STATISTICAL LEARNING SYSTEMS
LECTURE 12: LATENT DIRICHLET ALLOCATION

Jacek Koronacki

Institute of Computer Science, Polish Academy of Sciences

Ph. D. Program 2013/2014

PhD STUDIES

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

The project is co-financed by the European Union within the framework of European Social Fund

Latent Dirichlet Allocation (LDA)

Following Blei, Ng and Jordan, *J. ML Research* 3 (2003), 993-1022, we describe LDA, a generative probabilistic model for collections of discrete data such as text corpora.

Let us define:

- A word - the basic unit of discrete data, being an item from a vocabulary indexed by $\{1, \ldots, V\}$; the v-th word in the vocabulary is represented by a V-vector w such that $w^{(v)} = 1$ and $w^{(u)} = 0$ for $u \neq v$,
- A document is a sequence of N words denoted by $\mathbf{w} = (w_1, \ldots, w_N)$, where w_n is the n-th word in the sequence.
- A corpus is a collection of M documents denoted by $D = \{\mathbf{w}_1, \ldots, \mathbf{w}_M\}$.

PhD STUDIES
KAPITAŁ LUDZKI NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA EUROPEJSKI FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Latent Dirichlet Allocation (LDA)

Let us present first some simpler models for text and start with the simplest possible one,

the unigram model, under which the words of every document are drawn independently from a single multinomial distribution:

$$p(\mathbf{w}) = \prod_{n=1}^{N} p(w_n).$$

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

If we augment the unigram model with discrete random topic variable z, we obtain a mixture of unigrams model

under which each document is generated by first choosing a topic z and then generating N words independently from the conditional multinomial $p(w|z)$:

$$p(\mathbf{w}) = \sum_z p(z) \prod_{n=1}^{N} p(w_n|z).$$

PhD STUDIES

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

The project is co-financed by the European Union within the framework of European Social Fund

Latent Dirichlet Allocation (LDA)

Probabilistic latent semantic indexing model (pLSI) posits that a document identity (label) d and a word w_n are conditionally independent given an unobserved topic z:

$$p(d, w_n) = \sum_z p(z)p(d|z)p(w_n|z) = p(d)\sum_z p(w_n|z)p(z|d).$$

The model captures the possibility that a document may contain multiple topics since $p(z|d)$ serves as the mixture weights of the topics for a particular document d.

The parameters for the k-topic pLSI model are k multinomial distributions of size V and M mixtures over the k hidden topics ($kV + kM$ parameters).

(It is important to note that d is a dummy index into the list of documents in the training set; thus, the model learns the topic mixtures $p(z|d)$ only for those documents on which it is trained.)

Latent Dirichlet Allocation (LDA)

In the LDA generative probabilistic model, documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. The generative process for each document **w** is:

- Choose $N \sim \mathrm{Poisson}(\xi)$.
- Choose $\theta \sim \mathrm{Dir}(\alpha)$ with fixed and known dimension k.
- For each of the N words w_n:
  - Choose a topic $z_n \sim \mathrm{Mulitnomial}(\theta)$.
  - Choose a word w_n from $p(w_n|z_n,\beta)$, a multinomial probability conditioned on the topic z_n, with matrix parameter β.

The word probabilities are parameterized by a $k \times V$ matrix β, where $\beta_{ij} = p(w^{(j)} = 1|z^{(i)} = 1)$. N is independent of all the other data generating variables (θ and $\mathbf{z}$). Dirichlet r.v. θ is k-dimensional and takes values in the $(k-1)$-simplex $(\sum_{i=1}^{k} \theta_i = 1,\ \theta_i \geqslant 0)$,

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

where α is a k-vector with components $\alpha_i > 0$. Parameters α and β ($k + kV$ parameters) are to be estimated.

Latent Dirichlet Allocation (LDA)

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

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

where $p(z_n|\theta)$ is θ_i for the unique i such that $z_n^{(i)} = 1$. Hence we obtain the marginal distribution of a document:

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

and the probability of a corpus:

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

Strictly speaking, the inferential problem is intractable. Indeed, we want to maximize (w.r.t α and β) the (marginal) log likelihood of the data:

$$\ell(\alpha,\beta)=\sum_{d=1}^{M}\log p(\mathbf{w}_d|\alpha,\beta),$$

but

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

and this function is intractable due to the coupling between θ and β. However, approximate inference algorithms for LDA are well-known.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

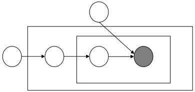
Latent Dirichlet Allocation (LDA)

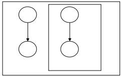
Graphical representation of the original model includes, for each of the N words in each of the M documents, edges between α, θ, z and w (depicted as a grey node), and an edge from β to w.

By dropping the edges between between θ, z and w, as well dropping the w node, we obtain a simplified graphical model with free variational parameters, γ and $(\phi_1, \ldots, \phi_N)$, which 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 is the Dirichlet parameter and $(\phi_1, \ldots, \phi_N)$ are the multinomial parameters:





Latent Dirichlet Allocation (LDA)

Having specified a simplified family of probability distributions, the next step is to set up an optimization problem that determines the values of the variational parameters γ and ϕ. This is done by minimizing the Kullback-Leibler divergence between the variational distribution and the true posterior $p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta)$:

$$(\gamma^\star, \phi^\star) = \arg\min_{(\gamma,\phi)} \mathrm{D}(q(\theta, \mathbf{z}|\gamma, \phi) \parallel p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta)).$$

Blei, Ng and Jordan have shown that

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

for some well-defined and computationally tractable L. Thus, maximizing L w.r.t. γ and ϕ is equivalent to minimizing the KL divergence. Maximizing then the resulting L w.r.t. α and β provides an approximation to the ML estimates for the latter two parameters.

During the lecture, some applications of LDA will be sketched.

PhD STUDIES

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

The project is co-financed by the European Union within the framework of European Social Fund