Analysis of a Convex Formulation for Distant Supervision and Fitting a Custom Kernel

Aman Madaan
Suvadeep Hajra
Supervised by: Prof. Ganesh Ramakrishnan
Department of CSE, IIT Bombay

October 9, 2018

Weakly Supervised Relation Extraction Problem

Knowledge Base

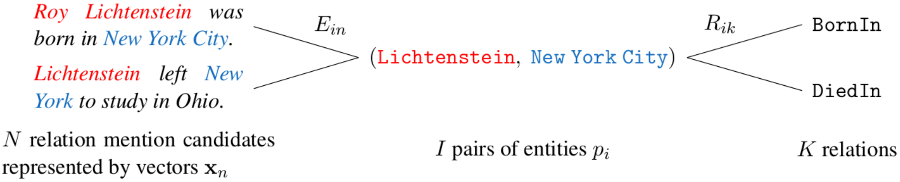
Knowledge base

r	e_1	e_2
BornIn	Lichtenstein	New York City
DiedIn	Lichtenstein	New York City

Sentences	Latent labels
Roy Lichtenstein was born in New York City, into an upper-middle-class family.	BornIn
In 1961, Leo Castelli started displaying Lichtenstein's work at his gallery in New York.	None
Lichtenstein died of pneumonia in 1997 in New York City.	DiedIn

Weakly Supervised Relation Extraction Problem

Matrices



Problem formulation

- We have
  - $X \in \mathbb{R}^{N \times D}$
  - $E \in \mathbb{R}^{I \times N}$
  - $R \in \mathbb{R}^{I \times K}$
- Need to find $Y \in \{0,1\}^{N \times (K+1)}$ such that

$$\min_{Y} \quad \min_{f} \quad \sum_{i=1}^{N} l(\mathbf{y}_n, f(\mathbf{x}_n)) + \Omega(f)$$

$$s.t. \quad Y \in \mathcal{Y}$$

Constrains on Y

- Y should satisfy
  1. $\forall n \in \{1, \ldots, N\}$, $\sum_{k=1}^{K} Y_{nk} = 1$
  2. $\forall (i, k)$ such that $R_{ik} = 1 \Rightarrow \sum_{n=1}^{N} E_{in} Y_{nk} \geq 1$
  3. $\forall (i, k)$ such that $R_{ik} = 0 \Rightarrow \sum_{n=1}^{N} E_{in} Y_{nk} = 0$
  4. $\forall i \in \{1, \ldots, I\}$, $\sum_{n=1}^{N} E_{in} Y_{n(K+1)} \leq c \sum_{n=1}^{N} E_{in}$
- All the above constraints can be written as

$$\mathrm{Y}\mathbf{1} = \mathbf{1}$$

$$(\mathrm{EY}) \circ \mathrm{S} \geq \tilde{R}$$

Primal problem

- Using linear classifier, squared loss and l_2-norm regularizer,

$$
\begin{aligned}
\min_{\mathrm{Y},\mathrm{W}} \quad & \frac{1}{2}||\mathrm{Y} - \mathrm{X}\mathrm{W}||_F^2 + \frac{\lambda}{2}||\mathrm{W}||_F^2, \\
\text{s.t.} \quad & \mathrm{Y} \in \{0,1\}^{N \times (K+1)}, \\
& \mathrm{Y}1 = 1, \\
& (\mathrm{E}\mathrm{Y}) \circ \mathrm{S} \geq \mathrm{R}.
\end{aligned}
$$

where $\mathrm{W} \in \mathbb{R}^{D \times (K+1)}$.

Primal problem (cont.)

- Replacing W by its optimum value, using Woodbury identity and relaxing the constrains $\mathrm{Y} \in \{0,1\}^{N\times(K+1)}$ into $\mathrm{Y} \in [0,1]^{N\times(K+1)}$,

$$\begin{aligned} \min_{\mathrm{Y}} \quad & \frac{1}{2}\mathrm{tr}(\mathrm{Y}^T(\mathrm{X}\mathrm{X}^T + \lambda \mathrm{I}_N)^{-1}\mathrm{Y}), \\ \text{s.t.} \quad & \mathrm{Y} \geq 0, \\ & \mathrm{Y}1 = 1, \\ & (\mathrm{E}\mathrm{Y}) \circ \mathrm{S} \geq \mathrm{R}. \end{aligned}$$

Primal problem (cont.)

- Finally adding slack variables $\xi \in \mathbb{R}^{I \times (K+1)}$,

$$\begin{aligned} \min_{\mathrm{Y},\xi} \quad & \frac{1}{2}\mathrm{tr}(\mathrm{Y}^T(\mathrm{X}\mathrm{X}^T + \lambda \mathrm{I}_N)^{-1}\mathrm{Y}) + \mu||\xi||_1, \\ \text{s.t.} \quad & \mathrm{Y} \geq 0, \quad \xi \geq 0, \\ & \mathrm{Y}1 = 1, \\ & (\mathrm{E}\mathrm{Y}) \circ \mathrm{S} \geq \mathrm{R} - \xi. \end{aligned}$$

Dual Problem

- Introducing Lagrangian and optimizing it against primal variables,dual problem can be given by,

$$\begin{aligned} \max_{\Lambda,\Sigma,\nu} \quad & -\frac{1}{2}\operatorname{tr}\left(\mathrm{Z}^T \mathrm{Q}\mathrm{Z}\right) + \operatorname{tr}\left(\Lambda^T \mathrm{R}\right) + \nu^T 1 \\ \text{s.t.} \quad & \Lambda_{ik} \geq 0, \quad \Sigma_{nk} \geq 0, \quad \Omega_{ik} \geq 0, \\ & \mu - \Lambda_{ik} - \Omega_{ik} = 0, \qquad \forall i, n, k. \end{aligned}$$

  where $\mathrm{Z} = \mathrm{E}^T(\mathrm{S} \circ \Lambda) + \Sigma + \nu 1^T$

- The dual problem has been solved using accelerated projected gradient descend algorithm

Difficulty in Using Custom Kernel

- Gradient of the dual cost function

$$\nabla_\Sigma f = (\mathrm{X}\mathrm{X}^T + \lambda \mathrm{I}_N)\mathrm{Z},$$
$$\nabla_\Lambda f = ((\mathrm{X}\mathrm{X}^T + \lambda \mathrm{I}_N)\mathrm{Z}\mathrm{E}^T) \circ \mathrm{S} - \mathrm{R},$$
$$\nabla_\nu f = (\mathrm{X}\mathrm{X}^T + \lambda \mathrm{I}_N)\mathrm{Z}1 - 1$$

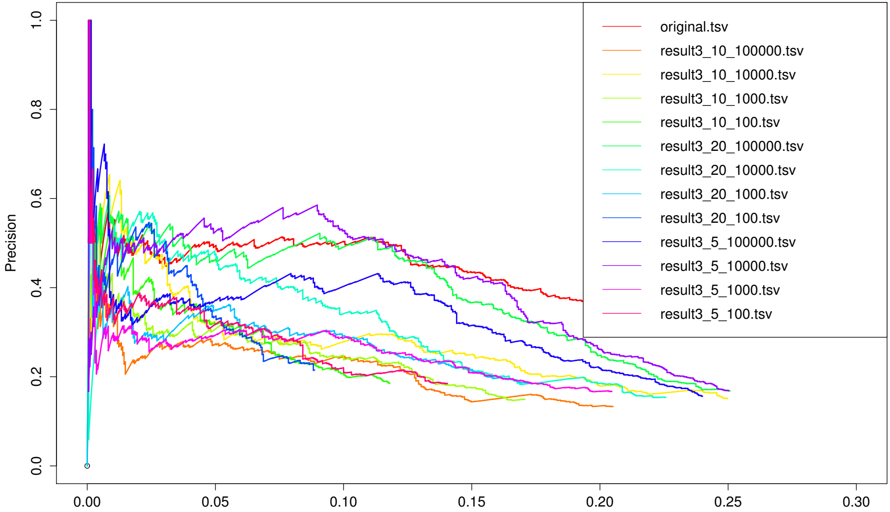
- Using sparsity of X, $\mathrm{X}\mathrm{X}^T\mathrm{Z}$ requires $\mathcal{O}(NFK)$ operations where F be the average number of features per example
- For kernelized algorithm, it requires $\mathcal{O}(N^2K)$ operations

Feature pruning

- Prunes irrelevant features
- What are the irrelevant features:
  - features appear in all instances
  - features appear rarely
- Motivation: feature appears across relation may not help in learning

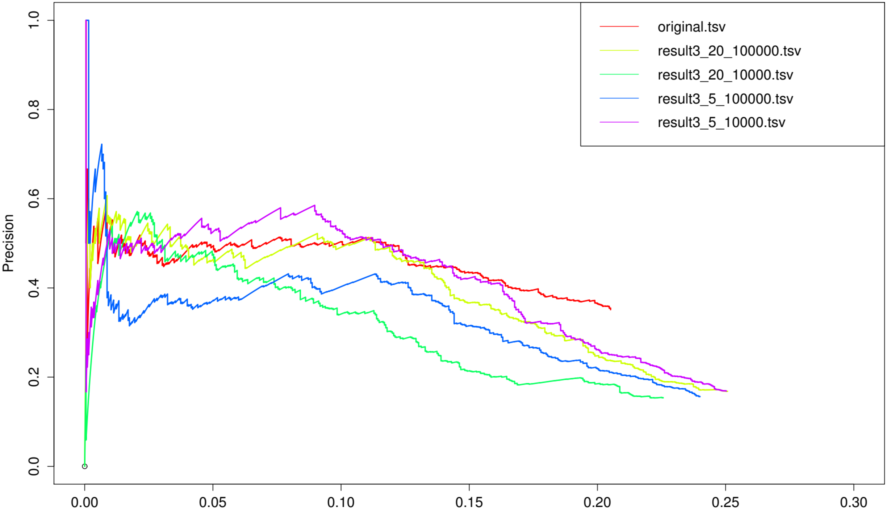
Experimental results

Precision Recall Curve for different Pruning Configurations



Experimental results (cont.)

Precision Recall Curve for the top 4 Pruning Configurations



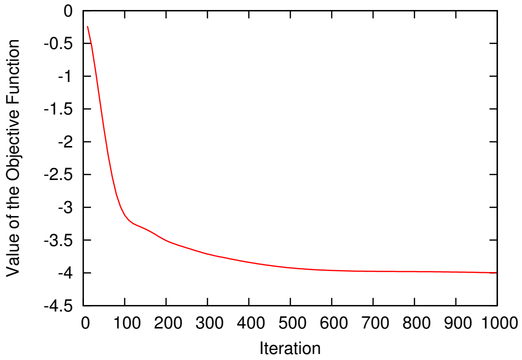
Experimental results (cont.)

AUC and time of completion

Min Pruning Cutoff	Max Pruning Cutoff	AUC	Time
20	100	0.03311205	211.513000
10	100	0.03736918	283.580000
5	100	0.04061753	386.337000
10	1000	0.04444769	633.979000
10	100000	0.04456001	1369.473000
20	1000	0.04935819	500.905000
5	1000	0.05199495	801.007000
10	10000	0.06930104	1131.839000
20	10000	0.07201145	965.479000
5	100000	0.08066457	1648.713000
No Pruning	No Pruning	0.09428642	551.461000
20	100000	0.09827496	1151.805000
5	10000	0.1033964	1365.037000

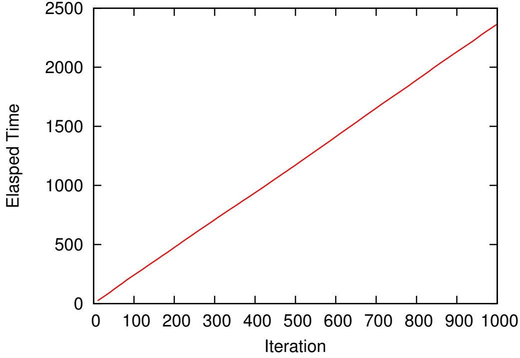
Experimental results (cont.)

Value of the objective function vs. no of iterations



Experimental results (cont.)

Elapsed time vs. no of iterations



Custom kernel using singular value decomposition (SVD)

- Let Φ be the feature matrix obtained by projecting X into new feature space
- Our proposed method works as follows:
  1. perform SVD of Φ

$$\Phi = V * \Sigma * U^T \tag{1}$$

  2. project the feature matrix into subspace obtained by the first F right singular vectors

$$\Phi' = \Phi * U_{(:,1:F)} \tag{2}$$

Thank you