Admin

Projects
- Status report due Sunday

Schedule for the rest of the semester
- Monday (4/29): text simplification
- Wednesday (5/1): ethics
  - Post 1-2 papers to read
  - Discussion
- Monday (5/6): 1 hr quiz + presentation info
- Wednesday (5/8): project presentations

Document Modeling

Modeling natural text

You’re goal is to create a probabilistic model of natural (human) text

What are some of the questions you might want to ask about a text?

What are some of the phenomena that occur in natural text that you might need to consider/model?

Questions
what are the key topics in the text?
what is the sentiment of the text?
who/what does the article refer to?
what are the key phrases?

Phenomena
synonym
sarcasm/hyperbole
variety of language (slang), mispellings
coreference (e.g. pronouns like he/she)

Document modeling:
learn a probabilistic model of documents

Model should capture text characteristics

Training a document model

model parameter estimation
document model
Applying a document model

Document model: what is the probability the new document is in the same "set" as the training documents?

Document model applications

Applications

search engines
Google
language generation
speech recognition
text classification and clustering
SPAM
Yahoo!
document hierarchies

Application: text classification

Category
sports
politics
entertainment
business
...
Sentiment
spam
not-spam
positive
negative

I think, therefore I am
Text classification: Training

SPAM

non-SPAM

model parameter estimation

Text classification: Applying

Is it SPAM or non-SPAM?

probability of document being SPAM

which is larger?

probability of document being non-SPAM

Representation and Notation

Standard representation: bag of words

- Fixed vocabulary ~50K words
- Documents represented by a count vector, where each dimension represents the frequency of a word

Clinton said banana repeatedly last week on tv, "banana, banana, banana"

Representation allows us to generalize across documents

Downside?

Representation and Notation

Standard representation: bag of words

- Fixed vocabulary ~50K words
- Documents represented by a count vector, where each dimension represents the frequency of a word

Clinton said banana repeatedly last week on tv, "banana, banana, banana"

Representation allows us to generalize across documents

Downside: lose word ordering information
Word burstiness

What is the probability that a political document contains the word "Clinton" exactly once?

The Stacy Koon-Lawrence Powell defense! The decisions of Janet Reno and Bill Clinton in this affair are essentially the moral equivalents of Stacy Koon’s. …

\[ p(\text{"Clinton"}=1|\text{political}) = 0.12 \]

Word burstiness in models

\[ p(\text{"Clinton"}=1|\text{political}) = \frac{n!}{\prod_{j=1}^{m} \theta_j} \]

Under the multinomial model, how likely is \( p(\text{"Clinton"} = 2 \mid \text{political}) \)?

Word burstiness in models

\[ p(\text{"Clinton"}=2|\text{political}) = 0.05 \]

Many models incorrectly predict:

\[ p(\text{"Clinton"}=2|\text{political}) = (p(\text{"Clinton"}=1|\text{political}))^2 \approx 0.0144 \]

And in general, predict:

\[ p(\text{"Clinton"}=i|\text{political}) = (p(\text{"Clinton"}=1|\text{political}))^i \]
p("Clinton" = x | political)

"Clinton" occurs exactly x times in document

Word count probabilities

common words – 71% of word occurrences and 1% of the vocabulary
average words – 21% of word occurrences and 10% of the vocabulary
rare words – 8% of word occurrences and 89% of the vocabulary

The models...

Multinomial model

20 rolls of a fair, 6-side die - each number is equally probable

(1, 10, 5, 1, 2, 1) vs (3, 3, 3, 4, 4)

Which is more probable?

https://xkcd.com/793/
Multinomial model

20 rolls of a fair, 6-side die - each number is equally probable

(1, 10, 5, 1, 2, 1) vs. (3, 3, 3, 3, 4, 4)

How much more probable?

0.000000764
0.000000891

1000 times more likely

Multinomial model for text

Many more "sides" on the die than 6, but the same concept...

(4, 1, 1, 0, 1, 0, 0, ...) vs. (4, 1, 1, 0, 1, 0, 0, ...)

Generative Story

To apply a model, we’re given a document and we obtain the probability.

We can also ask how a given model would generate a document.

This is the “generative story” for a model.
Multinomial Urn: Drawing words from a multinomial

Selected: $w_1$ $w_3$ $w_2$ $w_1$ $w_1$

Drawing words from a multinomial

Selected: $w_1$

Drawing words from a multinomial

Selected: $w_1$ $w_3$ $w_2$ $w_1$ $w_1$

Put a copy of $w_1$ back

Drawing words from a multinomial

Selected: $w_1$ $w_3$ $w_2$ $w_1$ $w_1$ $w_1$ $w_3$
Drawing words from a multinomial

Selected: \( w_1 \) \( w_2 \) 

Put a copy of \( w_1 \) back

Drawing words from a multinomial

Selected: \( w_1 \) \( w_2 \) \( w_3 \) 

Drawing words from a multinomial

Selected: \( w_1 \) \( w_2 \) \( w_3 \) 

Drawing words from a multinomial

Selected: \( w_1 \) \( w_2 \) \( w_3 \) 

Put a copy of \( w_2 \) back
Does the multinomial model capture burstiness?

$p(\text{word})$ remains constant, independent of which words have already been drawn (in particular, how many of this particular word have been drawn).

Multinomial probability simplex

Generate documents containing 100 words from a multinomial with just 3 possible words:

\[
\begin{align*}
\text{word 1} & \quad \text{word 2} & \quad \text{word 3} \\
(0.31, & \quad 0.44, & \quad 0.25)
\end{align*}
\]

Multinomial word count probabilities
Multinomial does not model burstiness of average and rare words

Better model of burstiness: DCM

Dirichlet Compound Multinomial

Polya Urn process

- **KEY**: Urn distribution changes based on previous words drawn
- Generative story:
  - Repeat until document length hit
    - Randomly draw a word from urn – call it \( w \)
    - Put 2 copies of \( w \) back in urn

Drawing words from a Polya urn

Selected:

Drawing words from a Polya urn

Selected: \( w_1 \)
Drawing words from a Polya urn

Selected: $w_1$

Put 2 copies of $w_1$ back

Adjust parameters

Drawing words from a Polya urn

Selected: $w_1$ $w_3$

Drawing words from a Polya urn

Selected: $w_1$ $w_3$

Put 2 copies of $w_1$ back

Adjust parameters

Drawing words from a Polya urn

Selected: $w_1$ $w_3$

Drawing words from a Polya urn

Selected: $w_1$ $w_3$ $w_2$

Put 2 copies of $w_1$ back

Adjust parameters

Drawing words from a Polya urn

Selected: $w_1$ $w_3$
Drawing words from a Polya urn

Selected: \( w_1 \), \( w_2 \), \( w_2 \)

Put 2 copies of \( w_2 \) back

Words already drawn are more likely to be seen again

Results in the Dirichlet Compound Multinomial (DCM) distribution

Controlled burstiness

Same distribution of words

Which is more bursty?

more bursty

less bursty
Polya urn

Words already drawn are more likely to be seen again.

Results in the DCM distribution.

We can modulate burstiness by increasing/decreasing the number of words in the urn while keeping distribution the same.

Burstiness with DCM

DCM

- Down scaled: \(0.31, 0.44, 0.25\)
- Medium scaled: \(0.32, 1.32, 0.75\)
- Up scaled: \(2.81, 3.94, 2.25\)

DCM word count probabilities

Reminder...

Data

Multinomial

DCM
DCM Model: another view

\[
p(x_1, x_2, \ldots, x_m | \theta_1, \theta_2, \ldots, \theta_m) = \frac{n!}{\prod_{j=1}^{m} x_j!} \prod_{j=1}^{m} \theta_j^{x_j}
\]

\[
p(x_1, x_2, \ldots, x_m | \alpha_1, \alpha_2, \ldots, \alpha_m) = \int_\theta p(x_1 \theta) p(\theta | \alpha) d\theta
\]

Generative story for a single class
- A class is represented by a Dirichlet distribution
- Draw a multinomial based on class distribution
- Draw a document based on the drawn multinomial distribution
Dirichlet Compound Multinomial

\[ p(x|\alpha) = \int \frac{\prod_{w=1}^{W} x_w^{\alpha_w} \prod_{w=1}^{W} \gamma(\alpha_w)}{\prod_{w=1}^{W} \gamma(x_w)} \frac{\prod_{w=1}^{W} \theta_w^{x_w}}{\prod_{w=1}^{W} \gamma(\theta_w)} \frac{d\theta}{\prod_{w=1}^{W} \theta_w^{x_w-1}} \]
Modeling a single class: the fruit bowl

Goal: predict what the fruit mix will be for the following Monday (assign probabilities to options)

Student 1

Student 2

Modeling a single class/group

How well does a model predict unseen data?

Model 1

Model 2

Which model is better?

How would you quantify how much better?

Modeling evaluation: perplexity

Perplexity is the average of the negative log of the model probabilities (likelihood) on test data

Model 1

Model 2

Use the same idea to measure the performance of the different models for modeling one set of documents

Perplexity results

20 newsgroups data set

Multinomial 92.1
DCM 58.7

Lower is better
- ideally the model would have a perplexity of 0!

Significant increase in modeling performance!
Classification results

Accuracy = number correct/ number of documents

<table>
<thead>
<tr>
<th></th>
<th>Industry</th>
<th>20 Newsgroups</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multinomial</td>
<td>0.600</td>
<td>0.853</td>
</tr>
<tr>
<td>DCM</td>
<td>0.806</td>
<td>0.890</td>
</tr>
</tbody>
</table>

(results are on par with state of the art discriminative approaches!)

Next steps in text modeling

Modeling textual phenomena like burstiness in text is important

Better grounded models like DCM ALSO perform better in applications (e.g. classification)

Better models
- text substitutability
- relax bag of words constraint (model co-occurrence)
- handling short phrases (tweets, search queries)
- hierarchical models

Applications of models
- multi-class data modeling (e.g. clustering)
- text similarity
- language generation applications (speech recognition, translation, summarization)