Introduction to Bayesian Methods for Text Analysis

Tim Van de Cruys

University of Groningen

CL reading group
March 26, 2010

Frequentist vs. Bayesian 1/2

- Frequentist view
  - Probability is the long-term expected frequency of an occurrence
  - There is a real (unknown) population mean that can be estimated from the data
  - Parameters are fixed
- Bayesian view
  - Probability is a degree of belief
  - The population mean is an abstraction based on the data and prior beliefs
  - Parameters are described probabilistically

Frequentist vs. Bayesian 2/2

- Frequentist view
  - Implicit perspective of many machine learning methods
  - support vector machines, decision trees, neural networks, LSA
- Bayesian view
  - Bayesian graphical models
  - Latent Dirichlet Allocation

Overview

Preliminaries

- data set $\mathcal{X} = \{x_i\}_{i=1}^{|\mathcal{X}|}$
  = sequence of independent and identically distributed (i.i.d.) realizations of random variable X, with ϑ being the parameters of the distribution
- Bayes' rule

$$p(\vartheta|\mathcal{X}) = \frac{p(\mathcal{X}|\vartheta) \cdot p(\vartheta)}{p(\mathcal{X})} \tag{1}$$

$$posterior = \frac{likelihood \cdot prior}{evidence} \tag{2}$$

Maximum likelihood estimation (MLE)

- Frequentist approach
- Find parameters that maximize the likelihood

$$L(\vartheta|\mathcal{X}) = p(\mathcal{X}|\vartheta) = \prod_{x \in \mathcal{X}} p(x|\vartheta) \tag{3}$$

$$\hat{\vartheta}_{ML} = \arg\max_{\vartheta} \mathcal{L}(\vartheta|\mathcal{X}) = \arg\max_{\vartheta} \sum_{x \in \mathcal{X}} \log p(x|\vartheta) \tag{4}$$

Maximum a posteriori estimation (MAP)

- Similar to MLE
- Allows to include prior belief on parameters by weighting with a prior distribution

$$\begin{aligned} \hat{\vartheta}_{ML} &= \arg\max_{\vartheta} p(\vartheta|\mathcal{X}) \\ &= \arg\max_{\vartheta} \frac{p(\mathcal{X}|\vartheta)p(\vartheta)}{p(\mathcal{X})} \\ &= \arg\max_{\vartheta} \{\sum_{x \in \mathcal{X}} \log p(x|\vartheta) + \log p(\vartheta)\} \end{aligned} \tag{5}$$

Bayesian inference

- Extension of MAP approach by allowing distribution over parameters ϑ (no direct estimate)
- Expectation and variance as measures of estimation quality
- Not an approximation (best possible value, like MLE and MAP) but a complete probability distribution

Introduction

- Treat data as observations that arise from a generative probabilistic process that includes hidden variables (hidden variables reflect thematic structure of the collection)
- Infer the hidden structure (topics) using posterior inference

Vlaamse woede over 'racistisch' Waals artikel

AMSTERDAM - Een Vlaamse politicus heeft de Franstalige Belgische krant Le Soir aangeklaagd wegens het aanzetten tot racisme en haat.

Bart De Wever, partijvoorzitter van de Vlaams-nationalistische partij N-VA, reageert daarmee op een opiniestuk in Le Soir, waarin een nieuwe Vlaamse wet in verband wordt gebracht met etnische zuiveringen.

In het gewraakte opiniestuk neemt Le Soir-columnist Jean-Paul Marthoz stelling tegen het Vlaamse decreet over wonen in eigen streek.

Vlaamse woede over 'racistisch' Waals artikel

AMSTERDAM - Een Vlaamse politicus heeft de Franstalige Belgische krant Le Soir aangeklaagd wegens het aanzetten tot racisme en haat.

Bart De Wever, partijvoorzitter van de Vlaams-nationalistische partij N-VA, reageert daarmee op een opiniestuk in Le Soir, waarin een nieuwe Vlaamse wet in verband wordt gebracht met etnische zuiveringen.

In het gewraakte opiniestuk neemt Le Soir-columnist Jean-Paul Marthoz stelling tegen het Vlaamse decreet over wonen in eigen streek.

$\longrightarrow$ BELGIUM

Vlaamse woede over 'racistisch' Waals artikel

AMSTERDAM - Een Vlaamse politicus heeft de Franstalige Belgische krant Le Soir aangeklaagd wegens het aanzetten tot racisme en haat.

Bart De Wever, partijvoorzitter van de Vlaams-nationalistische partij N-VA, reageert daarmee op een opiniestuk in Le Soir, waarin een nieuwe Vlaamse wet in verband wordt gebracht met etnische zuiveringen.

In het gewraakte opiniestuk neemt Le Soir-columnist Jean-Paul Marthoz stelling tegen het Vlaamse decreet over wonen in eigen streek.

$\longrightarrow$ POLITICS

Vlaamse woede over 'racistisch' Waals artikel
AMSTERDAM - Een Vlaamse politicus heeft de Franstalige Belgische krant Le Soir aangeklaagd wegens het aanzetten tot racisme en haat.
Bart De Wever, partijvoorzitter van de Vlaams-nationalistische partij N-VA, reageert daarmee op een opiniestuk in Le Soir, waarin een nieuwe Vlaamse wet in verband wordt gebracht met etnische zuiveringen.
In het gewraakte opiniestuk neemt Le Soir-columnist Jean-Paul Marthoz stelling tegen het Vlaamse decreet over wonen in eigen streek.

→ PRESS

Vlaamse woede over 'racistisch' Waals artikel

AMSTERDAM - Een Vlaamse politicus heeft de Franstalige Belgische krant Le Soir aangeklaagd wegens het aanzetten tot racisme en haat.

Bart De Wever, partijvoorzitter van de Vlaams-nationalistische partij N-VA, reageert daarmee op een opiniestuk in Le Soir, waarin een nieuwe Vlaamse wet in verband wordt gebracht met etnische zuiveringen.

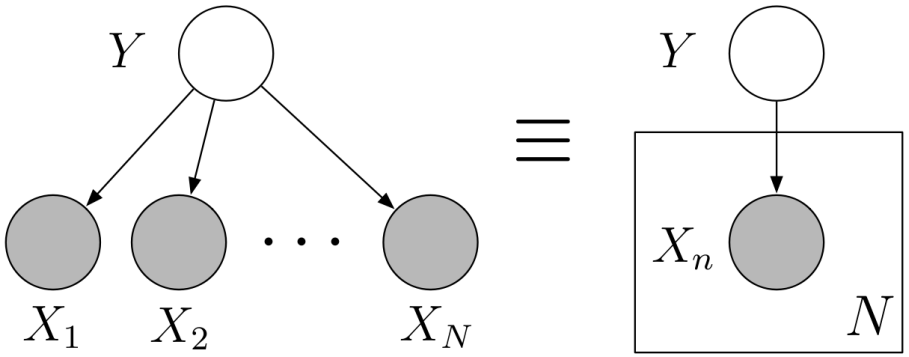
In het gewraakte opiniestuk neemt Le Soir-columnist Jean-Paul Marthoz stelling tegen het Vlaamse decreet over wonen in eigen streek.

→ NETHERLANDS, JUSTICE, HOUSING, ...

Intuition

- Each document is a mixture of corpus-wide topics
- Each word is drawn from one of those topics
- Only the documents are observed
- Goal is to try to infer the underlying topic structure

Graphical model

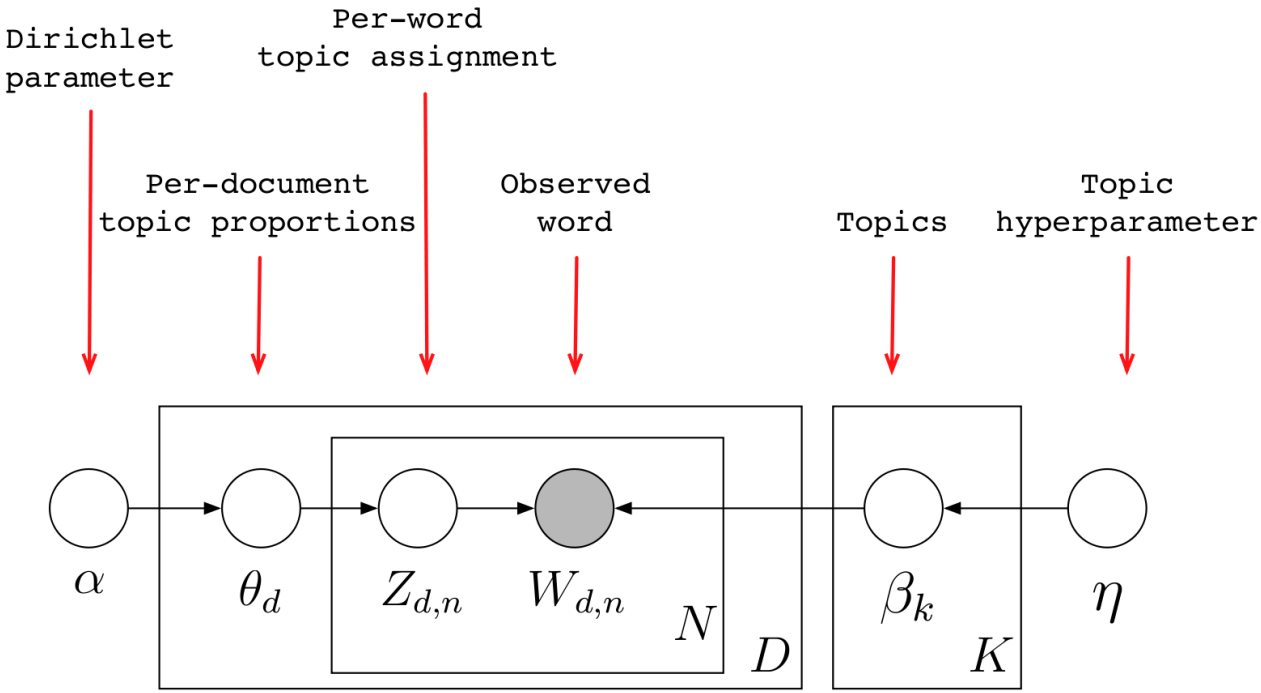


- Nodes are random variables
- Edges denote conditional dependencies
- Observed variables are shaded
- Plates denote replicated structure

Dirichlet distribution

- Formally:
  - **conjugate** to the multinomial distribution: given a multinomial observation, posterior distribution is also a Dirichlet distribution
- Practically:
  - Used to control sparsity of multinomial observations

LDA: graphical model



LDA: generative model

- For each topic $k \in [1, K]$:
  - sample mixture components $\overrightarrow{\beta_k} \sim Dir(\overrightarrow{\eta})$
    ($\rightarrow$ *topic* $\times$ *word* matrix indicating $p(w|z)$)
- For each document $d \in [1, D]$:
  - sample mixture proportion $\overrightarrow{\theta_d} \sim Dir(\overrightarrow{\alpha})$
    ($\rightarrow$ *document* $\times$ *topic* matrix indicating $p(z|d)$)
  - (sample document length $N_d \sim Poiss(\xi)$)
  - for each word $n \in [1, N_d]$ in document d:
    - sample topic index $z_{d,n} \sim Mult(\overrightarrow{\theta_d})$
    - sample word $w_{d,n} \sim Mult(\overrightarrow{\beta_{z_{d,n}}})$

How does it work?

- Word probabilities are maximized by dividing words among topics
- $\longrightarrow$ co-occurring words are found
- Dirichlet on topic proportions is used to encourage sparsity: document is penalized for using many topics
- Leads to sets of term that tightly co-occur
- Algorithms: variational inference or Gibbs sampling

Methodology

- LDA inference (Gibbs sampling, 200 iterations) on part of TWNC (60k documents)
- Filter stop words
- number of topics K = 50

Results

- onderzoek student jaar universiteit doe commissie krijg studie ga hoogleraar
- krant schrijf vraag Volkskrant journalist medium artikel doe blad lees
- Amerikaans VS Amerikaan Irak Verenigde_Staten Amerika wereld Washington dollar Bush
- politie rechter advocaat justitie onderzoek rechtbank verdachte Justitie slachtoffer straf
- eet water vis doe smaak kook bak restaurant wijn drink
- boek schrijf schrijver verhaal lees fl roman auteur pagina verschijn

Conclusion

- LDA is a rigorously Bayesian framework that yields state-of-the-art results
- Modular, can be easily extended (model to automatically find number of topics, incorporate syntax, . . .)
- Computationally rather heavy, especially compared to other 'latent semantic models' (latent semantic analysis, non-negative matrix factorization)