Making the Most of Bag-of-words: Sentence Regularization with Alternating Direction Method of Multipliers

Jesse Dodge
10-805

Motivation

On Forbes Avenue in Squirrel Hill. Across the street from the Kosher Dunkin Donuts. I like the food here. It has your standard Americanized type fried rice and other Chinese type fare. And it has the good stuff!

Motivation

On Forbes Avenue in Squirrel Hill. Across the street from the Kosher Dunkin Donuts. I like the food here. It has your standard Americanized type fried rice and other Chinese type fare. And it has the good stuff!

Are sentences in orange relevant to predicting whether this is a positive or negative review?

Motivation

On Forbes Avenue in Squirrel Hill. Across the street from the Kosher Dunkin Donuts. I like the food here. It has your standard Americanized type fried rice and other Chinese type fare. And it has the good stuff!

Are sentences in orange relevant to predicting whether this is a positive or negative review?

Very simple linguistic knowledge that a piece of text typically consists of multiple sentences (sentence boundaries)

Contributions

- Introduce a regularizer that exploits the intuition that only some parts of an input are important to the prediction task

- Show an efficient learning algorithm for sparse group lasso with millions of overlapping groups

Outline

- Notation
- Sentence regularizer
- Learning
- Experiments

Notation

- Learning

$$\min_{\mathbf{w}} \overbrace{\boxed{\mathcal{L}(\mathbf{w})}}^{\text{loss function}} + \underbrace{\boxed{\Omega(\mathbf{w})}}_{\text{regularizer}}$$

Notation

- Learning

loss function

$$\min_{\mathbf{w}} \boxed{\mathcal{L}(\mathbf{w})} + \boxed{\Omega(\mathbf{w})}$$

regularizer

loss function: fit the data

regularizer: prevent overfitting and encode prior knowledge

Notation

- Learning

loss function

$$\min_{\mathbf{w}} \boxed{\mathcal{L}(\mathbf{w})} + \boxed{\Omega(\mathbf{w})}$$

regularizer

loss function: fit the data

our focus is on the regularizer

regularizer: prevent overfitting and encode prior knowledge

Structured regularizers

- Structured regularizers promote structural patterns
  - Group lasso (Yuan and Lin, 2006)

$$\boxed{\Omega(\mathbf{w}) = \sum_{g=1}^{G} \|\mathbf{w}_g\|_2}$$

Structured regularizers

- Structured regularizers promote structural patterns

  - Group lasso (Yuan and Lin, 2006)

$$\boxed{\Omega(\mathbf{w}) = \sum_{g=1}^{G} \|\mathbf{w}_g\|_2}$$

Choice of groups: domain knowledge
If prior knowledge is correct, leads to statistical improvements
(Stojnic et al., 2009)

Sentence regularizer

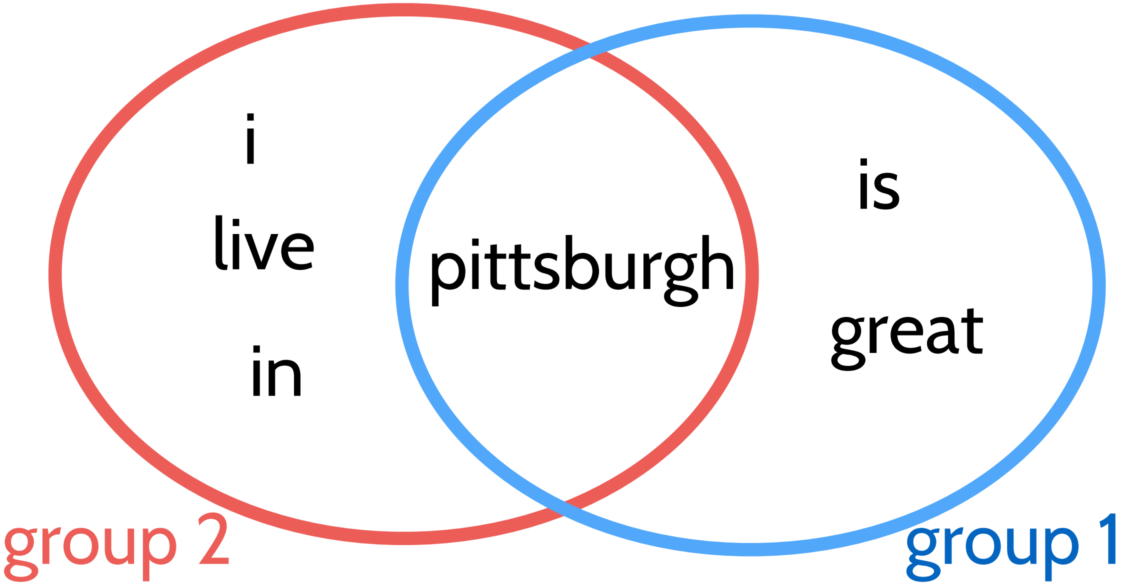
- One group for each sentence in the training corpus
- Intuition: most sentences are irrelevant to the prediction

Sentence regularizer

- One group for each sentence in the training corpus
- Intuition: most sentences are irrelevant to the prediction

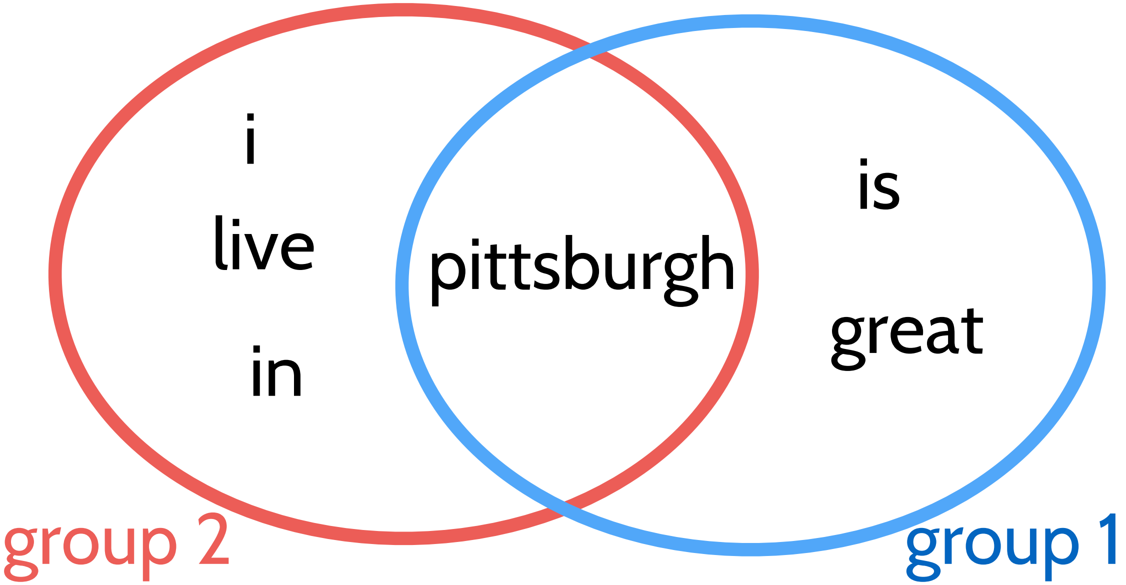
```
i live in pittsburgh . pittsburgh is great .
```

Sentence regularizer



`i live in pittsburgh . pittsburgh is great .`

Sentence regularizer



`i live in pittsburgh . pittsburgh is great .`

Can lead to millions of overlapping groups!

Learning

- Optimization problem in our framework

$$\boxed{\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}) + \Omega_{sen}(\mathbf{w}) + \Omega_{las}(\mathbf{w})}$$

- Note that we couple the sentence regularizer with a classic lasso regularizer (sparse group lasso; Friedman et al., 2010)

Learning

- Optimization problem in our framework

$$\boxed{\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}) + \Omega_{sen}(\mathbf{w}) + \Omega_{las}(\mathbf{w})}$$

- Note that we couple the sentence regularizer with a classic lasso regularizer (sparse group lasso; Friedman et al., 2010)

- How to optimize this efficiently when the structures are millions of overlapping groups (~2.5 million in our experiments)

Learning

$$\boxed{\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}) + \Omega_{sen}(\mathbf{w}) + \Omega_{las}(\mathbf{w})}$$

Learning

$$\boxed{\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}) + \Omega_{sen}(\mathbf{w}) + \Omega_{las}(\mathbf{w})}$$

alternating direction method of multipliers (Hestenes, 1969; Powell, 1969)

Learning

$$\boxed{\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}) + \Omega_{sen}(\mathbf{w}) + \Omega_{las}(\mathbf{w})}$$

rewrite

$$\boxed{\begin{aligned} &\min_{\mathbf{w},\mathbf{v}} \mathcal{L}(\mathbf{w}) + \Omega_{sen}(\mathbf{v}) + \Omega_{las}(\mathbf{w}) \\ &\qquad\qquad\qquad\text{s.t. } \mathbf{v} = \mathbf{M}\mathbf{w} \end{aligned}}$$

Learning

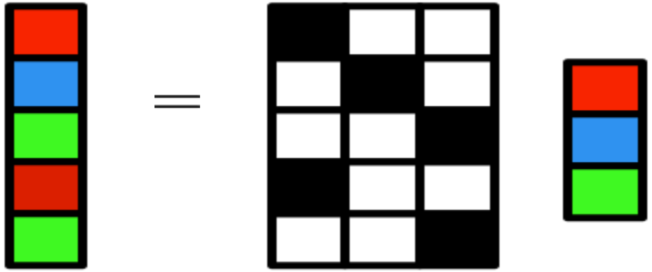
$$\boxed{\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}) + \Omega_{sen}(\mathbf{w}) + \Omega_{las}(\mathbf{w})}$$

rewrite

$$\boxed{\begin{aligned} &\min_{\mathbf{w},\mathbf{v}} \mathcal{L}(\mathbf{w}) + \Omega_{sen}(\mathbf{v}) + \Omega_{las}(\mathbf{w}) \\ &\qquad\qquad\qquad\qquad \text{s.t. } \mathbf{v} = \mathbf{M}\mathbf{w} \end{aligned}}$$

$\boldsymbol{M}$ is a binary matrix that maps elements of $\boldsymbol{v}$ into $\boldsymbol{w}$

2 groups



3 features

Learning

$$\boxed{\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}) + \Omega_{sen}(\mathbf{w}) + \Omega_{las}(\mathbf{w})}$$

rewrite

$$\boxed{\begin{aligned} &\min_{\mathbf{w},\mathbf{v}} \mathcal{L}(\mathbf{w}) + \Omega_{sen}(\mathbf{v}) + \Omega_{las}(\mathbf{w}) \\ &\qquad\qquad\qquad\qquad \text{s.t. } \mathbf{v} = \mathbf{M}\mathbf{w} \end{aligned}}$$

augment

$$\boxed{\begin{aligned} &\min_{\mathbf{w},\mathbf{v}} \mathcal{L}(\mathbf{w}) + \Omega_{sen}(\mathbf{v}) + \Omega_{las}(\mathbf{w}) \\ &\qquad + \mathbf{u}^\top(\mathbf{v} - \mathbf{M}\mathbf{w}) + \frac{\rho}{2}\|\mathbf{v} - \mathbf{M}\mathbf{w}\|_2^2 \end{aligned}}$$

Learning

- Iterate

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

Learning

- Iterate

$\boxed{\text{elastic net like minimization problem}}$

need not be carried out until convergence

Learning

- Iterate

elastic net like minimization problem

need not be carried out until convergence

$$\min_{\mathbf{v}} \Omega_{sen}(\mathbf{v}) + \mathbf{u}^\top \mathbf{v} + \frac{\rho}{2}||\mathbf{v} - \mathbf{M}\mathbf{w}||_2^2$$

Learning

- Iterate

elastic net like minimization problem

need not be carried out until convergence

proximal update

the step that deals with the structured regularizer, can be done in parallel!

Learning

- Iterate

elastic net like minimization problem

need not be carried out until convergence

proximal update

the step that deals with the structured regularizer, can be done in parallel!

$$\boxed{\mathbf{u} = \mathbf{u} + \rho(\mathbf{v} - \mathbf{M}\mathbf{w})}$$

Learning

- Iterate

elastic net like minimization problem

proximal update

dual variable update

Experiments

- Three text classification problems:
  - Topic classification: categorizing documents into two related categories
  - Sentiment analysis: predicting the polarity of a piece of text
  - Text forecasting: predicting a response variable revealed in the future from text

Baselines

- Ridge L2 regularization (Hoerl and Kennard, 1970)
- Lasso L1 regularization (Tibshirani, 1996)
- Elastic net regularization (Zou and Hastie, 2005)

Results

Dataset	m.f.c	lasso	ridge	elastic	sentence
science	50.13	90.63	91.90	91.65	**96.20**
sports	50.13	91.08	93.94	93.71	**95.10**
religion	55.51	90.52	92.47	92.47	**92.75**
computer	50.45	85.84	86.74	87.13	**90.86**

Classification accuracy, higher is better

Model size

Dataset	m.f.c.	lasso	ridge	elastic	sentence
science	-	1	100	34	12
sports	-	2	100	15	3
religion	-	0.3	100	48	94
computer	-	2	100	24	10

Percent of non-zero feature coefficients

Sentence group analysis

task: predicting whether an article is a macintosh or ibm article

blue = "selected" sentences

Sentence group analysis

task: predicting whether an article is a macintosh or ibm article

blue = "selected" sentences

from : *anonymized*

subject : accelerating the macplus ... ;)

lines : 15 we ' re about ready to take a bold step into the 90s around here by accelerating our rather large collection of stock macplus computer .

yes indeed , difficult to comprehend why anyone would want to accelerate a macplus , but that 's another story

suffuce it to say , we can get accelerators easier than new machines

hey , i don ' t make the rules ...

Conclusion

- Introduced a sparse overlapping group lasso regularizer that exploits the intuition that only some parts of an observation are important to the prediction task

- Showed an ADMM algorithm for sparse group lasso with millions of overlapping groups

Thanks!