Regularizing Text Categorization with Clusters of Words
Konstantinos Skiannis, François Rousseau, Michalis Vazirgiannis
LIX, École Polytechnique, France

I. Structured Regularization

Algorithm AUMM
Input: augmented Lagrangian parameter \( \rho \), \( \lambda \) and \( \lambda_{\text{var}} \)
1. while update in weights not small do
2. \( \theta = \text{argmin}_{\theta} \sum_{i} \frac{1}{2} \| d_i - \rho M_i \|_2 \)
3. for \( g = 1 \) to \( G \) do
4. \( v_g = \text{prox}_{\lambda_g} [v_g] \)
5. end for
6. \( u = u + \rho (v - M) \)
7. end while

II. Structured Regularization in NLP

STATISTICAL REGULARIZERS
- Network of features
  \( \Omega(w_i) = \lambda \sum_j \theta_{ij} M_{ij} \), where \( M = \alpha (I - P) (I - P) + \beta I \).
- Sentence Regularizer
  \( \Omega(s_i) = \sum_{w_j \in s_i} \theta_{ij} \theta_{kj} \| \theta_{kl} \|_2 \).

SEMANTIC REGULARIZERS:
- LDA regularizer
  \( \Omega_{\text{LDA}}(s_i) = \sum_{k=1}^K \| \theta_{ik} \|_2 \).
- LSI regularizer
  \( \Omega_{\text{LSI}}(s_i) = \sum_{k=1}^K \| \theta_{ik} \|_2 \).

GRAPHICAL REGULARIZERS:
- Graph-of-words regularizer
  - Community detection on document collection graph
  \( \Omega_{\text{GoW}}(s_i) = \sum_{k=1}^K \| \theta_{ik} \|_2 \).
- Word2vec regularizer
  - K-means clustering on word2vec
  \( \Omega_{\text{word2vec}}(s_i) = \sum_{k=1}^K \| \theta_{ik} \|_2 \).
- \( K \) is the number of clusters

Why? Clusters of words will capture same concepts & topics

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

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: \( \text{v} \) ∈ \( \mathcal{C} \), add the 5 or 10 nearest words

IV. Results

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

<table>
<thead>
<tr>
<th>dataset</th>
<th>GoW</th>
<th>word2vec</th>
</tr>
</thead>
<tbody>
<tr>
<td>science</td>
<td>0.946</td>
<td>0.965</td>
</tr>
<tr>
<td>sports</td>
<td>0.908</td>
<td>0.976</td>
</tr>
<tr>
<td>religion</td>
<td>0.894</td>
<td>0.911</td>
</tr>
<tr>
<td>computer</td>
<td>0.843</td>
<td>0.933</td>
</tr>
</tbody>
</table>

Table: Descriptive statistics of the datasets

<table>
<thead>
<tr>
<th>dataset</th>
<th>GoW</th>
<th>word2vec</th>
</tr>
</thead>
<tbody>
<tr>
<td>movie</td>
<td>0.160</td>
<td>0.200</td>
</tr>
<tr>
<td>books</td>
<td>1.140</td>
<td>1.200</td>
</tr>
<tr>
<td>kitch.</td>
<td>0.040</td>
<td>0.050</td>
</tr>
</tbody>
</table>

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 – 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

Empirical Methods on Natural Language Processing (EMNLP), 2016
ksklani@lix.polytechnique.fr