Sparse Additive Generative Models of Text

Jacob Eisenstein, Amr Ahmed, and Eric P. Xing

Carnegie Mellon University

June 29, 2011
Generative models of text

Generative models are a powerful tool for understanding document collections.

- Classification/clustering (Naive Bayes)
- Discover latent themes (LDA)
- Distinguish latent and observed factors (e.g. Topic-aspect models)
Generative models are a powerful tool for understanding document collections.

- Classification/clustering (Naive Bayes)
- Discover latent themes (LDA)
- Distinguish latent and observed factors (e.g. Topic-aspect models)

**Unifying idea**: each class or latent theme is represented by a distribution over tokens, $P(w|y)$
Redundancy

- A naïve Bayes classifier must estimate the parameter $Pr(w = \text{“the”} | y)$ for every class $y$. 
Redundancy

- A naïve Bayes classifier must estimate the parameter $Pr(w = “the” | y)$ for every class $y$.
- The probability $Pr(w = “the”) is a fact about English, not about any of the classes (usually).
A naïve Bayes classifier must estimate the parameter $Pr(w = \text{“the”} \mid y)$ for every class $y$.

The probability $Pr(w = \text{“the”})$ is a fact about English, not about any of the classes (usually).

Heuristic solutions like stopword pruning are hard to generalize to new domains.
A naïve Bayes classifier must estimate the parameter $Pr(w = "the" | y)$ for every class $y$.

The probability $Pr(w = "the")$ is a fact about English, not about any of the classes (usually).

Heuristic solutions like stopword pruning are hard to generalize to new domains.

It would be better to focus computation on parameters that distinguish the classes.
Overparametrization

- An LDA **model** with $K$ topics and $V$ words requires $K \times V$ parameters.
- An LDA **paper** shows 10 words per topic.
An LDA model with $K$ topics and $V$ words requires $K \times V$ parameters.

An LDA paper shows 10 words per topic.

What about the other $V - 10$ words per topic??
An LDA model with $K$ topics and $V$ words requires $K \times V$ parameters.

An LDA paper shows 10 words per topic.

What about the other $V - 10$ words per topic??

- These parameters affect the assignment of documents...
- But they may be unnoticed by the user.
- And there may not be enough data to estimate them accurately.
Latent topics may be combined with additional facets, such as sentiment and author perspective. “Switching” variables decide if a word is drawn from a topic or from another facet. Twice as many latent variables per document!
Sparse Additive Generative Models

- **Multinomial generative models**: each class or latent theme is represented by a distribution over tokens, \( P(w|y) = \beta_y \)
- **Multinomial generative models**: each class or latent theme is represented by a distribution over tokens, \( P(w|y) = \beta_y \)

- **Sparse Additive Generative models**: each class or latent theme is represented by its deviation from a background distribution.

\[
P(w|y, m) \propto \exp(m + \eta_y)
\]
Sparse Additive Generative Models

- **Multinomial generative models**: each class or latent theme is represented by a distribution over tokens, $P(w|y) = \beta_y$

- **Sparse Additive GEnerative models**: each class or latent theme is represented by its deviation from a background distribution.

\[ P(w|y, m) \propto \exp (m + \eta_y) \]

- $m$ captures the background word log-probabilities
- $\eta$ contains sparse deviations for each topic or class
- additional facets can be added in log-space
Sparse Additive Generative Models

A topic-perspective-background model using Dirichlet-multinomials:
Sparse Additive Generative Models

A topic-perspective-background model using SAGE:
A topic-perspective-background model using SAGE:
Sparsity deviation of log probabilities

- Sparsity: $\eta_i = 0$ for many $i$
Sparsity deviation of log probabilities

- Sparsity: $\eta_i = 0$ for many $i$
- Due to normalization, the generative probabilities will not be identical, $Pr(w = i | \eta + m) \neq Pr(w = i | m)$, even if $\eta_i = 0$. Different notion of sparsity from sparseTM (Wang & Blei, 2009), which sets $Pr(w = i | y) = 0$ for many $i$. 
Sparsity deviation of log probabilities

- Sparsity: $\eta_i = 0$ for many $i$

- Due to normalization, the generative probabilities will not be identical, $Pr(w = i|\eta + m) \neq Pr(w = i|m)$, even if $\eta_i = 0$.

- But for most pairs of words, $\frac{Pr(w=i|\eta+m)}{Pr(w=j|\eta+m)} = \frac{Pr(w=i|m)}{Pr(w=j|m)}$
Sparsity deviation of log probabilities

- Sparsity: $\eta_i = 0$ for many $i$
- Due to normalization, the generative probabilities will not be identical, $Pr(w = i|\eta + m) \neq Pr(w = i|m)$, even if $\eta_i = 0$.
- But for most pairs of words, $\frac{Pr(w=i|\eta+m)}{Pr(w=j|\eta+m)} = \frac{Pr(w=i|m)}{Pr(w=j|m)}$.

Different notion of sparsity from sparseTM (Wang & Blei, 2009), which sets $Pr(w = i|y) = 0$ for many $i$. 
Sparsity deviation of log probabilities

- Sparsity: $\eta_i = 0$ for many $i$
- Due to normalization, the generative probabilities will not be identical, $Pr(w = i | \eta + m) \neq Pr(w = i | m)$, even if $\eta_i = 0$.
- But for most pairs of words, $\frac{Pr(w=i | \eta + m)}{Pr(w=j | \eta + m)} = \frac{Pr(w=i | m)}{Pr(w=j | m)}$

Different notion of sparsity from sparseTM (Wang & Blei, 2009), which sets $Pr(w = i | y) = 0$ for many $i$. 
The Laplace distribution induces sparsity: $\eta \sim \mathcal{L}(0, \sigma)$
The Laplace distribution induces sparsity: $\eta \sim \mathcal{L}(0, \sigma)$

We can apply the integral:

$$\mathcal{L}(\eta; 0, \sigma) = \int \mathcal{N}(\eta; 0, \tau)\exp(\tau; \sigma)d\tau$$  

(Lange & Simsheimer, 1993)
The Laplace distribution induces sparsity: $\eta \sim \mathcal{L}(0, \sigma)$

- We can apply the integral:
  $$\mathcal{L}(\eta; 0, \sigma) = \int \mathcal{N}(\eta; 0, \tau) \text{Exp}(\tau; \sigma) d\tau$$  
  (Lange & Simsheimer, 1993)

- Other integrals also induce sparsity, e.g.
  $$\int \mathcal{N}(\eta; 0, \tau) \frac{1}{\tau} d\tau$$  
  (Figueiredo, 2001; Guan & Dy, 2009)
The Laplace distribution induces sparsity: \( \eta \sim \mathcal{L}(0, \sigma) \)

- We can apply the integral:
  \[
  \mathcal{L}(\eta; 0, \sigma) = \int \mathcal{N}(\eta; 0, \tau) \text{Exp}(\tau; \sigma) d\tau \quad \text{(Lange & Simsheimer, 1993)}
  \]

- Other integrals also induce sparsity, e.g.
  \[
  \int \mathcal{N}(\eta; 0, \tau) \frac{1}{\tau} d\tau \quad \text{(Figueiredo, 2001; Guan & Dy, 2009)}
  \]

- We solve this integral through coordinate ascent, updating:
The Laplace distribution induces sparsity: \( \eta \sim \mathcal{L}(0, \sigma) \)

We can apply the integral:
\[
\mathcal{L}(\eta; 0, \sigma) = \int \mathcal{N}(\eta; 0, \tau) \text{Exp}(\tau; \sigma) d\tau \quad \text{(Lange & Simsheimer, 1993)}
\]

Other integrals also induce sparsity, e.g.
\[
\int \mathcal{N}(\eta; 0, \tau) \frac{1}{\tau} d\tau \quad \text{(Figueiredo, 2001; Guan & Dy, 2009)}
\]

We solve this integral through coordinate ascent, updating:
- The variational distribution \( Q(\tau) \)
The Laplace distribution induces sparsity: $\eta \sim \mathcal{L}(0, \sigma)$

We can apply the integral:

$$\mathcal{L}(\eta; 0, \sigma) = \int \mathcal{N}(\eta; 0, \tau) \exp(\tau; \sigma) d\tau$$  \hspace{1cm} \text{(Lange & Simsheimer, 1993)}

Other integrals also induce sparsity, e.g.

$$\int \mathcal{N}(\eta; 0, \tau) \frac{1}{\tau} d\tau$$  \hspace{1cm} \text{(Figueiredo, 2001; Guan & Dy, 2009)}

We solve this integral through coordinate ascent, updating:

- The variational distribution $Q(\tau)$
- A **point estimate** of $\eta$
Applications

- Document classification
- Topic models
- Multifaceted topic models
Each document $d$ has a label $y_d$

Each token $w_{d,n}$ is drawn from a multinomial distribution $\beta$, where

$$\beta_i = \frac{\exp(\eta_{y_d,i}+m_i)}{\sum_j \exp(\eta_{y_d,j}+m_j)}$$

Each parameter $\eta_{k,i}$ is drawn from a distribution equal to $\mathcal{N}(0, \tau_{k,i})$, with $P(\tau_{k,i}) \sim 1/\tau_{k,i}$
We maximize the variational bound

\[ \ell = \sum_{d} \sum_{n} \log P(w_n^{(d)}|m, \eta_y) + \sum_{k} \langle \log P(\eta_k|0, \tau_k) \rangle + \sum_{k} \langle \log P(\tau_k|\gamma) \rangle - \sum_{k} \langle \log Q(\tau_k) \rangle, \]

where \( c_k \) are the observed counts for class \( k \)

\( C_k = \sum_i c_{ki} \)

\( \beta_k \propto \exp(\eta_k + m) \)
Inference

- We maximize the variational bound

\[
\ell = \sum_d \sum_n \log P(w_n^{(d)}|m, \eta_{y_d}) + \sum_k \langle \log P(\eta_k|0, \tau_k) \rangle \\
+ \sum_k \langle \log P(\tau_k|\gamma) \rangle - \sum_k \langle \log Q(\tau_k) \rangle,
\]

- The gradient wrt \( \eta \) is,

\[
\frac{\partial \ell}{\partial \eta_k} = c_k - C_k \beta_k - \text{diag} \left( \langle \tau_k^{-1} \rangle \right) \eta_k,
\]

where

- \( c_k \) are the observed counts for class \( k \)
- \( C_k = \sum_i c_{ki} \)
- \( \beta_k \propto \exp(\eta_k + m) \)
We maximize the variational bound

\[
\ell = \sum_d \sum_n \log P(w_n^{(d)}|m, \eta_{y_d}) + \sum_k \langle \log P(\eta_k|0, \tau_k) \rangle \\
+ \sum_k \langle \log P(\tau_k|\gamma) \rangle - \sum_k \langle \log Q(\tau_k) \rangle ,
\]
Inference

- We maximize the variational bound

\[
\ell = \sum_d \sum_n \log P(w_n^{(d)} | \mathbf{m}, \eta_y^d) + \sum_k \langle \log P(\eta_k | 0, \tau_k) \rangle
\]
\[
+ \sum_k \langle \log P(\tau_k | \gamma) \rangle - \sum_k \langle \log Q(\tau_k) \rangle ,
\]

- We choose \( Q(\tau_{k,i}) = \text{Gamma}(\tau_{k,i}; a_{k,i}, b_{k,i}) \)
Inference

- We maximize the variational bound

\[
\ell = \sum_d \sum_n \log P(w_n^{(d)} | m, \eta_d^n) + \sum_k \langle \log P(\eta_k | 0, \tau_k) \rangle \\
+ \sum_k \langle \log P(\tau_k | \gamma) \rangle - \sum_k \langle \log Q(\tau_k) \rangle,
\]

- We choose \( Q(\tau_{k,i}) = \text{Gamma}(\tau_{k,i}; a_{k,i}, b_{k,i}) \)

- Iterate between a Newton update to \( a \) and a closed-form update to \( b \)
Document classification evaluation

- 20 newsgroups data: 11K training docs, 50K vocab

![Graph showing the comparison between SAGE and Dirichlet with varying proportion of training set.](image)
Document classification evaluation

- 20 newsgroups data: 11K training docs, 50K vocab

![Graph showing accuracy vs proportion of training set.]

- Adaptive sparsity:
  - 10% non-zeros for full training set (11K docs)
  - 2% non-zeros for minimal training set (550 docs)
The gradient for $\eta_k$ now includes expected counts:

$$\frac{\partial \ell}{\partial \eta_k} = \langle c_k \rangle - \langle C_k \rangle - \beta_k - \text{diag}(\langle \tau_k - 1 \rangle) \eta_k,$$

where $\langle c_{ki} \rangle = \sum_n Q(z_n)(k) \delta(w_n = i)$. 

$i \in \{1, \ldots, W\}$
The gradient for $\eta$ now includes **expected** counts:

$$
\frac{\partial \ell}{\partial \eta_k} = \langle c_k \rangle - \langle C_k \rangle \beta_k - \text{diag} \left( \langle \tau_k^{-1} \rangle \right) \eta_k,
$$

where $\langle c_{ki} \rangle = \sum_n Q_{zn}(k) \delta(w_n = i)$. 

**SAGE in latent variable models**
Sparse topic model results

- NIPS dataset: 1986 training docs, 10K vocabulary

![Graph showing perplexity vs number of topics for Dirichlet-Multinomial and SAGE systems.](image)

- Adaptive sparsity: 5% non-zeros for 10 topics, 1% non-zeros for 50 topics.
Sparse topic model results

- NIPS dataset: 1986 training docs, 10K vocabulary

Adaptive sparsity:
- 5% non-zeros for 10 topics
- 1% non-zeros for 50 topics
Sparse topic model analysis

Total variation $= \sum_i |\beta_{k,i} - \bar{\beta}_i|$

Standard topic models assign the greatest amount of variation for the probabilities of the words with the least evidence!
Multifaceted generative models

- Combines latent topics $\beta^{(T)}$ with other facets $\beta^{(A)}$, e.g. ideology, dialect, sentiment
Multifaceted generative models

- Combines latent topics $\beta^{(T)}$ with other facets $\beta^{(A)}$, e.g. ideology, dialect, sentiment
- Typically, a **switching variable** determines which generative facet produces each token (Paul & Girju, 2010; Ahmed & Xing, 2010).
Multifaceted generative models

- Combines latent topics $\beta^{(T)}$ with other facets $\beta^{(A)}$, e.g. ideology, dialect, sentiment
- Typically, a **switching variable** determines which generative facet produces each token (Paul & Girju, 2010; Ahmed & Xing, 2010).
- There is one switching variable per token, complicating inference.
In SAGE, switching variables are not needed

\[
P(w | z, y) \propto \exp(\eta(T)z + \eta(A)y + m)
\]

The gradient for \(\eta(T)\) is now

\[
\frac{\partial \ell}{\partial \eta(T)k} = c(T)k - \sum_j C_{jk} \beta_{jk} - \text{diag}(\langle \tau - 1 \rangle_k) \eta_k, \theta
\]

\(w, \theta, \alpha \in \{1, \ldots, W\}\)
In SAGE, switching variables are not needed.

Instead, we just sum all the facets in log-space:

\[ P(w|z, y) \propto \exp \left( \eta_z^{(T)} + \eta_y^{(A)} + \mathbf{m} \right) \]
In SAGE, switching variables are not needed. Instead, we just sum all the facets in log-space:

\[ P(w|z, y) \propto \exp \left( \eta^{(T)}_z + \eta^{(A)}_y + m \right) \]

The gradient for \( \eta^{(T)} \) is now

\[
\frac{\partial \ell}{\partial \eta^{(T)}_k} = \langle c^{(T)}_k \rangle - \sum_j \langle C_{jk} \rangle \beta_{jk} - \text{diag} \left( \langle \tau_k^{-1} \rangle \right) \eta_k,
\]
Evaluation: Ideology prediction

- Task: predict blog ideology
- Model: latent topics, observed ideology labels
- Data: six blogs total (two held out), 21K documents, 5.1M tokens

Results match previous best of 69% for Multiview LDA and support vector machine (Ahmed & Xing, 2010).
Evaluation: Ideology prediction

- Task: predict blog ideology
- Model: latent topics, observed ideology labels
- Data: six blogs total (two held out), 21K documents, 5.1M tokens

Results match previous best of 69% for Multiview LDA and support vector machine (Ahmed & Xing, 2010).
Evaluation: Geographical Topic Model

- Task: location prediction from Twitter text
- Model: latent “region” generates text and locations
- 9800 weeklong twitter transcripts; 380K messages; 4.9M tokens
Evaluation: Geographical Topic Model

- Task: location prediction from Twitter text
- Model: latent “region” generates text and locations
- 9800 weeklong twitter transcripts; 380K messages; 4.9M tokens

<table>
<thead>
<tr>
<th>error in kilometers:</th>
<th>median</th>
<th>mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eisenstein et al, 2010 (5K word vocabulary)</td>
<td>494</td>
<td>900</td>
</tr>
<tr>
<td>Wing &amp; Baldridge, 2011 (22K word vocabulary)</td>
<td>479</td>
<td>967</td>
</tr>
</tbody>
</table>
Evaluation: Geographical Topic Model

- Task: location prediction from Twitter text
- Model: latent “region” generates text and locations
- 9800 weeklong twitter transcripts; 380K messages; 4.9M tokens

<table>
<thead>
<tr>
<th>error in kilometers:</th>
<th>median</th>
<th>mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eisenstein et al, 2010 (5K word vocabulary)</td>
<td>494</td>
<td>900</td>
</tr>
<tr>
<td>Wing &amp; Baldridge, 2011 (22K word vocabulary)</td>
<td>479</td>
<td>967</td>
</tr>
<tr>
<td>SAGE (5K)</td>
<td>501</td>
<td>845</td>
</tr>
</tbody>
</table>
Evaluation: Geographical Topic Model

- Task: location prediction from Twitter text
- Model: latent “region” generates text and locations
- 9800 weeklong twitter transcripts; 380K messages; 4.9M tokens

<table>
<thead>
<tr>
<th>error in kilometers:</th>
<th>median</th>
<th>mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eisenstein et al, 2010 (5K word vocabulary)</td>
<td>494</td>
<td>900</td>
</tr>
<tr>
<td>Wing &amp; Baldridge, 2011 (22K word vocabulary)</td>
<td>479</td>
<td>967</td>
</tr>
<tr>
<td>SAGE (5K)</td>
<td>501</td>
<td>845</td>
</tr>
<tr>
<td>SAGE (22K)</td>
<td>461</td>
<td>791</td>
</tr>
</tbody>
</table>
The Dirichlet-multinomial pair is computationally convenient, but does not adequately control model complexity.
Summary

- The Dirichlet-multinomial pair is computationally convenient, but does not adequately control model complexity.
- The Sparse Additive Generative model (SAGE):
  - gracefully handles extraneous parameters,
  - adaptively controls sparsity without a regularization constant,
  - facilitates inference in multifaceted models.
The Dirichlet-multinomial pair is computationally convenient, but does not adequately control model complexity.

The **Sparse Additive** **G**enerative model (SAGE):
- gracefully handles extraneous parameters,
- adaptively controls sparsity without a regularization constant,
- facilitates inference in multifaceted models.

Thanks!
Example Topics

20 Newsgroups, Vocab=20000, K=25

LDA (perplexity = 1131)

- health insurance smokeless tobacco smoked infections care meat
- wolverine punisher hulk mutants spiderman dy timucin bagged marvel
- gaza gazans glocks glock israeli revolver safeties kratz israel
- homosexuality gay homosexual homosexuals promiscuous optilink male
- god turkish armenian armenians gun atheists armenia genocide firearms
Example Topics

20 Newsgroups, Vocab=20000, K=25

LDA (perplexity = 1131)
- health insurance smokeless tobacco smoked infections care meat
- wolverine punisher hulk mutants spiderman dy timucin bagged marvel
- gaza gazans glocks glock israeli revolver safeties kratz israel
- homosexuality gay homosexual homosexuals promiscuous optilink male
- god turkish armenian armenians gun atheists armenia genocide firearms

SAGE (Perplexity = 1090)
- ftp pub anonymous faq directory uk cypherpunks dcr loren
- disease msg patients candida dyer yeast vitamin infection syndrome
- car cars bike bikes miles tires odometer mavenry altcit
- jews israeli arab arabs israel objective morality baerga amehdi hossien
- god jesus christians bible faith atheism christ atheists christianity