Text Mining
Statistical Modeling of Textual Data
Lecture 3

Måns Magnusson, Mattias Villani

Division of Statistics
Dept. of Computer and Information Science
Linköping University

Overview

Section 1

Distributional semantics

DISTRIBUTIONAL SEMANTICS RECAP

- The distributional semantics hypothesis

  "a word is characterized by the company it keeps"
  Firth (1957)

- Word meaning comes from **textual context**

  “cold”

  “It’s cold outside.”

  “I’m having a cold”

- Different **contexts** (sentence, word windows, documents)
- Different **context size** - different properties
  - Short distance context, syntagmatic similarities
  - Long distance context, *topical* similarities

Section 2

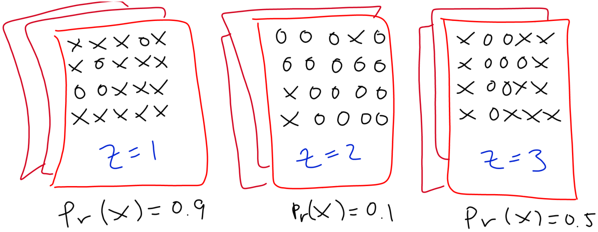
TOPIC MODELS

TOPIC MODELS

- **Problem:** Identifying underlying themes in documents
- Models for **unsupervised learning**, but more recently also for **supervised learning**.
- **Probabilistic generative** unsupervised model.
- The most common topic model is **Latent Dirichlet Allocation (LDA)**.
- **Very popular** model in applications and research. > 20000 Google scholar citations in ~ 15 years.
- **Many extensions** in recent years: n-grams LDA, supervised LDA, nonparametric LDA, relational topics, correlated topics, dynamically time-varying topics...

LDA: Mixture of unigrams

- The basic topic models are extensions of the **bag-of-words** (unigram) model.
- **Mixture of unigrams**:
  1. Draw a *topic indicator* z_d for the dth document from a topic distribution $\theta = (\theta_1, ..., \theta_K)$.
  2. Conditional on the drawn topic z_d draw words w_d from a word distribution for that topic.



- Topic models are **mixed-membership models**:
  each document can belong to **several topics simultaneously**.

Connection to document clustering

- In the Multinomial mixture model **a document** belongs to a cluster

$$p(\mathbf{w}|s) = \prod_{j=1}^{n} p(\mathbf{w}_j|s_j)$$

where $p(\mathbf{w}_j|s_j) \sim MN(\theta_{s_j}, n_j)$

- In Latent Dirichlet Allocation (LDA) **all tokens** belongs to its own cluster

$$p(\mathbf{w}|\mathbf{z}, \Theta, \Phi) = \prod_{j=1}^{D} \prod_{j=1}^{N_i} p(w_j|z_j, \phi_k) p(z_j, \theta_d)$$

where $p(w_j|z_j) \sim Categorical(\phi_{z_j})$ and $p(z_j|\theta_d) \sim Categorical(\theta_d)$

Generating a corpus from a topic model*

- Assume that we have:
  - A fixed vocabulary V
  - D documents
  - N_d words in each document
  - K topics

1. **For each topic** $(k = 1, ..., K)$:
   1.1 Draw a distribution over the words $\phi_k \sim Dir(\beta)$
2. **For each document** $(d = 1, ..., D)$:
   2.1 Draw a vector of topic proportions $\theta_d \sim Dir(\alpha)$
   2.2 **For each word** $(n = 1, ..., N)$:
      2.2.1 Draw a topic assignment $z_{d,n} \sim Multinomial(\theta_d)$
      2.2.2 Draw a word $w_{d,n} \sim Multinomial(\beta_{z_{d,n}})$

Example of $\mathbf{z}$, Θ and Φ

$\mathbf{w}_1$	boat	shore	bank		
$\mathbf{z}_1$	1	1	1		
$\mathbf{w}_2$	Zlatan	boat	shore	money	bank
$\mathbf{z}_2$	2	1	1	3	3
$\mathbf{w}_3$	money	bank	soccer	money	
$\mathbf{z}_3$	3	3	2	3	

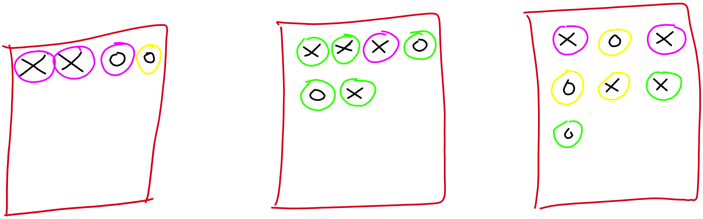
$\Phi =$

	boat	shore	soccer	Zlatan	bank	money
Topic 1	0.35	0.35	0.05	0.05	0.15	0.05
Topic 2	0.025	0.025	0.45	0.45	0.025	0.025
Topic 3	0.025	0.025	0.025	0.025	0.45	0.45

$\Theta =$

	Topic 1	Topic 2	Topic 3
doc 1	0.96	0.02	0.02
doc 2	0.3	0.2	0.5
doc 3	0.05	0.35	0.6

EXAMPLE OF TOPIC MODEL



$\theta_1 = (0.9 \quad 0.1 \quad 0)$

$\theta_2 = (0.1 \quad 0.1 \quad 0.8)$

$\theta_3 = (0.3 \quad 0.4 \quad 0.3)$

$\beta_1 = (0.9 \quad 0.1)$

$\beta_2 = (0.1 \quad 0.9)$

$\beta_3 = (0.5 \quad 0.5)$

EXAMPLE - SIMULATION FROM TWO TOPICS

Topic	Word distr.	probability	dna	gene	data	distribution
1	β_1	0.5	0.1	0.0	0.2	0.2
2	β_2	0.0	0.5	0.4	0.1	0.0

Doc 1	$\theta_1 = (0.2, 0.8)$		
	Word 1:	Topic=2	Word='gene'
	Word 2:	Topic=2	Word='gene'
	Word 3:	Topic=1	Word='data'

Doc 2	$\theta_2 = (0.9, 0.1)$		
	Word 1:	Topic=1	Word='probability'
	Word 2:	Topic=1	Word='data'
	Word 3:	Topic=1	Word='probability'

Doc 3	$\theta_2 = (0.5, 0.5)$

Subsection 1

The linear algebra view of a topic models

CO-OCCURANCE MATRIX

A friend in need is a friend indeed.
She is my friend indeed.

	a	friend	in	indeed	is	my	need	she
Doc 1	2	2	1	1	1	0	1	0
Doc 2	0	1	0	1	1	1	0	1

Co-occurance matrix II

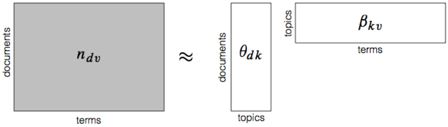
A friend in need is a friend indeed.
She is my friend indeed.

- Context window (of one step)

	a	friend	in	indeed	is	my	need	she
a	2	2	0	0	1	0	0	0
friend	2	3	1	2	0	1	0	0
in	0	1	1	0	0	0	1	0
indeed	0	2	0	2	0	0	0	0
is	1	0	0	0	2	1	1	1
my	0	1	0	0	1	1	0	0
need	0	0	1	0	1	0	1	0
she	0	0	0	0	1	0	0	1

TOPIC MODELS - THE LINEAR ALGEBRA VIEW

- Reduce co-occurance matrix to a **lower dimension**
  - The linear algebra view



FIGUR: Matrix decomposition (taken from talk by David Blei at NIPS 2013)

- A kind of probabilistic Non-Negative Matrix factorization
- n_{dv} often very large - how to do this efficiently (and without creating n_{dv})
- Can be done with SVD $\rightarrow$ Latent Semantic Analysis

Section 3

Inference in the LDA model

LEARNING/INFERENCE IN TOPIC MODELS

- What do we know?
  - The words in the documents: $\mathbf{w}$
- What do we not know?
  - Topic proportions for each document: θ_d
  - Topic assignments for each word in each document: $\mathbf{z}$
  - Word distributions for each topic: ϕ_k
- Do the **Bayes dance**: Posterior distribution

$$p(\Theta, \mathbf{z}, \Phi | \mathbf{w})$$

- The posterior is mathematically untractable. **Solutions**:
  - Gibbs sampling (MCMC) [Correct, but can be slow]
  - Variational Bayes EM [Crude approximation of the posterior *distribution*, but typically rather accurate about posterior mode (MAP)]

GIBBS SAMPLER FOR LDA I

Bayes theorem

$$p(B|A) = \frac{p(A|B) \cdot p(B)}{p(B)}$$

For the topic model

$$\begin{aligned} p(\mathbf{z}, \Theta, \Phi|\mathbf{w}) &= \frac{p(\mathbf{z}, \Theta, \Phi|\mathbf{w}) \cdot p(\mathbf{z}, \Theta, \Phi)}{p(\mathbf{w})} \\ &\propto p(\mathbf{z}, \Theta, \Phi|\mathbf{w}) \cdot p(\mathbf{z}, \Theta, \Phi) \end{aligned}$$

GIBBS SAMPLER FOR LDA II

The basic Gibbs sampler:

$$p(z = k|\Phi, \Theta) \propto \phi_{v,k}\theta_{k,d}$$

$$\theta_d|\mathbf{z} \sim Dir(\mathbf{n}^{(d)} + \alpha)$$

$$\phi_k|\mathbf{z} \sim Dir(\mathbf{n}^{(v)} + \beta)$$

where $\mathbf{n}_d$ is the number of tokens by topic in document d and $\mathbf{n}_v$ is the number of tokens by topic for word type $\mathbf{z}$.

GIBBS SAMPLER FOR LDA III

Integrating out (collapsing) Θ and Φ (Griffiths and Steyvers (2004)):

$$p(\mathbf{z}|\mathbf{w}) \quad = \quad \int\int p(\mathbf{z},\Theta,\Phi|\mathbf{w}) \cdot p(\mathbf{z},\Theta,\Phi) d\Phi d\Theta$$

will result in the following gibbs sampler

$$p(z_i = k|w_i, \mathbf{z}_{\neg i}) \propto \underbrace{\frac{n_k^{(v)} + \beta}{n_k^{(v)} + V\beta}}_{type-topic\ (\Phi)} \cdot \underbrace{(n_k^{(d)} + \alpha)}_{topic-doc\ (\Theta)}$$

where $n^{(v)}$ and $n^{(d)}$ are count matrices of size $D \times K$ and $K \times V$.

Example of $n^{(v)}$ and $n^{(d)}$

$\mathbf{w}_1$	boat	shore	bank		
$\mathbf{z}_1$	1	1	1		
$\mathbf{w}_2$	Zlatan	boat	shore	money	bank
$\mathbf{z}_2$	2	1	1	3	3
$\mathbf{w}_3$	money	bank	soccer	money	
$\mathbf{z}_3$	3	3	2	3	

$$n^{(v)} = \begin{array}{cccccc} \text{boat} & \text{shore} & \text{soccer} & \text{Zlatan} & \text{bank} & \text{money} \\ 2 & 2 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 2 \end{array}$$

$$n^{(d)} = \begin{bmatrix} 3 & 0 & 0 \\ 2 & 1 & 3 \\ 0 & 2 & 3 \end{bmatrix}$$

COLLAPSED GIBBS SAMPLER FOR LDA

```
# Initialization
Sample all topic indicators randomly
Calculate n^(w) and n^(d)
# Gibbs sampler
for each gibbs iteration do
    for each token w_i do
        remove z_i from n^(v) and n^(d)
        for each k in 1 to K do
            prob_k[k] = (n_k^(v) + β) / (n_k^(w) + Vβ) · (n_k^(d) + α)
        end for
        z_i <- draw multinomial(prob_k)
        add z_i to n^(v) and n^(d)
    end for
end for
return n^(w), n^(d)
```

$$\texttt{prob\_k[k]} = \frac{n_k^{(v)}+\beta}{n_k^{(w)}+V\beta} \cdot \left(n_k^{(d)} + \alpha\right)$$

(NAIVE) COLLAPSED GIBBS SAMPLER ALGORITHM II

- Estimation of Φ and Θ

$$\begin{aligned}\hat{\phi}_{k,v} &= \frac{n_k^{(v)} + \beta}{n_k^{(v)} + V\beta} \\ \hat{\theta}_{d,k} &= \frac{n_d^{(d)} + \alpha}{n_d^{(d)} + K\alpha}\end{aligned}$$

- Sort $\hat{\phi}_{k,v}$ by largest values.

- Computational complexity is $O(K)$ for each token
- Slow for larger corpuses... as you will see...

Section 4

EVALUATION

Evaluation of topic models

- Convergence:
  - Log marginal posterior $p(\mathbf{w}|\mathbf{z}^{(n)})$
- Evaluating and comparing models:
  - Held-out marginal likelihood (Wallach et al. (2009b))

$$p(\hat{\mathbf{w}}) = \sum^{\mathbf{z}} p(\mathbf{w}|\mathbf{z})$$

  - NPMI, topic distributions (junk topics)
  - Topic coherence

$$C(t, V^{(t)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D(v_m^{(t)}, v_l^{(t)}) + 1}{D(v_l^{(t)})}$$

EVALUATION OF TOPIC MODELS

Table 1: Example topics (good/general, good/research, chained/research) with different coherence scores (numbers closer to zero indicate higher coherence). The chained topic combines words related to aging (indicated in plain text) and words describing blood and blood-related diseases (bold). The only connection is the common word *human*.

-167.1	students, program, summer, biomedical, training, experience, undergraduate, career, minority, student, careers, underrepresented, medical_students, week, science
-252.1	neurons, neuronal, brain, axon, neuron, guidance, nervous_system, cns, axons, neural, axonal, cortical, survival, disorders, motor
-357.2	aging, lifespan, **globin**, age_related, longevity, human, age, **erythroid**, **sickle_cell**, **beta_globin**, **hb**, senescence, adult, older, lcr

FIGUR: Example of topic coherence. (Mimno et. al. 2011)

Section 5

PRACTICAL DECISIONS

HYPER PARAMETERS

- α and β defines sparsity in Θ and Φ respectively
- Can be learned from data with hyperparameter optimization (Wallach et al., 2009a)
- Common values are $\alpha = 0.1$ and $\beta = 0.01$ or $1/K$ - remeber the Dirichlet prior
- K - think of it as resolution or the corpus

Defining tokens

- What is a token?
- "Moderata samlingspartiet"
- **Recommendation:** Start without combining collocations - add collocations later on that is of importance (anecdotal)

Stop words

- Common words with less (?) thematic/semantic meaning
- Big part of corpus ($\sim 50\%$)
- **Recommendation:** Remove a small set of very common words (Xanda Shoefield and Mimno, 2017)

Rare words

- Thematic words (will mainly be affected by the context (documents))
- Large part of Φ
- **Recommendation:** Remove very rare words...

STEMMING AND LOWERCASING

- Reduce the vocabulary
- Stemming: Reduce each token to its (word) stem "running" $\rightarrow$ "run"
- **Recommendation:** Do not stem, unless really small corpora (Schofield and Mimno, 2016)
- **Recommendation:** Reduce to lower case (anecdotal)

"Junk" topics

- Some topics just capture language structure AlSumait et al. (2009)
- Some topics combine two topics with a few common words.
- Often called junk topics
- **Recommendation:** Ignore these topics (anecdotal)

Document segmentation

- Segment documents to smaller pieces

Master thesis proposal - Polylingual topic model

- Storytel master thesis project:
  - Match books.
  - Combine with character names (Named Entity Recognition).
- Polyalingual topic model for multiple languages Mimno et al. (2009)

MASTER THESIS PROPOSAL - POLYLINGUAL TOPIC MODEL

1. **For each topic** ($k = 1, ..., K$):
   1.1 **For each language** ($l = 1, ..., L$)
      1.1.1 Draw a distribution over the words $\phi_{k,l} \sim Dir(\beta)$
2. **For each document** ($d = 1, ..., D$):
   2.1 Draw a vector of topic proportions $\theta_d \sim Dir(\alpha)$
   2.2 **For each language** ($l = 1, ..., L$):
      2.2.1 **For each word** ($n_l = 1, ..., N_L$):
      2.2.2 Draw a topic assignment $z_{d,n} \sim Multinomial(\theta_d)$
      2.2.3 Draw a word $w_{d,l,n} \sim Multinomial(\phi_{z_{d,n},l})$

References

AlSumait, L., Barbará, D., Gentle, J., Domeniconi, C., 2009. Topic significance ranking of lda generative models. Machine Learning and Knowledge Discovery in Databases, 67–82.

Firth, J., 1957. A synopsis of linguistic theory 1930-1955. Studies in linguistic analysis, 1–32.

Griffiths, T., Steyvers, M., 2004. Finding scientific topics. ...academy of Sciences of the United ....
URL `http://www.pnas.org/content/101/suppl.1/5228.short`

Mimno, D., Wallach, H. M., Naradowsky, J., Smith, D. A., McCallum, A., 2009. Polylingual topic models. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2. Association for Computational Linguistics, pp. 880–889.

Schofield, A., Mimno, D., 2016. Comparing apples to apple: The effects of stemmers on topic models. Transactions of the Association for Computational Linguistics 4, 287–300.

Wallach, H. M., Mimno, D. M., McCallum, A., 2009a. Rethinking lda: Why priors matter. In: Advances in neural information processing systems. pp.