Mixed Membership Markov Models for Unsupervised Conversation Modeling

MICHAEL J. PAUL
JOHNS HOPKINS UNIVERSITY
Conversation Modeling: High Level Idea

- We’ll be modeling sequences of documents
  - e.g. a sequence of email messages from a conversation

- We’ll use $M^4 = \text{Mixed Membership Markov Models}$
- $M^4$ is a combination of:
  - **Topic models** (LDA, PLSA, etc.)
    - Documents are mixtures of latent classes/topics
  - **Hidden Markov models**
    - Documents in a sequence depend on the previous document

M.J. Paul. Mixed Membership Markov Models. EMNLP 2012. Jeju Island, Korea.
Generative Models of Text

- Some distinctions to consider...

| Intra-document structure | Inter-document structure |
|--------------------------|--------------------------|
| **Single-Class**         | **Independent**          |
|                          | **Naïve Bayes**          |
| **Mixed-Membership**     |                          |
|                          | **Markov**               |
|                          | **HMM**                  |
|                          | **LDA**                  |
|                          | **This talk! 😊**         |

M.J. Paul. Mixed Membership Markov Models.

EMNLP 2012. Jeju Island, Korea.
Overview

- Unsupervised Content Models
  - Naïve Bayes
  - Topic Models
- Unsupervised Conversation Modeling
  - Hidden Markov Models
- Mixed Membership Markov Models (M⁴)
  - Overview
  - Inference
- Experiments with Conversation Data
  - Thread reconstruction
  - Speech act induction

M.J. Paul. Mixed Membership Markov Models.

EMNLP 2012. Jeju Island, Korea.
Motivation: Unsupervised Models

- Huge amounts of unstructured and unannotated data on the Web
- Unsupervised models can help manage this data and are robust to variations in language and genre
- Tools like topic models can uncover interesting patterns in large corpora
(Unsupervised) Naïve Bayes

- Each document belongs to some category/class $z$
- Each class $z$ is associated with its own distribution over words
(Unsupervised) Naïve Bayes

football 0.03
team 0.01
hockey 0.01
baseball 0.005

charge 0.02
court 0.02
police 0.015
robbery 0.01

congress 0.02
president 0.02
election 0.015
senate 0.01

probability distributions over words
imaginary class labels

“SPORTS”
“CRIME”
“POLITICS”
(Unsupervised) Naïve Bayes

| football  | 0.03 |
| team      | 0.01 |
| hockey    | 0.01 |
| baseball  | 0.005 |
| charge    | 0.02 |
| court     | 0.02 |
| police    | 0.015 |
| robbery   | 0.01 |
| congress  | 0.02 |
| president | 0.02 |
| election  | 0.015 |
| senate    | 0.01 |

---

**Spanish team honored by fans, royal family in Madrid**

Spain's national team received a joyous welcome at a parade through the streets of Madrid after winning Euro 2012 over Italy. MADRID (AP) -- Swathed in the red-and-yellow colors of Spain, hundreds of thousands packed central Madrid to give a rapturous welcome.

**$21M lawsuit filed in NY police shooting**

WHITE PLAINS, NY - Police in suburban New York responding to a medical alert used excessive force when they killed an emotionally disturbed 68-year-old ex-Marine, the man's son claimed in a $21 million lawsuit Monday.

**Voters encouraged to use 'MyVote' before primary**

Many voters want a quick and easy way to learn more about the candidates they'll see on their primary ballot. Others simply want a fast and convenient way to register to vote or update their registration status in time for the primary.
What if an article belongs to more than one category?
(Unsupervised) Naïve Bayes?

Jury Finds Baseball Star Roger Clemens Not Guilty On All Counts

A jury found baseball star Roger Clemens not guilty on six charges against. Clemens was accused of lying to Congress in 2008 about his use of performance enhancing drugs.
M.J. Paul. Mixed Membership Markov Models.

EMNLP 2012. Jeju Island, Korea.
Topic Models

- One class distribution $\theta_d$ per document
- One class value per token
  - (rather than per document)

T. Hofmann. Probabilistic Latent Semantic Indexing. SIGIR 1999.
Latent Dirichlet Allocation (LDA)

- One class distribution $\theta_d$ per document
- One class value per token (rather than per document)

D. Blei, A. Ng, M. Jordan. Latent Dirichlet Allocation. JMLR 2003.
Overview

- Unsupervised Content Models
- **Unsupervised Conversation Modeling**
- Mixed Membership Markov Models
- Experiments with Conversation Data
- Conclusion
Documents on the web are more complicated than news articles
Documents on the web are more complicated than news articles
What’s missing from Naïve Bayes and LDA?
- They assume documents are generated independently of each other

Messages in conversations aren’t at all independent
- Doesn’t make sense to pretend that they are
- But we’d like to represent this dependence in a reasonably simple way

Solution: Hidden Markov Models
Block HMM

- Message emitted at each time step of Markov chain

- Each message in thread depends on the message to which it is a response
Bayesian Block HMM

- Each message in thread depends on the message to which it is a response

A. Ritter, C. Cherry, B. Dolan. Unsupervised Modeling of Twitter Conversations. HLT-NAACL 2010.

Message 1  Message 2  Message 3
### Block HMM

| GREETING          | SPORTS         | CRIME          | POLITICS       |
|-------------------|----------------|----------------|----------------|
| hey               | football       | charge         | congress       |
| sup               | team           | court          | president      |
| hi                | hockey         | police         | election       |
| hello             | baseball       | robbery        | senator        |
|                   |                |                |                |

| QUESTION          | LAUGHTER       |                |                |
|-------------------|----------------|----------------|----------------|
| what              | lol            |                 | congress       |
| what’s            | haha           |                 | president      |
| how               | :)             |                 | election       |
| is                | lmao           |                 | senator        |
|                   |                |                 |                |

- Hey: 0.1
- Sup: 0.06
- Hi: 0.04
- Hello: 0.01
- ...: ...

- Football: 0.03
- Team: 0.01
- Hockey: 0.01
- Baseball: 0.005
- ...: ...

- Charge: 0.02
- Court: 0.02
- Police: 0.015
- Robbery: 0.01
- ...: ...

- Congress: 0.02
- President: 0.02
- Election: 0.015
- Senate: 0.01
- ...: ...

| QUESTION          | LAUGHTER       |                |                |
|-------------------|----------------|----------------|----------------|
| what              | lol            |                 | congress       |
| what’s            | haha           |                 | president      |
| how               | :)             |                 | election       |
| is                | lmao           |                 | senator        |
|                   |                |                 |                |

- What: 0.03
- What’s: 0.025
- How: 0.02
- Is: 0.02
- ...: ...

- Lol: 0.04
- Haha: 0.04
- :) : 0.03
- Lmao: 0.01
- ...: ...

M.J. Paul. Mixed Membership Markov Models.

EMNLP 2012. Jeju Island, Korea.
Block HMM

- Nice and simple way to model dependencies between messages
  - This is similar to Naïve Bayes
    - One class per document!

- Let’s make it more like LDA
  - Documents are *mixtures* of classes
Generative Models of Text

**Inter-document structure**

| Single-Class          | Independent | Markov |
|-----------------------|-------------|--------|
| **Mixed-Membership**  | ![Diagram](image1) | ![Diagram](image2) |
| **Intra-document structure** | ![Diagram](image3) | ![Diagram](image4) |

This talk! 😊
Overview

- Unsupervised Content Models
- Unsupervised Conversation Modeling
- Mixed Membership Markov Models
- Experiments with Conversation Data
- Conclusion
Mixed Membership Markov Models (M^4)

- Like LDA
  - One distribution $\pi_d$ per doc
  - One class $z$ per token

- But now each message’s distribution depends on the class assignments of previous message

Message 1
Message 2
Message 3

transition parameters

class distribution (function of $z$ and $\lambda$)

M.J. Paul. Mixed Membership Markov Models.
**Core of M⁴:**

Probability of class $j$ in message $d$

$$\pi_{dj} \propto \exp(\lambda_j^T z_{d-1})$$

- Transition parameters
- Class distribution (function of $z$ and $\lambda$)
• Why not transition directly from $\pi$ to $\pi$?

• Makes more sense for next message to depend on actual classes of previous message (not the distribution over all possible classes)
Suppose documents are mixtures of 4 classes: \( \mathbf{Y} \ G \ R \ B \).

Then \( \Lambda \) is a 4x4 matrix with values such as:

\[
\lambda_{G \rightarrow R} = -0.2 \quad \text{“The presence of } G \text{ in doc 1 slightly decreases the likelihood of having } R \text{ in doc 2”}
\]

\[
\lambda_{B \rightarrow B} = 5.0 \quad \text{“The presence of } B \text{ in doc 1 greatly increases the likelihood of having } B \text{ in doc 2”}
\]
Example

- Multinomial parameters $\pi$
- Repeatedly sample $z$ from $\pi$
  - i.e. sample class histogram for doc 1

Counts of $z$:

- $Y$: 0
- $G$: 2
- $R$: 5
- $B$: 2
Example

Counts of $z$:

- $Y$: 0
- $G$: 2
- $R$: 5
- $B$: 2

\[
\begin{align*}
\pi_2^Y & \propto \exp \left( 0 \times \lambda_{Y \rightarrow Y} + 2 \times \lambda_{G \rightarrow Y} + 5 \times \lambda_{R \rightarrow Y} + 2 \times \lambda_{B \rightarrow Y} \right) = \\
\pi_2^G & \propto \exp \left( 0 \times \lambda_{Y \rightarrow G} + 2 \times \lambda_{G \rightarrow G} + 5 \times \lambda_{R \rightarrow G} + 2 \times \lambda_{B \rightarrow G} \right) = \\
\pi_2^R & \propto \exp \left( 0 \times \lambda_{Y \rightarrow R} + 2 \times \lambda_{G \rightarrow R} + 5 \times \lambda_{R \rightarrow R} + 2 \times \lambda_{B \rightarrow R} \right) = \\
\pi_2^B & \propto \exp \left( 0 \times \lambda_{Y \rightarrow B} + 2 \times \lambda_{G \rightarrow B} + 5 \times \lambda_{R \rightarrow B} + 2 \times \lambda_{B \rightarrow B} \right) =
\end{align*}
\]
**Example**

Counts of $z$:

- Doc 1:
  - $Y$: 0
  - $G$: 2
  - $R$: 5
  - $B$: 2

- Doc 2:
  - $Y$: 0
  - $G$: 2
  - $R$: 5
  - $B$: 2

$$\pi_{2Y} \propto \exp(0 \times \lambda_{Y \rightarrow Y} + 2 \times \lambda_{G \rightarrow Y} + 5 \times \lambda_{R \rightarrow Y} + 2 \times \lambda_{B \rightarrow Y}) =$$

$$\pi_{2G} \propto \exp(0 \times \lambda_{Y \rightarrow G} + 2 \times \lambda_{G \rightarrow G} + 5 \times \lambda_{R \rightarrow G} + 2 \times \lambda_{B \rightarrow G}) =$$

$$\pi_{2R} \propto \exp(0 \times \lambda_{Y \rightarrow R} + 2 \times \lambda_{G \rightarrow R} + 5 \times \lambda_{R \rightarrow R} + 2 \times \lambda_{B \rightarrow R}) =$$

$$\pi_{2B} \propto \exp(0 \times \lambda_{Y \rightarrow B} + 2 \times \lambda_{G \rightarrow B} + 5 \times \lambda_{R \rightarrow B} + 2 \times \lambda_{B \rightarrow B}) =$$
Example

Doc 1

\[ \pi \]

\[ \begin{align*}
Y: & 0 \\
G: & 2 \\
R: & 5 \\
B: & 2 \\
\end{align*} \]

\[ \begin{align*}
z_1 & \ z_2 \ z_3 \ z_4 \ z_5 \\
z_6 & \ z_7 \ z_8 \ z_9 \\
\end{align*} \]

Doc 2

\[ \pi \]

\[ \begin{align*}
Y: & 3 \\
G: & 1 \\
R: & 1 \\
B: & 5 \\
\end{align*} \]

\[ \begin{align*}
z_1 & \ z_2 \ z_3 \ z_4 \ z_5 \\
z_6 & \ z_7 \ z_8 \ z_9 \ z_{10} \\
\end{align*} \]

Doc 3

\[ \pi \]

\[ \begin{align*}
Y: & 1 \\
G: & 2 \\
R: & 2 \\
B: & 3 \\
\end{align*} \]

\[ \begin{align*}
z_1 & \ z_2 \ z_3 \ z_4 \ z_5 \\
z_6 & \ z_7 \ z_8 \\
\end{align*} \]
**Mixed Membership Markov Models (M⁴)**

- **M⁴** is a Markov chain where the state space is the set of all possible class histograms
  - If no bound on document length, then the size of this space is countably infinite!
  - But the transition matrix is given in terms of the same number parameters as in a standard HMM
(Approximate) Inference

- Monte Carlo EM
  - E-step: Sample from posterior over class assignments ($z$)
  - M-step: Direct optimization of transition parameters ($\lambda$)

- Inference algorithm alternates between:
  - 1 iteration of collapsed Gibbs sampling
  - 1 iteration (step) of gradient ascent

- Sampler is similar to LDA Gibbs sampler
  - Slower because computing the relative probability of each class involves summing over all classes to compute $\exp(\lambda_j^T z_{d-1})$
Overview

- Unsupervised Content Models
- Unsupervised Conversation Modeling
- Mixed Membership Markov Models
- Experiments with Conversation Data
- Conclusion
Data

- Two sets of asynchronous web conversations
  - CNET forums
    - Technical help and discussion
    - Labeled with speech acts
      - S.N. Kim, L. Wang, T. Baldwin. Tagging and Linking Web Forum Posts. CoNLL 2010.
  - Twitter
    - More personal communication
    - Short messages

| # threads | # messages | # tokens per message |
|-----------|------------|----------------------|
| 321       | 1309       | 78                   |
| 36K       | 100K       | 13                   |

M.J. Paul. Mixed Membership Markov Models.

EMNLP 2012. Jeju Island, Korea.
Experimental Details

- Baselines:
  - Bayesian Block HMM (BHMM)
  - Latent Dirichlet Allocation (LDA)

- Symmetric Dirichlet prior on word distributions
  - Fancy way of describing smoothing
  - Concentration parameter sampled via Metropolis-Hastings

- $\mu$-mean Gaussian prior on transition parameters $\lambda$
  - Independent weights (diagonal covariance)
  - Acts as L2 regularizer on weights

- All Dirichlet hyperparameters are optimized
  - Applies to LDA and BHMM
Pretend we don’t know the thread structure of a conversation. Can we figure out which messages are in response to which?

- Treat “parent” of each message as a hidden variable
  - Sample using simulated annealing
- Evaluate on held-out test data
  - Metric: accuracy (% of messages correctly aligned to parent)
  - Results pooled over many trials

M.J. Paul. Mixed Membership Markov Models. EMNLP 2012. Jeju Island, Korea.
**Thread Reconstruction**

- $M^4$ is a lot better than Block HMM on CNET corpus
  - Twitter messages are short, so single-class assumption is probably reasonable

---

M.J. Paul. Mixed Membership Markov Models. EMNLP 2012. Jeju Island, Korea.
Speech Act Induction

- Messages in CNET corpus are annotated with speech act labels

- 12 labels
  - Question (broken into subclasses)
  - Answer (broken into subclasses)
  - Resolution, Reproduction, Other

- We measured how well the latent classes induced by $M^4$ matched the human labels
  - Metric: variation of information (VI)
M₄ is significantly better.
What Does $M^4$ Learn?

- Top words from a subset of classes
- Arrows show sign of $\lambda$ from going from one class to another
Overview

- Unsupervised Content Models
- Unsupervised Conversation Modeling
- Mixed Membership Markov Models
- Experiments with Conversation Data
- Conclusion
Conclusion

- $\mathbf{M^4}$
  - Combines properties of topic models and Markov models
  - Outperforms LDA and HMM individually

- Room for extensions
  - Richer model of \textit{intra}-message structure
  - Bayesian formulations

- Code is available
  - \texttt{http://cs.jhu.edu/~mpaul}
Acknowledgements

- **Advice:**
  - Mark Dredze
  - Jason Eisner
  - Nick Andrews
  - Matt Gormley
  - Frank Ferraro, Wes Filardo, Adam Teichert, Tim Viera

- **$$$:**

---

M.J. Paul. Mixed Membership Markov Models.

EMNLP 2012. Jeju Island, Korea.
Thank You 감사합니다
### Perplexity

| # classes: | 5 | 10 | 15 | 20 | 25 |
|------------|---|----|----|----|----|
| **CNET**   |    |    |    |    |    |
| Unigram    | 63.1 | 63.1 | 63.1 | 63.1 | 63.1 |
| LDA        | 57.2 | 54.4 | 52.9 | 51.6 | 50.5 |
| BHMM       | 61.3 | 61.1 | 60.9 | 60.9 | 60.9 |
| M⁴         | 60.4 | 59.6 | 59.3 | 59.2 | 59.3 |
| **Twitter**|    |    |    |    |    |
| Unigram    | 93.0 | 93.0 | 93.0 | 93.0 | 93.0 |
| LDA        | 83.7 | 78.4 | 74.0 | 70.9 | 70.2 |
| BHMM       | 90.5 | 89.9 | 89.7 | 89.6 | 89.4 |
| M⁴         | 88.4 | 86.2 | 85.5 | 85.6 | 86.31 |

- **M⁴** more predictive than the block HMM