Regularizing Text Categorization with Clusters of Words

Konstantinos Skianis, François Rousseau, Michalis Vazirgiannis

LIX, École Polytechnique, France

DaSciM Data Science and Mining Team École Polytechnique

l'X ÉCOLE POLYTECHNIQUE UNIVERSITÉ PARIS-SACLAY

Motivation

Text categorization is hard:

- high dimensionality
- prone to overfitting
- state-of-the-art structured regularization is slow due to overlapping clusters

Regularization is necessary:

- Critical for language modeling, structured prediction, and classification
- Prior on the feature weights
- Find the optimal weights: $\theta^* = \operatorname{argmin}_\theta \underbrace{\underbrace{\sum_{i=1}^{N} \mathcal{L}(\mathcal{X}^i, \theta, y^i)}_{\text{empirical risk}} + \underbrace{\lambda\Omega(\theta)}_{\text{penalty term}}}_{\text{expected risk}}$

I. Structured Regularization

Group lasso: $\Omega(\theta) = \lambda \sum_g \lambda_g \|\theta_g\|_2$

Objective: $\Omega_{las}(\theta) + \Omega_{glas}(\mathrm{v}) + \mathcal{L}(\theta) + \mathrm{u}^\top(\mathrm{v} - M\theta) + \frac{\rho}{2}\|\mathrm{v} - M\theta\|_2^2$

Iterative update of θ, v and u:

$$\min_\theta \Omega_{las}(\theta) + \mathcal{L}(\theta) + \mathrm{u}^\top M\theta + \frac{\rho}{2}\|\mathrm{v} - M\theta\|_2^2$$

$$\min_{\mathrm{v}} \Omega_{glas}(\mathrm{v}) + \mathrm{u}^\top \mathrm{v} + \frac{\rho}{2}\|\mathrm{v} - M\theta\|_2^2$$

$$\mathrm{u} = \mathrm{u} + \rho(\mathrm{v} - M\theta)$$

Algorithm ADMM

Input: augmented Lagrangian variable ρ, λ_{glas} and λ_{las}

```
while update in weights not small do
    θ = argmin_θ Ω_las(θ) + L(θ) + (ρ/2) Σ_{i=1}^{V} N_i(θ_i − μ_i)^2
    for g = 1 to G do
        v_g = prox_{Ω_glas, λ_g/ρ}(z_g)
    end for
    u = u + ρ(v − Mθ)
end while
```

II. Structured Regularization in NLP

STATISTICAL REGULARIZERS

- Network of features
  - $\Omega_{net}(\theta) = \lambda_{net} \sum \theta_k^\top M \theta_k$, where $M = \alpha(I - P)^\top(I - P) + \beta I$.
- Sentence Regularizer
  - $\Omega_{sen}(\theta) = \sum_{d=1}^{D} \sum_{s=1}^{S_d} \lambda_{d,s} \|\theta_{d,s}\|_2$

SEMANTIC REGULARIZERS:

- LDA regularizer
- **LSI regularizer**
  - $\Omega_{LDA,LSI}(\theta) = \sum_{k=1}^{K} \lambda \|\theta_k\|_2$

GRAPHICAL REGULARIZERS

- **Graph-of-words regularizer**
  - Community detection on document collection graph
  - $\Omega_{gow}(\theta) = \sum_{c=1}^{C} \lambda \|\theta_c\|_2$
  - c ranges over the C communities.
- **Word2vec regularizer**
  - Kmeans clustering on word2vec
  - $\Omega_{word2vec}(\theta) = \sum_{k=1}^{K} \lambda \|\theta_k\|_2$
  - K is the number of clusters

$\hookrightarrow$ **Why?** Clusters of words will capture same concepts & topics

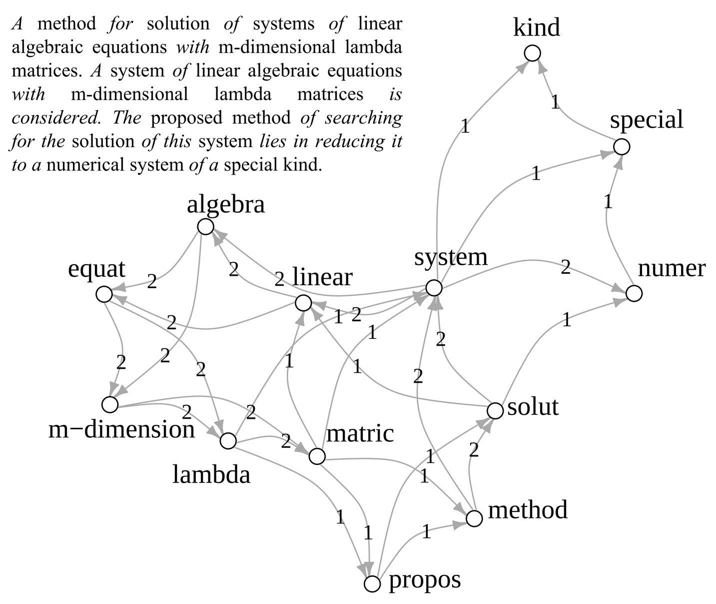

Figure: A Graph-of-words example.

III. Datasets & Setup

DATA

- Topic categorization on 20NG dataset
  - Four binary classification tasks
- Sentiment analysis
  - U.S. Congress floor speeches
  - Movie reviews
  - Amazon product reviews

	dataset	train	dev	test	# words	# sents
20NG	science	949	238	790	25787	16411
	sports	957	240	796	21938	14997
	religion	863	216	717	18822	18853
	comp.	934	234	777	16282	10772
Sentiment	vote	1175	257	860	19813	43563
	movie	1600	200	200	43800	49433
	books	1440	360	200	21545	13806
	dvd	1440	360	200	21086	13794
	electr.	1440	360	200	10961	10227
	kitch.	1440	360	200	9248	8998

Table: Descriptive statistics of the datasets

SETTINGS

- Logistic regression
- 80% for training and 20% for validation with stratified split
- Parameter tuning on development set
- LDA: 1000 topics, 10 most probable words of each topic
- Non-overlapping Louvain community detection for Graph-of-words
- LSI: 1000 latent dimensions, 10 most significant words per topic
- Minibatch K-Means clustering on word2vec with max 2000 clusters
- word2vec: $\forall$ words $\in$ cluster, add the 5 or 10 nearest words

IV. Results

	dataset	no reg.	lasso	ridge	elastic	group lasso LDA	LSI	sentence	GoW	word2vec
20NG	science	0.946	0.916	0.954	0.954	**0.968**	**0.968***	0.942	0.967*	**0.968***
	sports	0.908	0.907	0.925	0.920	0.959	0.964*	**0.966**	0.959*	0.946*
	religion	0.894	0.876	0.895	0.890	0.918	0.907*	**0.934**	0.911*	0.916*
	computer	0.846	0.843	0.869	0.856	0.891	0.885*	0.904	0.885*	**0.911***
Sentiment	vote	0.606	0.643	0.616	0.622	**0.658**	0.653	0.656	0.640	0.651
	movie	0.865	0.860	0.870	0.875	**0.900**	0.895	0.895	0.895	0.890
	books	0.750	0.770	0.760	0.780	0.790	0.795	0.785	0.790	**0.800**
	dvd	0.765	0.735	0.770	0.760	0.800	**0.805***	0.785	0.795*	0.795*
	electr.	0.790	0.800	0.800	**0.825**	0.800	0.815	0.805	0.820	0.815
	kitch.	0.760	0.800	0.775	0.800	0.845	**0.860***	0.855	0.840	0.855*

Table: Bold font marks the best performance. * indicates statistical significance of improvement over lasso at $p < 0.05$ using micro sign test for one of our models LSI, GoW and word2vec (underlined).

	dataset	no reg.	lasso	ridge	elastic	group lasso LDA	LSI	sentence	GoW	word2vec
20NG	science	100	1	100	63	19	20	86	19	21
	sports	100	1	100	5	60	11	6.4	55	44
	religion	100	1	100	3	94	31	99	10	85
	computer	100	2	100	7	40	35	77	38	18
Sentiment	vote	100	1	100	8	15	16	13	97	13
	movie	100	1	100	59	72	81	55	90	62
	books	100	3	100	14	41	74	72	90	99
	dvd	100	2	100	28	64	8	8	58	64
	electr.	100	4	100	6	10	8	43	8	9
	kitch.	100	5	100	79	73	44	27	75	46

Table: Fraction (in %) of non-zero feature weights in each model for each dataset: the smaller, the more compact the model.

V. Discussion & Future Work

- Superior proposed regularizers: more effective, more efficient and sparser
- GoW-based regularization although very fast, did not outperform the other methods
  - Overlapping community detection algorithms failed to identify "good" groups

CONCLUSION

- Find and extract semantic and syntactic structures that lead to sparser feature spaces $\rightarrow$ faster learning times
- Linguistic prior knowledge in the data can be used to improve categorization performance for baseline bag-of-words models, by mining inherent structures
- No significant change in results with different loss functions as the proposed regularizers are not log loss specific

FUTURE WORK

- How to create and cluster graphs, i. e. covering weighted and/or signed cases
- Find better clusters in word2vec (+overlapping with GMM)
- Explore alternative regularization algorithms diverging from group-lasso

	dataset	GoW	word2vec
20NG	science	79	691
	sports	137	630
	religion	35	639
	computer	95	594

Table: Number of groups.

	dataset	lasso	ridge	elastic	group lasso LDA	LSI	sentence	GoW	word2vec
20NG	science	10	1.6	1.6	15	11	76	12	19
	sports	12	3	3	7	20	67	5	9
	religion	12	3	7	10	4	248	6	20
	computer	7	1.4	0.8	8	6	43	5	10

Table: Time (in seconds) for learning with best hyperparameters.

= 0	left-handedness abilities lubin acad sci obesity page erythromycin bottom
≠ 0	space cancer and nasa hiv health shuttle for tobacco that cancer that research center space hiv aids are use theory keyboard data telescope available are from system information space ftp

Table: Examples with LSI regularizer.

= 0	village town points guard guarding crown title champion champions
≠ 0	numbness tingling dizziness fevers laryngitis bronchitis undergo undergoing undergoes undergone healed mankind humanity civilization planet nasa kunin lang tao kay kong

Table: Examples with word2vec regularizer.

kskianis@lix.polytechnique.fr