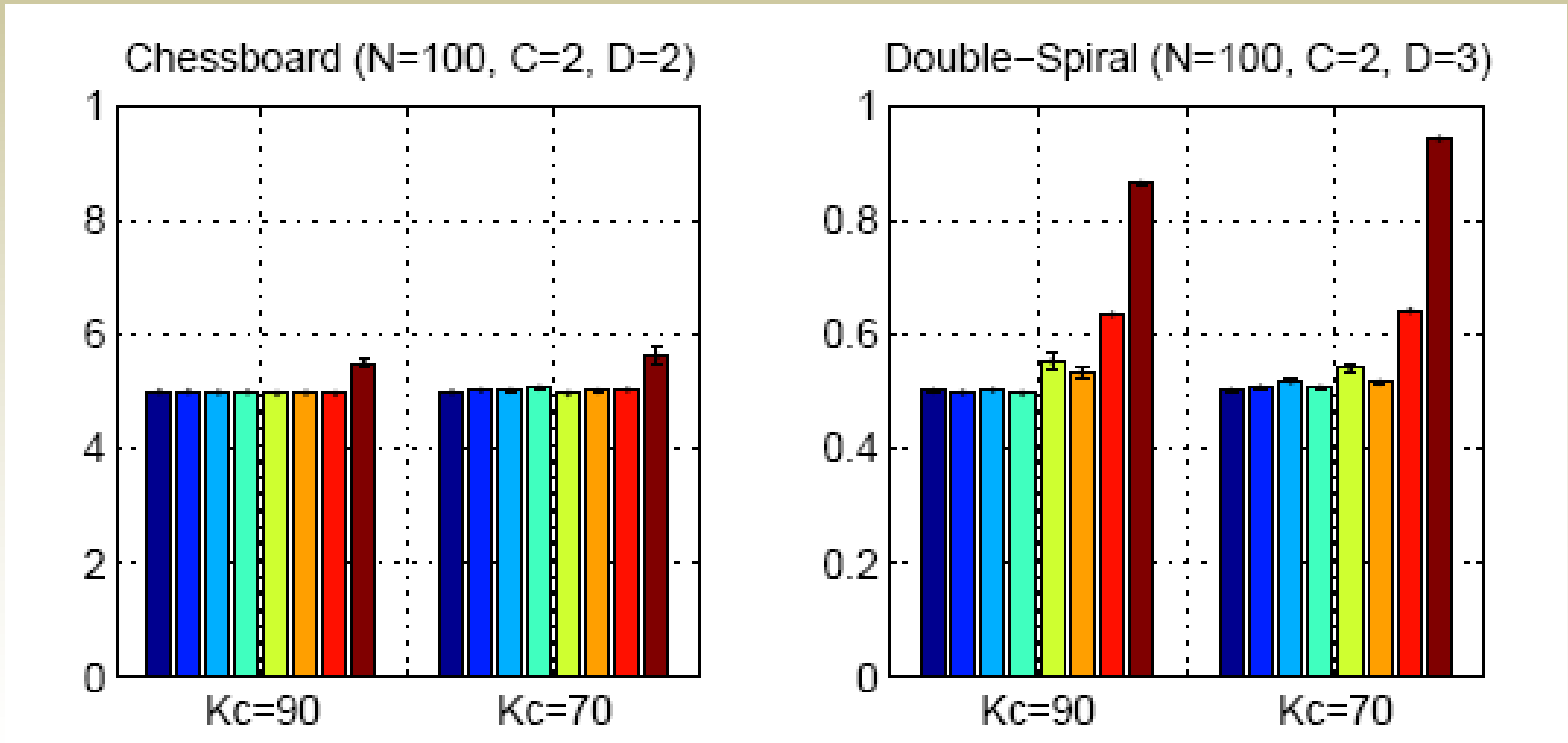
$$Q_{k,l} = \min\left(c, \mu_{k,l} - T_{k,l}G_{1,2} + T_{k,l}L_{k,l}\right)$$

- In this computation, avoid the computation of matrix inverse:

$$\mathbf{w}_a^\top A_2^{-1}\mathbf{w}_b = \frac{1}{2}\Big((\mathbf{w}_a + \mathbf{w}_b)^\top A_2^{-1}(\mathbf{w}_a + \mathbf{w}_b) - \mathbf{w}_a^\top A_2^{-1}\mathbf{w}_a - \mathbf{w}_b^\top A_2^{-1}\mathbf{w}_b\Big)$$

Experiments

- K-means
- Constrained K-means + RCA, Xing, RBF, MPK, LRK, NPK



Efficiency Comparison

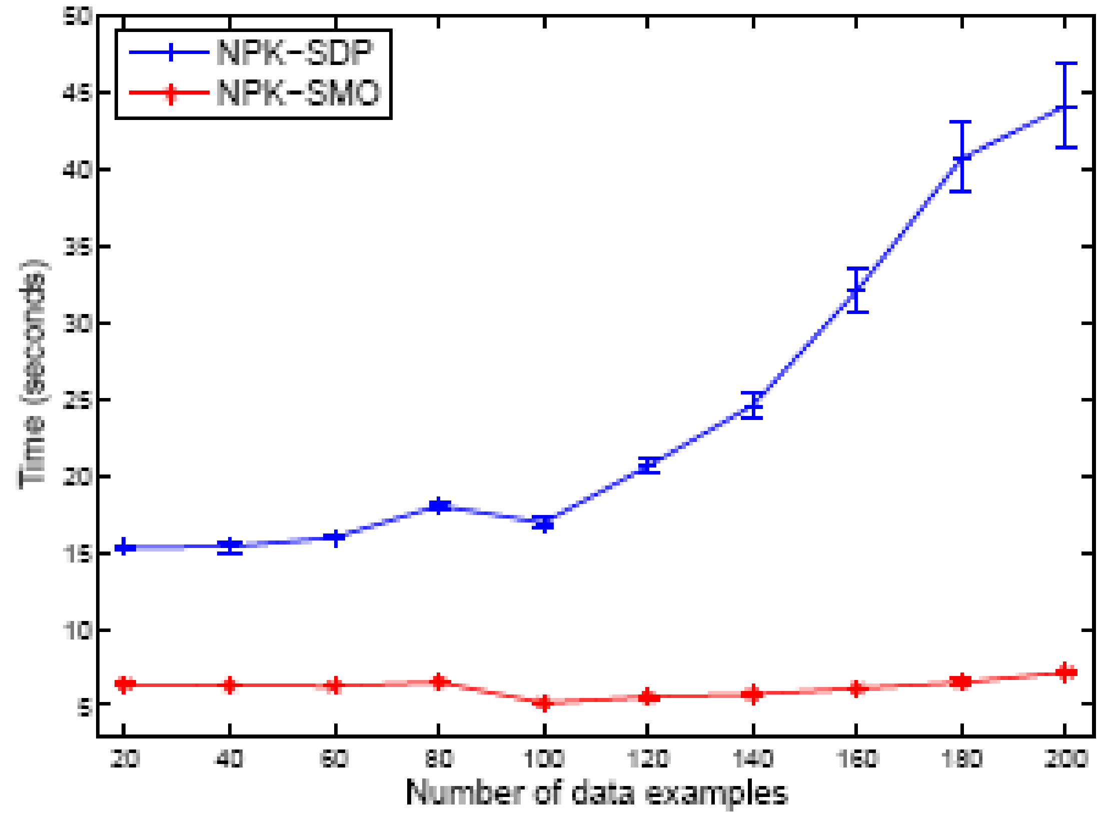


Figure 3. Time cost of different numbers of data examples. The number of pairwise constraints is fixed to 100.

Conclusions

- Use Laplacian as regularization
- Efficient SDP solver.

- But this method is transductive.
- Still too costly.
- Is it sensitive to the Laplacian?
- How to construct an optimal Laplacian?