Latent Dirichlet Allocation

BAINSA Research Presentation

Matei-Gabriel Cosa & Kassym Mukhanbetiyar

September 29, 2023

Outline

- Topic modelling
- Intuition
- Notation
- Preliminaries
- Inference (EM)
- Conclusion (Q&A)

Topic modelling

- **NLP problem**: extract topics from a set of documents without prior knowledge of what the topics might be $\Rightarrow$ **unsupervised task**
- **Use cases**: text summarization, content analysis, document clustering, recommender systems
- **Models**: latent Dirichlet allocation, latent semantic analysis, word embedding models (Word2Vec, Doc2Vec, etc.), LLM's

Intuition

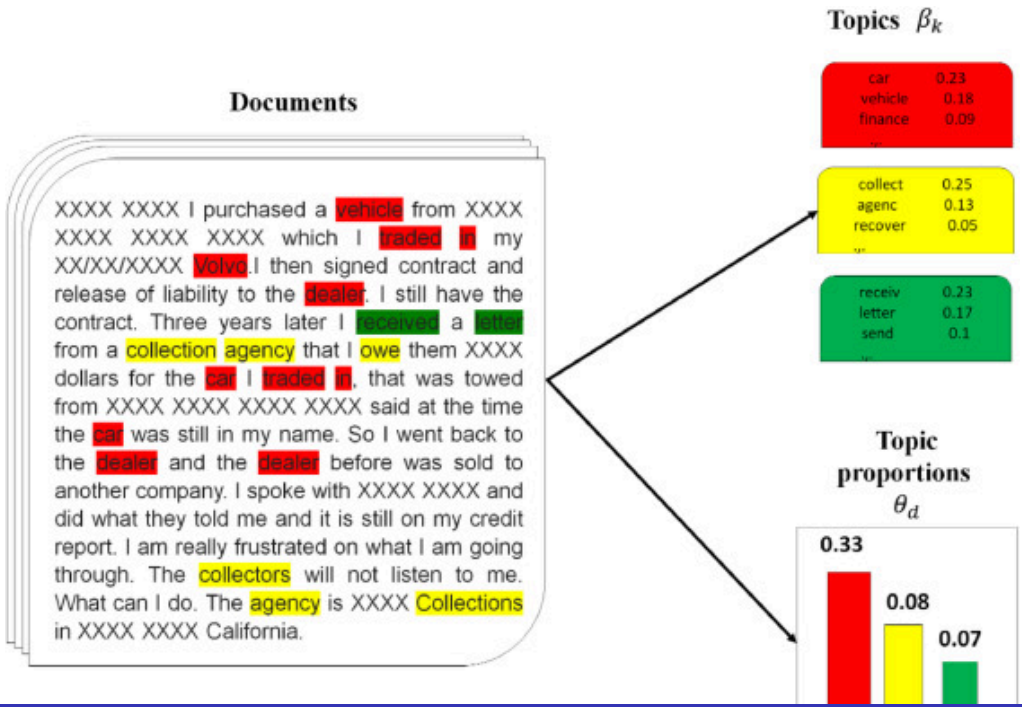
Bag of words

We disregard grammar and word order, but keep track of the number of occurrences of each word ⇒ **dictionary-like structure**.

	the	red	dog	cat	eats	food
1. the red dog →	1	1	1	0	0	0
2. cat eats dog →	0	0	1	1	1	0
3. dog eats food →	0	0	1	0	1	1
4. red cat eats →	0	1	0	1	1	0

Documents, topics, words, ...

We assume each document is a mixture of topics and each topic is a mixture of words. Therefore, we can view topics as a probability distribution over the bag of words and documents as a probability distribution over the topics $\Rightarrow$ **several layers**.



An example

A cat and a dog fly through a galaxy of stars.
A cat wants to fly towards a mouse.

Topic level:

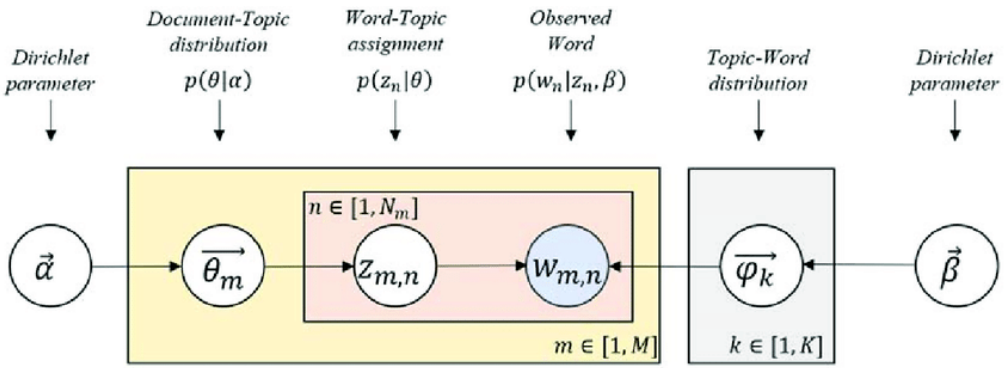
	cat	dog	fly	galaxy	stars	mouse
topic 1 (animals?)	0.5	0.1	0.1	0.0	0.0	0.3
topic 2 (space?)	0.0	0.0	0.3	0.4	0.3	0.0

Document level:

	Probability
topic 1 (animals?)	0.6
topic 2 (space?)	0.4

The big picture

We have a corpus comprised of several documents, each with its own distribution of topics, with each topic having a distribution of words $\Rightarrow$ **hierarchical Bayesian model**.



In Bayesian Statistics there isn't a "true" underlying distribution over the r.v.. We update our beliefs as we see more data.
Example.

- 1 in 1000 people are ill.
- θ - 0 if the patient is healthy, 1 if ill.
- X - test for disease, 0 if negative, 1 if positive.
- $P[X = 1|\theta = 1] = P[X = 0|\theta = 0] = 0.95$

Quantity of interest:

$$P[\theta = 1|X = 1] = \frac{P[X = 1, \theta = 1]}{P[X = 1]}$$

$$= \frac{P[X = 1|\theta = 1]P[\theta = 1]}{P[X = 1|\theta = 0]P[\theta = 0] + P[X = 1|\theta = 1]P[\theta = 1]} = 0.019$$

Thus, we went from prior $Bern(0.001)$ to posterior $Bern(0.019)$ on θ.

- Let k be a fixed finite number.
- We have k possible mutually exclusive outcomes, with corresponding probabilities $p_1, p_2, ..., p_k$ and n trials

$$\forall 1 \leqslant i \leqslant k: \quad p_i \geqslant 0 \quad \text{and} \quad \sum_{i=1}^{k} p_i = 1$$

- Probability Mass Function:

$$f(x_1, ..., x_k; n, p_1, ...p_k) = \frac{n!}{x_1!...x_n!} p_1^{x_1} ... p_k^{x_k} \quad \text{given} \sum_{i=1}^{k} x_i = n$$

- A k-vector lies in a $k-1$ simplex if $\theta_i \geqslant 0, \quad \sum_{i=1}^{k} \theta_i = 1$
- A k-dimensional Dirichlet random variable θ can take values in $(k-1)$-simplex and has the following distribution on this simplex:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} ... \theta_k^{\alpha_k - 1}$$

Notation

General framework

- Let $\mathcal{V}$ be the set of all words present in our vocabulary and let $|\mathcal{V}| = V$. Then each word is indexed by $i \in \{1, 2, \ldots, V\}$ and represented as a unit basis vector w_i such that $w_i^i = 1$ and $w_i^j = 0$ for $j \neq i$.
- A document is a sequence of N words: $\mathbf{w} = (w_1, w_2, \ldots, w_N)$.
- A corpus is a collection of M documents: $\mathbf{D} = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M\}$.
- A topic is a probability vector z_i with $i \in \{1, 2, \ldots, k\}$, where k is the total number of topics.

Probabilistic latent semantic indexing (pLSI)

Suppose d is the label of some document in $\mathbf{D}$ and w_n is a word present in the document. Then by assuming d and w_n are conditionally independent given some unobserved topic z_i:

$$p(d, w_n) = p(d) \sum_{i=1}^{k} p(w_n|z_i)p(z|d)$$

In this model, d is a dummy variable corresponding to the document in our training corpus. Therefore, the model only learns conditional distributions for previously seen documents, making it unfeasible for predicting new words in unseen documents.

Latent Dirichlet Allocation (LDA)

Generative model

Consider the following generative process:

- For each document i:
  - Draw $\theta_i \sim Dir(\alpha)$
  - For word j in document i:
    - Draw topic $z_{ij} \sim Multinomial(\theta_i)$
    - Draw $w_{ij} \sim p(w_{ij}|z_{ij}, \beta)$

Some observations

- N is independent of all other paramaters and we can thus ignore its randomness
- Dimensionality k of θ is considered fixed a priori
- Conditional word probabilities β_{ij} are considered fixed (we will see later how to estimate them). For clarity, $\beta \in \mathbb{R}^{k \times V}$ such that:

$$\beta_{ij} = p(w^j = 1 | z^i = 1)$$

Joint distribution

Given the model parameters α, and β, a topic mixture θ, a set of N topics $\mathbf{z}$, and a set N words $\mathbf{w}$ follows the distribution given by:

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{i=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta)$$

Notice that $p(z_n|\theta)$ is θ_i for the unique i that yield $z_n^i = 1$, since θ is the mixture of topics.

Marginal distribution

Since we are interested in the marginal distribution of the document w, conditional on the model parameters, let us integrate out θ and sum over all the topics in $\mathbf{z}$.

$$p(\mathbf{w}|\alpha,\beta) = \int p(\theta|\alpha)\left(\prod_{i=1}^{N}\sum_{z_n} p(z_n|\theta)p(w_n|z_n,\beta)\right)d\theta$$

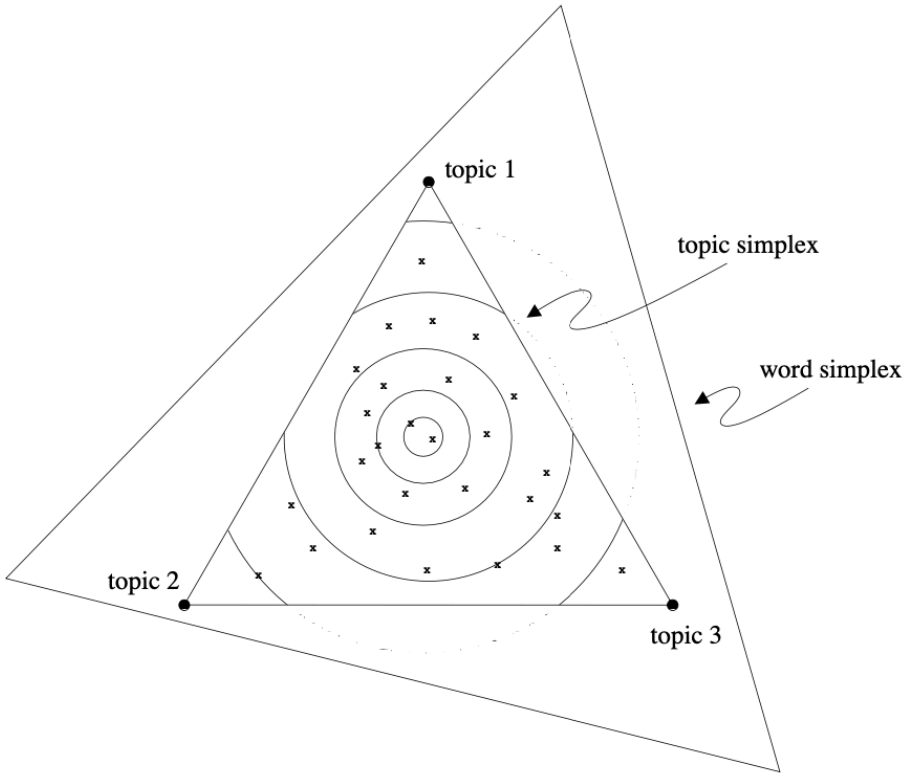
To obtain the probability of the entire corpus D, we need to take the product of all the marginal probabilties of the individual documents. Therefore we obtain:

$$p(D|\alpha,\beta) = \prod_{d=1}^{M}\int p(\theta_d|\alpha)\left(\prod_{i=1}^{N_d}\sum_{z_{dn}} p(z_{dn}|\theta_d)p(w_{dn}|z_{dn},\beta)\right)d\theta_d$$

Some remarks

- α and β are corpus-level parameters, meaning they are only sampled once when generating the corpus.
- θ_d are sampled for every document, meaning they represent a document-specific mixture of topics.
- z_{dn} and w_{dn} are sampled for each word in each document.
- Key point: **documents are associated with several topics** and **words can "belong" to several topics**.

Geometric interpretation



Inference

The key inferential problem that we need to solve in order to use LDA is that of computing the posterior distribution of the hidden variables given a document:

$$p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{w}|\alpha, \beta)}$$

Unfortunately, this distribution is intractable to compute in general.

Indeed, to normalize the distribution we marginalize over the hidden variables and write Eq. (3) in terms of the model parameters:

$$p(\mathbf{w}|\alpha,\beta) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \int \left(\prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \right) \left(\prod_{n=1}^{N} \sum_{i=1}^{k} \prod_{j=1}^{V} (\theta_i \beta_{ij})^{w_n^j} \right) d\theta$$

a function which is intractable due to the coupling between θ and β in the summation over latent topics (Dickey, 1983)

Inference

Although the posterior distribution is intractable for exact inference, a wide variety of approximate inference algorithms can be considered for LDA, including Laplace approximation, variational approximation, and Markov chain Monte Carlo (Jordan, 1999)

Variational Inference

- Jensen's Inequality: obtain a familiy of lower bounds on the log likelihood.
- Each lower bound is indexed by the *variational parameters*
- Variational parameters are chosen by an optimization procedure that attempts to find the tightest lower bound.

Variational Inference

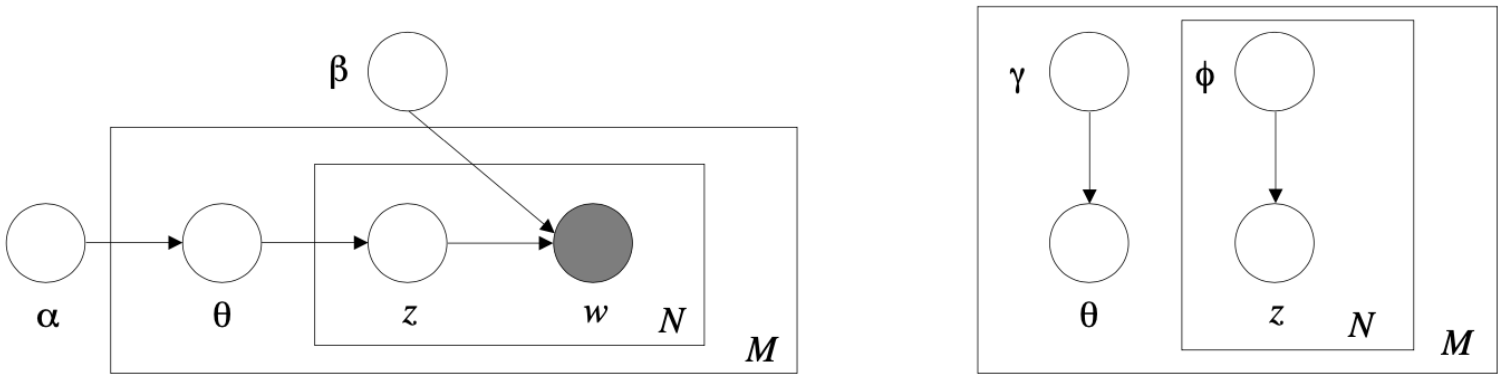


Figure 5: (Left) Graphical model representation of LDA. (Right) Graphical model representation of the variational distribution used to approximate the posterior in LDA.

The family is characterized by the following variational distribution:

$$q(\theta, \mathbf{z}|\gamma, \phi) = q(\theta|\gamma) \prod_{n=1}^{N} q(z_n|\phi_n)$$

Where, γ - Dirichlet parameter, $(\phi_1, ..., \phi_N)$ - Multinomial parameters.

Variational Inference. E-step

Following Jordan et al. (1999), using Jensen Inequality, for any variational distribution $q(\theta, \mathbf{z}|\gamma, \phi)$ we get:

$$\log p(\mathbf{w}|\alpha, \beta) \geqslant \mathbb{E}_q[\log p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)] - \mathbb{E}_q[\log q(\theta, \mathbf{z})] \quad (1)$$

The difference between LHS and RHS is the Kullback–Leibler divergence (KL) between the variational posterior probability and the true posterior probability. That is, letting $L(\gamma, \phi; \alpha, \beta)$ denote the RHS, we have:

$$\log p(\mathbf{w}|\alpha, \beta) = L(\gamma, \phi; \alpha, \beta) + D\left(q(\theta, \mathbf{z}|\theta, \phi) \parallel p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta)\right) \quad (2)$$

Variational Inference. E-step

- Maximizing the lower bound $L(\gamma, \phi; \alpha, \beta)$ is equivalent to minimizing the KL divergence between the variational posterior probability and the true posterior probability. This minimization can be achieved via an **iterative fixed-point method**
- By computing the derivatives of the KL divergence and setting them to zero, we get the update step:

$$\phi_{ni} \propto \beta_{iw_n} exp\{\mathbb{E}_q[\log(\theta_i)|\gamma]\} \tag{3}$$

$$\gamma_i = \alpha_i + \sum_{n=1}^{N} \phi_{ni} \tag{4}$$

where ϕ_{ni} - probability that the n-th word is generated by latent topic i.

The multinomial update can be computed as follows:

$$\mathbb{E}_q[\log(\theta_i)|\gamma] = \Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j)$$

where Ψ is the first derivative of the $\log \Gamma$ function.

Thus, we get the variational inference algorithm for LDA:

(1) initialize $\phi_{ni}^0 := 1/k$ for all i and n
(2) initialize $\gamma_i := \alpha_i + N/k$ for all i
(3) **repeat**
(4) **for** $n = 1$ **to** N
(5) **for** $i = 1$ **to** k
(6) $\phi_{ni}^{t+1} := \beta_{iw_n} \exp(\Psi(\gamma_i^t))$
(7) normalize ϕ_n^{t+1} to sum to 1.
(8) $\gamma^{t+1} := \alpha + \sum_{n=1}^{N} \phi_n^{t+1}$
(9) **until** convergence

Equations 3 and 4 have an appealing intuitive interpretation. The Dirichlet update is a posterior Dirichlet given expected observations taken under the variational distribution, $\mathbb{E}[z_n|\phi_n]$. The multinomial update is akin to using Bayes' theorem, $p(z_n|w_n) \propto p(w_n|z_n)p(z_n)$

Equation 2 can be rewritten as:

$$L(\theta, \phi; \alpha, \beta) = \log \Gamma(\sum_{j=1}^{k} \alpha_j) - \sum_{i=1}^{k} \log \Gamma(\alpha_i)$$

$$+ \sum_{i=1}^{k} (\alpha_i - 1)(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j)) + \sum_{n=1}^{N} \sum_{i=1}^{k} \phi_{ni}(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j))$$

$$+ \sum_{n=1}^{N} \sum_{i=1}^{k} \sum_{j=1}^{V} \phi_{dni} w_{dn}^{j} \log \beta_{ij} - \log \Gamma(\sum_{j=1}^{k} \gamma_j)$$

$$+ \sum_{i=1}^{k} (\gamma_i - 1)(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j))$$

$$- \sum_{n=1}^{N} \sum_{i=1}^{k} \phi_{ni} \log \phi_{ni}$$

Parameter Estimation

- We obtain empirical Bayes estimates of the model parameters α and β by using the variational lower bound as a surrogate for the (intractable) marginal log likelihood, with the variational parameters ϕ and γ fixed to the values found by variational inference. We then obtain (approximate) empirical Bayes estimates by maximizing this lower bound with respect to the model parameters.
- We have thus far considered the log likelihood for a single document. Given our assumption of exchangeability for the documents, the overall log likelihood of a corpus $D = \{\mathbf{w}_1, ..., \mathbf{w}_M\}$ is the sum of the log likelihoods for individual documents; moreover, the overall variational lower bound is the sum of the individual variational bounds.

Parameter Estimation

In the variational E-step, discussed before, we maximize the bound $L(\gamma, \phi; \alpha, \beta)$ with respect to the variational parameters γ and ϕ. In the M-step, which we describe in the following slides, we maximize the bound with respect to the model parameters α and β. The overall procedure can thus be viewed as coordinate ascent in L.

We isolate terms and add Lagrange multipliers:

$$\mathbf{L}_{[\beta]} = \sum_{d=1}^{M} \sum_{n=1}^{N_d} \sum_{i=1}^{k} \sum_{j=1}^{V} \phi_{dni} w_{dn}^{j} \log \beta_{ij} + \sum_{i=1}^{k} \gamma_i \left(\sum_{j=1}^{V} \beta_{ij} - 1 \right)$$

We take the derivative w.r.t. β_{ij}, set it to zero, and find:

$$\beta_{ij} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{dni} w_{dn}^{j}$$

Parameter estimation. M-step. Maximizing w.r.t. α

- Take the derivative of the terms containing α

$$\frac{\partial L}{\partial \alpha_i} = M(\Psi(\sum_{j=1}^{k} \alpha_j) - \Psi(\alpha_i)) + \sum_{d=1}^{M} (\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{k} \gamma_{dj}))$$

- We must use an iterative method to find the maximal α. We can invoke Newton-Raphson algorithm using the Hessian matrix:

$$\frac{\partial L}{\partial \alpha_i \alpha_j} = \delta(i,j) M \Psi'(\alpha_i) - \Psi'(\sum_{j=1}^{k} \alpha_j)$$

Conclusion

Conclusion

- LDA provides a flexible framework for topic modelling by leveraging Bayesian statistics
- The model allows us to learn a distribution of topics without prior information on how these topic might look like
- Variational inference for the model is rather complex, but several implementations exist in popular ML libraries
- In spite of being 20 years old, LDA remains a powerful tool for summarizing text data

Thank you!