Linguistic Structured Sparsity in Text Categorization

Dani Yogatama and Noah A. Smith
Language Technologies Institute
Carnegie Mellon University
{dyogatama,nasmith}@cs.cmu.edu
Dani Yogatama
Summary

• Words of a feather (should) flock together

• Idea: use linguistic structure to define *feathers* (flocks) instead of *features*

• Math: sparse group lasso regularization

• Results: text classification (sentiment, forecasting, topic)
this film is one big joke: you have all the basics elements of romance (love at first sight, great passion, etc.) and gangster flicks (brutality, dangerous machinations, the mysterious don, etc.), but it is all done with the crudest humor. it’s the kind of thing you either like visually and immediately “get” or you don’t. that is a matter of taste and expectations.
i enjoyed it and it took me back to the mid80s, when nicolson and turner were in their primes. the acting is very good, if a bit obviously tongue-in-cheek.
this film is one big joke: you have all the basics elements of romance (love at first sight, great passion, etc.) and gangster flicks (brutality, dangerous machinations, the mysterious don, etc.), but it is all done with the crudest humor.
it’s the kind of thing you either like viserally and immediately ”get” or you don’t.
that is a matter of taste and expectations.
i enjoyed it and it took me back to the mid80s, when nicolson and turner were in their primes.
the acting is very good, if a bit obviously tongue-in-cheek.
<table>
<thead>
<tr>
<th></th>
<th>Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>acting</td>
</tr>
<tr>
<td>1</td>
<td>at</td>
</tr>
<tr>
<td>1</td>
<td>back</td>
</tr>
<tr>
<td>1</td>
<td>basics</td>
</tr>
<tr>
<td>1</td>
<td>big</td>
</tr>
<tr>
<td>1</td>
<td>bit</td>
</tr>
<tr>
<td>1</td>
<td>brutality</td>
</tr>
<tr>
<td>1</td>
<td>but</td>
</tr>
<tr>
<td>1</td>
<td>cheek</td>
</tr>
<tr>
<td>1</td>
<td>crudest</td>
</tr>
<tr>
<td>1</td>
<td>dagerous</td>
</tr>
<tr>
<td>6</td>
<td>the</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Linear Classifier

\[
\hat{y} = \text{sign} \left( f(\text{document}) \cdot \mathbf{w} \right)
\]
this film is one big joke: you have all the basics elements of romance (love at first sight, great passion, etc.) and gangster flicks (brutality, dagerous machinations, the mysterious don, etc.), but it is all done with the crudest humor.

it's the kind of thing you either like viserally and immediately ”get” or you don’t.

that is a matter of taste and expectations.
i enjoyed it and it took me back to the mid80s, when nicolson and turner were in their primes.

the acting is very good, if a bit obviously tongue-in-cheek.
Text is Not a Bag of Words!

- Sentences
- Phrases

this film is one big joke: you have all the basics elements of romance (love at first sight, great passion, etc.) and gangster flicks (brutality, dangerous machinations, the mysterious don, etc.), but it is all done with the crudest humor.

It's the kind of thing you either like viserally and immediately "get" or you don't. That is a matter of taste and expectations.

I enjoyed it and it took me back to the mid80s, when nicolson and turner were in their primes. The acting is very good, if a bit obviously tongue-in-cheek.
Text is Not a Bag of Words!

- Sentences
- Phrases
- Fine-grained syntactic classes

this film is one big joke: you have all the basics elements of romance (love at first sight, great passion, etc.) and gangster flicks (brutality, dangerous machinations, the mysterious don, etc.), but it is all done with the crudest humor.

it’s the kind of thing you either like viserally and immediately ”get” or you don’t.

that is a matter of taste and expectations.

i enjoyed it and it took me back to the mid80s, when nicolson and turner were in their primes.

the acting is very good, if a bit obviously tongue-in-cheek.
this film is one big joke: you have all the basics elements of romance (love at first sight, great passion, etc.) and gangster flicks (brutality, dangerous machinations, the mysterious don, etc.), but it is all done with the crudest humor.
it’s the kind of thing you either like viserally and immediately ”get” or you don’t.
that is a matter of taste and expectations.
i enjoyed it and it took me back to the mid80s, when nicolson and turner were in their primes.
the acting is very good, if a bit obviously tongue-in-cheek.
Learning the Weights $w$

“fit the data”
(e.g., log-likelihood of $y_n$ given $d_n$, hinge loss, ...)

$$\hat{w} = \arg \min_w \sum_{n=1}^{N} L(f(d_n), y_n; w) + R(w)$$

“generalize”
(e.g., $\lambda \|w\|_2^2$; $\lambda \|w\|_1$)
Group Lasso (Yuan & Lin ‘06)

\[ R(w) = \sum_{g} \lambda_g \|w_g\|_2 \]

- **w\_1**: no sparsity
- **w\_2**: classic sparsity
- **w\_3**: group sparsity
Group Lasso (Yuan & Lin ‘06)

\[ R(w) = \sum_{g} \lambda_g \|w_g\|_2 \]

In NLP:
• chunking and parsing (Martins et al., 2011)
• language modeling (Nelakanti et al., 2013)
Learning the Weights $w$

\[
\hat{w} = \arg \min_w \sum_{n=1}^{N} L(f(d_n), y_n; w) + R(w)
\]
Learning the Weights $\mathbf{w}$

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w}} \sum_{n=1}^{N} L(f(d_n), y_n; \mathbf{w}) + R(\mathbf{w})$$

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w}} \sum_{n=1}^{N} L(f(d_n), y_n; \mathbf{w})$$

s.t. $R(\mathbf{w}) \leq \tau$

“Tikhonov” regularization

“Ivanov” regularization
Lasso vs. Group Lasso

\[ R(\mathbf{w}) = |w_1| + |w_2| + |w_3| \]

\[ \hat{\mathbf{w}} = \arg \min_{\mathbf{w}} \sum_{n=1}^{N} L(f(d_n), y_n; \mathbf{w}) \]

\[ \text{s.t. } R(\mathbf{w}) \leq \tau \]

Martins et al., EACL 2014 tutorial on structured sparsity in NLP
Lasso vs. Group Lasso

\[ R(w) = |w_1| + |w_2| + |w_3| \]

\[ R(w) = ||\langle w_1, w_2 \rangle||_2 + |w_3| \]

Martins et al., EACL 2014 tutorial on structured sparsity in NLP
Whence Groups?

*Back to NLP ...*
Sentence Regularizer

\[ R(w) = \sum_{n=1}^{N} \sum_{s=1}^{S_n} \lambda_{n,s} \|w_{n,s}\|_2 \]

- Every sentence \(s\) in every document \(n\) gets a group.
- If \(w_{n,s}\) can be driven to zero, that means the sentence is irrelevant to the task.
- Many overlapping groups!

Yogatama and Smith (ICML 2014)
| 1 | acting |
| 1 | at |
| 1 | back |
| 1 | basics |
| 1 | big |
| 1 | bit |
| 1 | brutality |
| 1 | but |
| 1 | cheek |
| 1 | crudest |
| 1 | dangerous |
| 6 | the |

This film is one big joke: you have all the basics elements of romance (love at first sight, great passion, etc.) and gangster flicks (brutality, dangerous machinations, the mysterious don, etc.), but it is all done with the crudest humor.

It's the kind of thing you either like viserally and immediately "get" or you don't. That is a matter of taste and expectations.

I enjoyed it and it took me back to the mid80s, when nicolson and turner were in their primes.

The acting is very good, if a bit obviously tongue-in-cheek.
this film is one big joke: you have all the basics elements of romance (love at first sight, great passion, etc.) and gangster flicks (brutality, dangerous machinations, the mysterious don, etc.), but it is all done with the crudest humor.

it’s the kind of thing you either like viserally and immediately ”get” or you don’t. that is a matter of taste and expectations.

i enjoyed it and it took me back to the mid80s, when nicolson and turner were in their primes.

the acting is very good, if a bit obviously tongue-in-cheek.
More Linguistic Structure Regularizers

- Parse tree regularizer

<table>
<thead>
<tr>
<th>words/features</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>actors</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>are</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>fantastic</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>.</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
</tbody>
</table>
More Linguistic Structure Regularizers

- Parse tree regularizer

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>actors</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>are</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>fantastic</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

- Each of 5,000 hierarchical Brown clusters
More Linguistic Structure Regularizers

- Parse tree regularizer

<table>
<thead>
<tr>
<th>words/features</th>
<th>groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>✓</td>
</tr>
<tr>
<td>actors</td>
<td>✓</td>
</tr>
<tr>
<td>are</td>
<td>✓</td>
</tr>
<tr>
<td>fantastic</td>
<td>✓</td>
</tr>
</tbody>
</table>

- Each of 5,000 hierarchical Brown clusters
- Top ten words in each of 1,000 LDA topics
Sparse Group Lasso

$$\min_{\mathbf{w}} R(\mathbf{w}) + \lambda \| \mathbf{w} \|_1 + \sum_{n=1}^{N} L(f(d_n), y_n; \mathbf{w})$$
Optimization

\[
\min_{\mathbf{w}} R(\mathbf{w}) + \lambda \| \mathbf{w} \|_1 + \sum_{n=1}^{N} L(f(d_n), y_n; \mathbf{w})
\]
Optimization

\[
\min_{\mathbf{w}, \mathbf{v}} R(\mathbf{v}) + \lambda \| \mathbf{w} \|_1 + \sum_{n=1}^{N} L(\mathbf{f}(d_n), y_n; \mathbf{w}) \quad \text{s.t.} \quad \mathbf{v} = M \mathbf{w}
\]

separate \( \mathbf{w} \) from "copies" \( \mathbf{v} \), constraint forces agreement

\[
\min_{\mathbf{w}} R(\mathbf{w}) + \lambda \| \mathbf{w} \|_1 + \sum_{n=1}^{N} L(\mathbf{f}(d_n), y_n; \mathbf{w})
\]
Optimization

\[
\min_{w,v} R(v) + \lambda \|w\|_1 + \sum_{n=1}^{N} L(f(d_n), y_n; w)
\]

s.t. \( v = Mw \)

separate \( w \) from “copies” \( v \),
constraint forces agreement
Optimization

\[
\min_{w,v} R(v) + \lambda \|w\|_1 + \sum_{n=1}^{N} L(f(d_n), y_n; w) \quad \text{s.t.} \quad v = Mw
\]

\[
\min_{w,v} \max_{u} R(v) + \lambda \|w\|_1 + \sum_{n=1}^{N} L(f(d_n), y_n; w) + u \cdot (v - Mw) + \frac{\rho}{2} \|v - Mw\|_2^2
\]

“augmented Lagrangian”
Optimization

\[
\begin{align*}
\min_{w, v} & \quad R(v) + \lambda \|w\|_1 + \sum_{n=1}^{N} L(f(d_n), y_n; w) \\
\text{s.t.} & \quad v = Mw
\end{align*}
\]

ADMM: Alternating Directions

Method of Multipliers

alternating, blockwise updates of \( w \) and \( v \)
a “faster” version of dual ascent for solving the augmented Lagrangian (Hestenes ’69; Powell ’69)

(Glowinski & Marrocco ‘75; Gabay & Mercier ’76)
“Blockwise” Updates

\[ \min_{w} \max_{v, u} R(v) + \lambda \| w \|_1 + \sum_{n=1}^{N} L(f(d_n), y_n; w) + u \cdot (v - Mw) + \frac{\rho}{2} \| v - Mw \|_2^2 \]

w update ≈ loss minimization with elastic net regularization (Zou & Hastie ’05)
"Blockwise" Updates

\[
\min_{w,v} \max_u R(v) + \lambda \|w\|_1 + \sum_{n=1}^{N} L(f(d_n), y_n; w) + u \cdot (v - Mw) + \frac{\rho}{2} \|v - Mw\|_2^2
\]

\text{v updates: proximal operator for each group:}

\[
z_{n,s} = M_{d,s} w - \frac{u_{d,s}}{\rho}
\]

\[
v_{n,s} = \begin{cases} 
0 & \text{if } \|z_{n,s}\|_2 \leq \tau \\
\frac{\|z_{n,s}\|_2 - \tau}{\|z_{n,s}\|_2}z_{n,s} & \text{otherwise}
\end{cases}
\]
“Blockwise” Updates

\[
\min_{w,v} \max_u R(v) + \lambda \|w\|_1 + \sum_{n=1}^{N} L(f(d_n), y_n; w) + u \cdot (v - Mw) + \frac{\rho}{2} \|v - Mw\|_2^2
\]

simple dual update \( u \)
Implications

• Group sparsity and strong sparsity

• Model class is still a (fast) bag of words ... but somehow “informed” by structure

• Learning is more expensive ... but still convex

• A new kind of interpretability ...
this film is one big joke: you have all the basics elements of romance (love at first sight, great passion, etc.) and gangster flicks (brutality, dagerous machinations, the mysterious don, etc.), but it is all done with the crudest humor.

It’s the kind of thing you either like viserally and immediately ”get” or you don’t.

That is a matter of taste and expectations.

I enjoyed it and it took me back to the mid80s, when nicolson and turner were in their primes.

The acting is very good, if a bit obviously tongue-in-cheek.
Classification Experiments

- \( L \): Bag of words logistic regression
- Baselines: m.f.c., lasso, ridge, elastic
- Eight datasets
Sentiment

![Sentiment Analysis Diagram](image)

- **Movies (Socher et al., 2013)**
- **Votes (Thomas et al., 2006)**

- m.f.c.
- lasso
- ridge
- elastic

Baselines:
- sentence
- parse
- Brown
- LDA

* *
20 Newsgroups Binary Tasks

![Bar chart showing performance metrics for various tasks and models.](Image)
Brown as features or regularizer?

![Bar chart showing performance comparisons for Science, Sports, Religion, and Computer categories. The chart includes bars for best baseline, lasso + Brown, ridge + Brown, elastic + Brown, and our Brown regularizer. The y-axis represents performance scores from 50 to 100.]
LDA as features or regularizer?

![Bar chart showing performance of different models across different categories.](chart.png)
Summary

• Words of a feather (should) flock together
• Idea: use linguistic structure to define *feathers* (flocks) instead of features
• Math: sparse group lasso regularization
• Results: text classification (topics, sentiment, forecasting)

Acknowledgments: Google, IARPA, Pittsburgh Supercomputing Center