Co-occurrences

Course of Machine Learning
Master Degree in Computer Science
University of Rome “Tor Vergata”

Giorgio Gambosi

a.a. 2017-2018

Latent semantic analysis

General aspects

Definition

All models refer to co-occurrence data.

They consider, given two collections $\mathbf{V}, \mathbf{D}$ (for example, terms and documents, or customers and items) a sequence of observations $\mathbf{W} = \{(w_1, d_1), \ldots, (w_N, d_N)\}$, with $w_i \in \mathbf{V}, d_i \in \mathbf{D}$ (for example, occurrences of terms in documents, customers accessing at item description, etc.)

Introduction to LSA

Basic assumptions

The approach of LSA (Latent Semantic Analysis) refers to three assumptions:

- sematic information can be derived from the $\mathbf{V}, \mathbf{D}$ matrix
- dimensionality reduction is a key aspect for such derivation
- “terms” and “documents” can be modeled as points (vectors) in a euclidean space

Framework

1. Dictionary $\mathbf{V}$ of $V = |\mathbf{V}|$ terms $t_1, t_2, \ldots, t_V$
2. Corpus $\mathbf{D}$ of $D = |\mathbf{D}|$ documents $d_1, d_2, \ldots, d_D$
3. Each document d_i is a sequence of N_i occurrences of terms from $\mathbf{V}$

Model

Idea

1. Each document d_i is considered as a multiset of N_i terms from $\mathbf{V}$ (hypothesis “bag of words”)
2. There exists a correspondance between $\mathbf{V}$ and $\mathbf{D}$, and a vector space $\mathcal{S}$. To each term t_i a vector $\mathbf{u}_i$ is associated, hence to each document d_j it is associated a vector $\mathbf{v}_j$ in $\mathcal{S}$

Occurrence matrix

Matrix $\mathbf{W} \in \mathbb{R}^{V \times D}$: $\mathbf{W}(i, j)$ is associated to the occurrences of term t_i in document d_j. The value of $\mathbf{W}(i, j)$ depends from the measure function predefined (tf, tf-idf, entropy, etc.).

- Terms: row vectors (dimension D)
- Documents: column vectors (dimension V)

Model

Problems

1. The values V and D are very large
2. The vectors for t_i and $d_j$are very sparse
3. The space for terms and documents are different

Solution

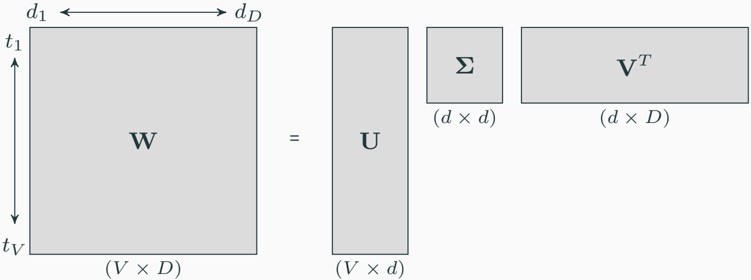
Applying singular value decomposition.

Let $\mathbf{W} \in \mathbb{R}^{n \times m}$ a matrix of rank $d \leq \min(n, m)$ and let $n > m$. Then, there exist

- $\mathbf{U} \in \mathbb{R}^{n \times d}$ orthonormal ($\mathbf{U}^T\mathbf{U} = \mathbf{I}_d$)
- $\mathbf{V} \in \mathbb{R}^{m \times d}$ orthonormal ($\mathbf{V}\mathbf{V}^T = \mathbf{I}_d$)
- $\mathbf{\Sigma} \in \mathbb{R}^{d \times d}$ diagonal

such that $\mathbf{W} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$

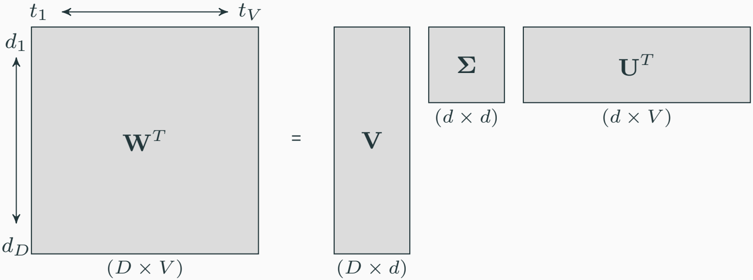
Application of SVD



Effect

The rows of $\mathbf{W}$ (terms) are projected on a d-dimensional subspace of $\mathbb{R}^D$ having the set of columns of $\mathbf{V}$ as basis: this defines for each term a new representation (row of $\mathbf{U\Sigma} \in \mathbb{R}^d$) as a vector of the coordinates with respect to this basis

Application of SVD



Effect

The rows of $\mathbf{W}^T$ (documents) are projected on a d-dimensional subspace of $\mathbb{R}^V$ having the set of columns of $\mathbf{U}$ as basis: this defines for each document a new representation (row of $\mathbf{V\Sigma} \in \mathbb{R}^d$) as a vector of the coordinates with respect to this basis

LSA

Dimensionality reduction

The dimension d of the projection space may be predefined, and less than the rank of $\mathbf{W}$. In this case,

$$\mathbf{W} \approx \overline{\mathbf{W}} = \mathbf{U\Sigma V}^T$$

Approximation

The property

$$\min_{\mathbf{A}:\text{rank}(\mathbf{A})=d} ||\mathbf{W} - \mathbf{A}||_2 = ||\mathbf{W} - \overline{\mathbf{W}}||_2$$

holds. The matrix $\overline{\mathbf{W}}$ is the matrix that best approximates $\mathbf{W}$ among all matrices of rank d according to the norm L_2 or of Frobenius

$$||\mathbf{A}||_2 = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2}$$

LSA

Effect

SVD defines a transformation from two discrete vector spaces $\mathcal{V} \in \mathbb{Z}^D$ and $\mathcal{D} \in \mathbb{Z}^V$, to one smaller continuous vector space, $\mathcal{T} \in \mathbb{R}^d$.

The dimension of $\mathcal{T}$ is less than or equal to the rank (unknown) of $\mathbf{W}$, and it is lower bounded from the amount of distortion acceptable in the projection.

Interpretation

$\hat{\mathbf{W}}$ captures the largest part of the associations between terms and documents $\mathbf{W}$, neglecting the least significative relations.

- Each term is represented as a (linear) combinations of hidden concepts, corresponding to the columns of $\mathbf{V}$: terms with projections near to each other tend to appear in the same documents (or in semantically similar documents)
- Each document is represented as a (linear) combinations of hidden topics, corresponding to the columns of $\mathbf{U}$: documents with projections near to each other tend to include the same terms (or semantically similar terms)

LSA and clustering

Co-occurrences

- $\mathbf{W}\mathbf{W}^T \in \mathbb{Z}^{V \times V}$ represents the co-occurrences between terms in $\mathbf{V}$ (number of documents where the two terms both occur)
- $\mathbf{W}^T\mathbf{W} \in \mathbb{Z}^{D \times D}$ represents the co-occurrences between documents in $\mathbf{D}$ (number of terms in common between them)

SVD and co-occurrence matrix

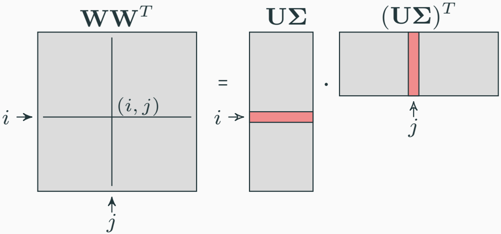
By applying SVD,

$$\mathbf{W}\mathbf{W}^T = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\mathbf{V}\mathbf{\Sigma}\mathbf{U}^T = \mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^T$$

and

$$\mathbf{W}^T\mathbf{W} = \mathbf{V}\mathbf{\Sigma}\mathbf{U}^T\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^T$$

Term clustering



Term proximity

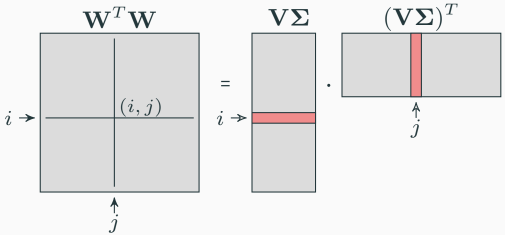
Reasonable measure of the proximity between terms t_i and t_j: value of item (i, j) of $\mathbf{W}\mathbf{W}^T$, hence of the inner product between $\mathbf{u}_i$ (i-th row of $\mathbf{U}\mathbf{\Sigma}$) and $\mathbf{u}_j$ (j-th row of $\mathbf{U}\mathbf{\Sigma}$).

In particolar,

$$\mathcal{D}(t_i, t_j) = \frac{1}{\cos(\mathbf{u}_i, \mathbf{u}_j)} = \frac{||\mathbf{u}_i|| \cdot ||\mathbf{u}_j||}{\mathbf{u}_i \mathbf{u}_j^T}$$

can be assumed as a measure of the distance between terms

Document clustering



Document proximity

Reasonable measure of the proximity between documents d_i and d_j: value of item (i,j) of $\mathbf{W}^T\mathbf{W}$, hence of the inner product between $\mathbf{v}_i$ (i-th row of $\mathbf{V\Sigma}$) and $\mathbf{v}_j$ (j-th row of $\mathbf{V\Sigma}$). In particolar,

$$\mathcal{D}(d_i,d_j)=\frac{1}{\cos(\mathbf{v}_i,\mathbf{v}_j)}=\frac{||\mathbf{v}_i||\cdot||\mathbf{v}_j||}{\mathbf{v}_i\mathbf{v}_j^T}$$

can be assumed as a measure of the distance between documents

Classification

Objective

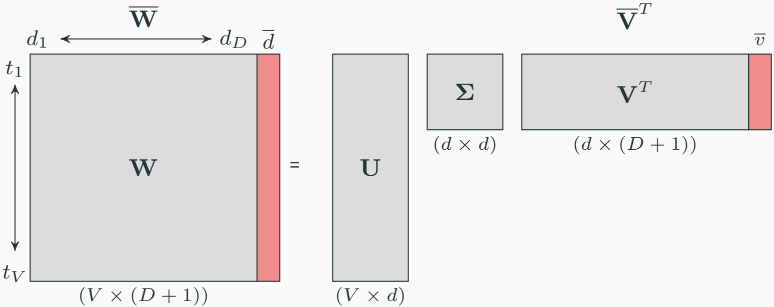
Determining, given a document, to which topic (in un predefined collection) its content is most related.

Approach

Construction of a vector of (possibly weighted) terms, to describe the class: can be seen as an additional document $\overline{d}$ (template of the class)

$\mathbf{W}$ can be extended by appending $\overline{d}$ as the $D+1$-th column of $\mathbf{W}$ (thus obtaining $\overline{\mathbf{W}} \in \mathbb{Z}^{V \times (D+1)}$)

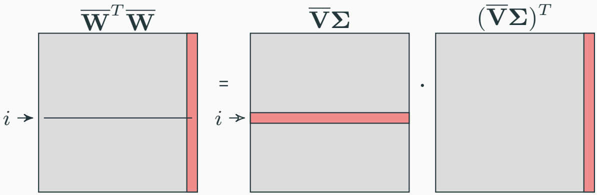
Classification



Effect

SVD introduces an additional vector $\overline{\mathbf{v}} \in \mathbb{R}^d$ as $D+1$-th row of $\mathbf{V}$, where $\overline{d} = \mathbf{U}\boldsymbol{\Sigma}\overline{\mathbf{v}}^T$

Classification


$\overline{\mathbf{W}}^T\overline{\mathbf{W}}$
$\overline{\mathbf{V}}\mathbf{\Sigma}$
$(\overline{\mathbf{V}}\mathbf{\Sigma})^T$
=
·
i
i

Proximity between document and class

Reasonable measure of the proximity between a document d_i and a class $\overline{d}$: value of item $(i, D+1)$ of $\overline{\mathbf{W}}^T\overline{\mathbf{W}}$, hence of the inner product between $\mathbf{v}_i$ (i-th row ofi $\overline{\mathbf{V}}\mathbf{\Sigma}$) and $\overline{\mathbf{v}}$ ($(D+1)$-th row of $\overline{\mathbf{V}}\mathbf{\Sigma}$).

In particolar,

$$\mathcal{D}(d_i, \overline{d}) = \frac{1}{\cos(\mathbf{v}_i, \overline{\mathbf{v}})} = \frac{||\mathbf{v}_i|| \cdot ||\overline{\mathbf{v}}||}{\mathbf{v}_i \overline{\mathbf{v}}^T}$$

can be assumed as a measure of the distance between a document and a class

Text modeling

Text modeling

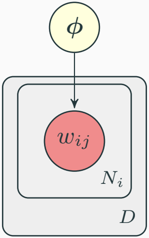
Exchangeability

- Term occurrences in a document can be seen as a set of random variables $w_i, i = 1, \ldots, N$.
- In general, we may assume that they are identically distributed, but not that they are independent. However, they are exchangeable: that is, their joint distribution $p(w_1, \ldots, w_n)$ is independent from their order.
- More formally, for any permutation $\pi : \{1, \ldots, N\} \mapsto \{1, \ldots, N\}$, we have $p(w_1, \ldots, w_n) = p(w_{\pi(1)}, \ldots, w_{\pi(n)})$. This is the bag of words assumption in language modeling and information retrieval.

Unigram model

Definition

Terms, in all documents, are instances of i.i.d. random variables, distributed according to a multinomial (with parameter $\boldsymbol{\phi}$) on the elements of dictionary $\mathbf{V}$.



$w_{ij} \sim \mathsf{Mult}(t|\boldsymbol{\phi})$

Note

N_i is the length document d_i.

Distribution

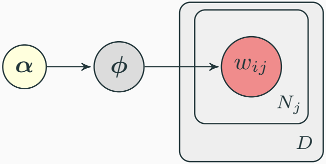
$$p(\mathbf{W}|\boldsymbol{\phi}) = \prod_{j=1}^{D}\prod_{i=1}^{N_i} p(w_{ij}|\boldsymbol{\phi}) = \prod_{k=1}^{V} p(t_k|\boldsymbol{\phi})^{n_k} = \prod_{k=1}^{V} \mathsf{Mult}(t_k|\boldsymbol{\phi})^{n_k}$$

where $V = |\mathbf{V}|$ and n_k is the number of occurrences of term $t_k \in \mathbf{V}$ into $\mathbf{W}$

Unigram model

Extension

We may also assume that the multinomial distribution is instance of another random variable with its own distribution, for example a Dirichlet with parameter $\boldsymbol{\alpha}$



$$\boldsymbol{\phi} \sim \mathsf{Dir}(\boldsymbol{\phi}|\boldsymbol{\alpha})$$
$$w_{ij} \sim \mathsf{Mult}(w|\boldsymbol{\phi})$$

Distribution

$$p(\mathbf{W}|\boldsymbol{\alpha}) = \int_{\boldsymbol{\phi}} \prod_{j=1}^{D}\prod_{i=1}^{N_j} p(w_{ij}|\boldsymbol{\phi})p(\boldsymbol{\phi}|\boldsymbol{\alpha})d\boldsymbol{\phi} = \int_{\boldsymbol{\phi}} p(\boldsymbol{\phi}|\boldsymbol{\alpha})\prod_{j=1}^{D}\prod_{i=1}^{N_j} p(w_{ij}|\boldsymbol{\phi})d\boldsymbol{\phi}$$

$$= \int_{\boldsymbol{\phi}} p(\boldsymbol{\phi}|\boldsymbol{\alpha})\prod_{k=1}^{V} p(t_k|\boldsymbol{\phi})^{n_k} d\boldsymbol{\phi} = \int_{\boldsymbol{\phi}} \mathsf{Dir}(\boldsymbol{\phi}|\boldsymbol{\alpha})\prod_{k=1}^{V} \mathsf{Mult}(t_k|\boldsymbol{\phi})^{n_k} d\boldsymbol{\phi}$$

Bayesian inference in the bag of words case

By Bayes theorem,

$$p(\boldsymbol{\phi}|\mathbf{W},\boldsymbol{\alpha}) = \frac{p(\mathbf{W}|\boldsymbol{\phi},\boldsymbol{\alpha})p(\boldsymbol{\phi}|\boldsymbol{\alpha})}{p(\mathbf{W}|\boldsymbol{\alpha})}$$

by the bayesian network structure, we have that $\mathbf{W}$ is conditionally independent from $\boldsymbol{\alpha}$, given $\boldsymbol{\phi}$, that is

$$p(\mathbf{W},\boldsymbol{\alpha}|\boldsymbol{\phi}) = p(\mathbf{W}|\boldsymbol{\phi})p(\boldsymbol{\alpha}|\boldsymbol{\phi})$$

then

$$p(\mathbf{W}|\boldsymbol{\alpha},\boldsymbol{\phi}) = \frac{p(\mathbf{W},\boldsymbol{\alpha}|\boldsymbol{\phi})}{p(\boldsymbol{\alpha}|\boldsymbol{\phi})} = \frac{p(\mathbf{W}|\boldsymbol{\phi})p(\boldsymbol{\alpha}|\boldsymbol{\phi})}{p(\boldsymbol{\alpha}|\boldsymbol{\phi})} = p(\mathbf{W}|\boldsymbol{\phi})$$

and

$$\begin{aligned} p(\boldsymbol{\phi}|\mathbf{W},\boldsymbol{\alpha}) &= \frac{p(\mathbf{W}|\boldsymbol{\phi})p(\boldsymbol{\phi}|\boldsymbol{\alpha})}{p(\mathbf{W}|\boldsymbol{\alpha})} = \frac{\prod_{j=1}^{D}\prod_{i=1}^{N_j} p(w_{ij}|\boldsymbol{\phi})p(\boldsymbol{\phi}|\boldsymbol{\alpha})}{p(\mathbf{W}|\boldsymbol{\alpha})} \\ &= \frac{\prod_{j=1}^{D}\prod_{i=1}^{N_j} \text{Mult}(w_{ij}|\boldsymbol{\phi})\text{Dir}(\boldsymbol{\phi}|\boldsymbol{\alpha})}{p(\mathbf{W}|\boldsymbol{\alpha})} = \frac{\prod_{k=1}^{V} \text{Mult}(t_k|\boldsymbol{\phi})^{n_k}\text{Dir}(\boldsymbol{\phi}|\boldsymbol{\alpha})}{p(\mathbf{W}|\boldsymbol{\alpha})} \end{aligned}$$

Bayesian inference in the bag of words case

Let $Z = p(\mathbf{W}|\boldsymbol{\alpha})$, which is constant with respect to $\boldsymbol{\phi}$, and let $\Delta(\boldsymbol{\alpha})$ be the inverse of the normalization constant of the Dirichlet distribution Dir$(\boldsymbol{\phi}|\boldsymbol{\alpha})$

$$\Delta(\boldsymbol{\alpha}) = \frac{\prod_{k=1}^{V} \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{V} \alpha_k)}$$

By the definition of Multinomial and Dirichlet distributions, we then have

$$p(\boldsymbol{\phi}|\mathbf{W}, \boldsymbol{\alpha}) = \frac{1}{Z} \prod_{k=1}^{V} \phi_k^{n_k} \frac{1}{\Delta(\boldsymbol{\alpha})} \prod_{k=1}^{V} \phi_k^{\alpha_k - 1} = \frac{1}{Z} \frac{1}{\Delta(\boldsymbol{\alpha})} \prod_{k=1}^{V} \phi_k^{\alpha_k + n_k - 1}$$

which, apart from the normalizing constant, is

$$\text{Dir}(\boldsymbol{\phi}|\boldsymbol{\alpha} + \mathbf{n}) = \frac{1}{\Delta(\boldsymbol{\alpha} + \mathbf{n})} \prod_{k=1}^{V} \phi_k^{\alpha_k + n_k - 1}$$

where $\mathbf{n} = (n_1, \dots, n_V)$.

Bayesian inference in the bag of words case

Evidence

Assigning to the normalizing constant Z a value suitable to obtain a probability distribution makes it possible to derive the evidence

$$Z = p(\mathbf{W}|\boldsymbol{\alpha}) = \frac{\Delta(\boldsymbol{\alpha} + \mathbf{n})}{\Delta(\boldsymbol{\alpha})} = \frac{\Gamma(\sum_{j=1}^{V} \alpha_j) \cdot \prod_{k=1}^{V} \Gamma(\alpha_k + n_k)}{\Gamma(\sum_{j=1}^{V} (\alpha_j + n_j)) \cdot \prod_{k=1}^{V} \Gamma(\alpha_k)}$$

Point estimate

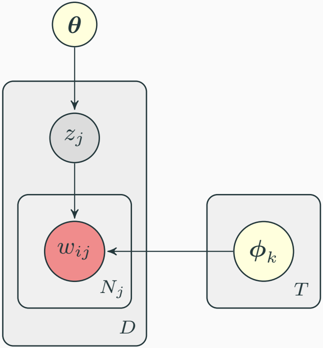
A point estimate for $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_V)$ can be derived, in case, by considering the expected value of the posterior distribution

$$\hat{\boldsymbol{\phi}} = E[\mathrm{Dir}(\boldsymbol{\phi}|\boldsymbol{\alpha} + \mathbf{n})] = \frac{\boldsymbol{\alpha} + \mathbf{n}}{\sum_{k=1}^{V} (\alpha_k + n_k)}$$

Mixture of unigrams model

Definition

- Latent variable model: one variable z with domain $1, \ldots, T$
- For each document, a multinomial $\boldsymbol{\phi}_j$ (corresponding to a topic) is sampled from a set of T predefined topics
- Each term occurrence in a document is sampled from the multinomial associated to that document



$w_{ij} \sim \text{Mult}(w|\boldsymbol{\phi})$

Mixture of unigrams model

Distribution

$$
\begin{aligned}
p(\mathbf{W}|\boldsymbol{\theta}, \boldsymbol{\phi}_{1:T}) &= \prod_{j=1}^{D} \left(\sum_{k=1}^{T} p(z_j = k|\boldsymbol{\theta}) \prod_{i=1}^{N_j} p(w_{ij}|z_j = k, \boldsymbol{\phi}_{1:T}) \right) \\
&= \prod_{j=1}^{D} \left(\sum_{k=1}^{T} \pi_k \prod_{i=1}^{N_j} p(w_{ij}|\boldsymbol{\phi}_k) \right) = \prod_{j=1}^{D} \left(\sum_{k=1}^{T} \pi_k \prod_{i=1}^{N_j} \mathsf{Mult}(w_{ij}|\boldsymbol{\phi}_k) \right) \\
&= \prod_{j=1}^{D} \left(\sum_{k=1}^{T} \pi_k \prod_{i=1}^{V} \mathsf{Mult}(t_i|\boldsymbol{\phi}_k)^{n_i} \right) = \prod_{j=1}^{D} \left(\sum_{k=1}^{T} \pi_k \prod_{i=1}^{V} \phi_{ki}^{n_i} \right)
\end{aligned}
$$

with $\pi_k = p(z = k|\boldsymbol{\theta})$ and $\phi_{ki} = p(t_i|\boldsymbol{\phi}_k)$

Classification

The characterization in terms of mixture can be used also for classification: a mixture component = a class

Mixture of unigrams model

Parameter estimate

The evidence corresponds to the probability distribution for the observed dataset in a mixture model

$$p(\mathbf{X}|\boldsymbol{\psi}) = \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k q_k(x_i|\theta_k)$$

whose maximal likelihood is evaluated by means of the EM algorithm.

In this case, we have that $N = D$ (dataset items correspond to documents) and the class-conditioned distributions $q_k(x_i|\theta_k)$ correspond to the probabilities of a document under the multinomial distribution associated to a topic

$$q_k(d_i|\boldsymbol{\phi}_k) = \prod_{j=1}^{V} \phi_{kj}^{n_{ij}}$$

where n_{ij} is the number of occurrences of term t_j in document d_i.

Mixture of unigrams model

Applying EM to mixtures

By the general discussion on the application of the EM algorithm to mixtures models, we have

$$\overline{\gamma}_k(x_i) = \frac{\overline{\pi}_k q_k(x_i|\overline{\theta}_k)}{\sum_{j=1}^{K} \overline{\pi}_j q_j(x_i|\overline{\theta}_j)}$$

$$\pi_k = \frac{1}{N}\sum_{i=1}^{N} \overline{\gamma}_k(x_i)$$

and θ_k the solutions of

$$\sum_{i=1}^{N} \frac{\overline{\gamma}_k(x_i)}{q_k(x_i|\theta_k)} \frac{\partial q_k(x_i|\theta_k)}{\partial \theta_k} = \sum_{i=1}^{N} \overline{\gamma}_k(x_i) \frac{\partial \log q_k(x_i|\theta_k)}{\partial \theta_k} = 0$$

Mixture of unigrams model

Applying EM here

In this framework, it results

$$\overline{\gamma}_k(d_i) = \frac{\overline{\pi}_k \prod_{j=1}^{V} \overline{\phi}_{kj}^{n_{ij}}}{\sum_{t=1}^{T} \overline{\pi}_t \prod_{j=1}^{V} \overline{\phi}_{tj}^{n_{ij}}}$$

$$\pi_k = \frac{1}{D} \sum_{i=1}^{D} \overline{\gamma}_k(d_i)$$

For what regards ϕ_{kj}, we must take into account the constraints

$$\sum_{j=1}^{V} \phi_{kj} = 1$$

for $k = 1, \ldots, T$.

Mixture of unigrams model

Applying EM here

By applying the lagrangian multipliers method, we have that the ϕ_{kj} values are the solutions of

$$\begin{aligned} 0 &= \sum_{i=1}^{D} \frac{\overline{\gamma}_k(d_i)}{\prod_{t=1}^{V} \phi_{kt}^{n_{it}}} \frac{\partial \prod_{t=1}^{V} \phi_{kt}^{n_{it}}}{\partial \phi_{kj}} - \frac{\partial}{\partial \phi_{kj}} \lambda \left(\sum_{j=1}^{V} \phi_{kj} - 1 \right) \\ &= \sum_{i=1}^{D} \frac{\overline{\gamma}_k(d_i)}{\prod_{t=1}^{V} \phi_{kt}^{n_{it}}} \frac{n_{ij} \prod_{t=1}^{V} \phi_{kt}^{n_{it}}}{\phi_{kj}} - \lambda = \frac{1}{\phi_{kj}} \sum_{i=1}^{D} n_{ij} \overline{\gamma}_k(d_i) - \lambda \\ 0 &= \frac{\partial}{\partial \lambda} \lambda \left(\sum_{j=1}^{V} \phi_{kj} - 1 \right) = \sum_{j=1}^{V} \phi_{kj} - 1 \end{aligned}$$

From which we easily obtain

$$\phi_{kj} = \frac{\sum_{i=1}^{D} n_{ij} \overline{\gamma}_k(d_i)}{\sum_{i=1}^{D} \sum_{j=1}^{V} n_{ij} \overline{\gamma}_k(d_i)}$$

Let us also remind that a ML estimate of the probability that document d_i is associated to the k-th topic, that is that $z_i = k$ is given by the responsibility $\gamma_k(d_i)$.

Probabilistic latent semantic analysis

Aspect model

Definition

Latent variables statistical model for co-occurence data:

- It considers, given two sets $\mathbf{V}, \mathbf{D}$ (for example, terms and documents) a sequence of observations $\mathbf{W} = \{(w_1, d_1), \ldots, (w_N, d_N)\}$, with w_i instance of a random variable with domain $\mathbf{V}$ and d_i instance of a random variable with domain $\mathbf{D}$ (for example, occurrences of terms in documents)
- It defines a set of T latent classes, associated to the integers $1, \ldots, T$, that provides a partition of the set $\mathbf{W}$ of observations
- It associates to each observation (w_i, d_i) a value $z_i \in \{1, \ldots, T\}$ of a latent (unobserved) class variable, denoting the class to which the observation belongs, obtaining a sequence $\{z_1, \ldots, z_N\}$

The set of observations in a same class is denoted as aspect

Relation with clustering

Clustering models provide partitions of set of objects (set $\mathbf{V}$ or set $\mathbf{D}$): the aspect model provides a partition of co-occurrences

Aspect model

Probabilistic assumptions

The model relies on two independence hypothesis

- Observations in $\mathbf{W}$ are independent and identically distributed

$$p((w_i, d_i), (w_j, d_j)) = p((w_i, d_i))p((w_j, d_j)) \qquad 1 \leq i, j \leq N, i \neq j$$

- For each observation $(w_i, d_i) \in \mathbf{W}$, w_i e d_i are conditionally independent, given the value z_i of the latent variable

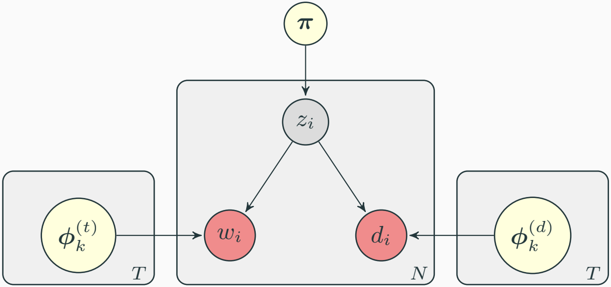
$$p(w_i, d_i|z_i) = p(w_i|z_i)p(d_i|z_i) \qquad 1 \leq i \leq N$$

Aspect model

Generative model

The model can be seen as a generative model on data, with the following behaviour:

- For $i = 1$ to N
  - Sample a value (class) z_i with probability $p(z_i)$
  - Sample a term w_i with probability $p(w_i|z_i)$
  - Sample a document d_i with probability $p(d_i|z_i)$



Joint probability

Probability

Joint probability of observations (evidence):

$$p(\mathbf{W}) = \prod_{i=1}^{N} p(w_i, d_i)$$

Joint probability of observations and latent variables:

$$p(\mathbf{W}, \mathbf{Z}) = \prod_{i=1}^{N} p(w_i, d_i, z_i) = \prod_{i=1}^{N} p(z_i) p(w_i|z_i) p(d_i|z_i)$$

Mixtures and aspect model

Mixture of distributions for the aspect model

$$p(w_i, d_i) = \sum_{z_i=1}^{T} p(w_i, d_i, z_i) = \sum_{z_i=1}^{T} p(w_i, d_i|z_i)p(z_i) = \sum_{z_i=1}^{T} p(w_i|z_i)p(d_i|z_i)p(z_i)$$

Effect on evidence

$$\begin{aligned} p(\mathbf{W}) &= \prod_{i=1}^{N} \sum_{z_i=1}^{T} p(w_i|z_i)p(d_i|z_i)p(z_i) \\ &= \prod_{w \in \mathbf{V}} \prod_{d \in \mathbf{D}} \left(\sum_{z=1}^{T} p(w|z)p(d|z)p(z) \right)^{n(w,d)} \end{aligned}$$

where

$$n(w, d) = |\{(w_i, d_i) \in \mathbf{W} : w_i = w \wedge d_i = d\}|$$

Parameter estimation through EM

Log likelihood of observations probability

$$
\begin{aligned}
l(\mathbf{W}) = \sum_{i=1}^{N} \log p(w_i, d_i) &= \sum_{w \in \mathbf{V}} \sum_{d \in \mathbf{D}} n(w, d) \log p(w, d) \\
&= \sum_{w \in \mathbf{V}} \sum_{d \in \mathbf{D}} n(w, d) \log \left(\sum_{z=1}^{T} p(w|z) p(d|z) p(z) \right)
\end{aligned}
$$

E-step

From the current parameter estimate, compute the posterior probabilities of latent variables

$$
p(z|w, d) = \frac{p(w, d|z) p(z)}{\sum_{z'=1}^{T} p(w, d|z') p(z')} = \frac{p(w|z) p(d|z) p(z)}{\sum_{z'=1}^{T} p(w|z') p(d|z') p(z')}
$$

Parameter estimation through EM

M-step

From the current estimate of the posterior probability of latent variables, determine (by applying ML) an estimate of parameter values, taking into account whenever necessary the constraint that probabilities should sum to 1.

The following estimates are obtained:

$$p(w|z) = \frac{1}{N(z)} \sum_{d \in \mathbf{D}} n(w,d) p(z|w,d)$$

$$p(d|z) = \frac{1}{N(z)} \sum_{w \in \mathbf{V}} n(w,d) p(z|w,d)$$

$$p(z) = \frac{N(z)}{N}$$

where

$$N(z) = \sum_{w \in \mathbf{V}} \sum_{d \in \mathbf{D}} n(w,d) p(z|w,d) \qquad N = \sum_{w \in \mathbf{V}} \sum_{d \in \mathbf{D}} n(w,d)$$

Parameter estimation through EM

Proof for $p(w|z)$

$$F = \sum_{w \in \mathbf{V}} \sum_{d \in \mathbf{D}} n(w,d) \log \left(\sum_{z=1}^{T} p(w|z)p(d|z)p(z) \right) - \sum_{z=1}^{T} \lambda_z \left(\sum_{w \in \mathbf{V}} p(w|z) - 1 \right)$$

$$\frac{\partial F}{\partial p(w|z)} = \sum_{d \in \mathbf{D}} n(w,d) \frac{p(d|z)p(z)}{\sum_{z'=1}^{T} p(w|z')p(d|z')p(z')} - \lambda_z = 0$$

hence

$$p(w|z) = \frac{\sum_{d \in \mathbf{D}} n(w,d)p(z|w,d)}{\lambda_z}$$

since

$$p(w,d) = \sum_{z'=1}^{T} p(w|z')p(d|z')p(z')$$

$$p(w,d,z) = p(w|z)p(d|z)p(z) = p(z|w,d)p(w,d)$$

Parameter estimation through EM

Proof for $p(w|z)$

Since

$$\frac{\partial F}{\partial \lambda_z} = \sum_{w \in \mathbf{V}} p(w|z) - 1 = 0$$

we have

$$1 = \sum_{w \in \mathbf{V}} p(w|z) = \frac{\sum_{w \in \mathbf{V}} \sum_{d \in \mathbf{D}} n(w,d) p(z|w,d)}{\lambda_z}$$

and

$$\lambda_z = \sum_{w \in \mathbf{V}} \sum_{d \in \mathbf{D}} n(w,d) p(z|w,d) = N(z)$$

and, finally,

$$p(w|z) = \frac{1}{N(z)} \sum_{d \in \mathbf{D}} n(w,d) p(z|w,d)$$

Use of discriminative terms

Model

$$p(w,d) = \lambda p_B(w) + (1-\lambda)\sum_{z=1}^{T} p(w|z)p(d|z)p(z)$$

- $p_B(w)$ is a background model, that describes the general statistics of the document collection

$$p_B(w) = \frac{1}{N}\sum_{d\in\mathbf{D}} n(w,d)$$

- λ is a mixing factor between background model and classes

Effect

The background model assigns higher probabilities to less informative terms. Such terms will tend to be assigned to the background model, and only rarely to some classes.

PLSA and LSA

Probability matrices

Let

$$\overline{\mathbf{U}} = [p(d_i|z_k)]_{i,k} \qquad \overline{\mathbf{V}} = [p(w_j|z_k)]_{j,k} \qquad \overline{\mathbf{\Sigma}} = \text{diag}[p(z_k)]_k$$

with $i = 1, \dots, D$, $j = 1, \dots, V$, $k = 1, \dots, T$

$$\mathbf{P} = [p(w_j, d_i)]_{j,i} = \left[\sum_{k=1}^{T} p(w_j|z_k)p(d_i|z_k)p(z_k)\right]$$

then $\mathbf{P} = \overline{\mathbf{U}\mathbf{\Sigma}\mathbf{V}}^T$

- Rows in $\overline{\mathbf{V}}$ represent the multinomial distribution associated to classes
- Rows in $\overline{\mathbf{U}}$ represent the distributions of classes on documents
- Items in $\overline{\mathbf{\Sigma}}$ represent class proportions
- Items in the product represent the conditional independence of w and d on z

Open problems

Limits of PLSA

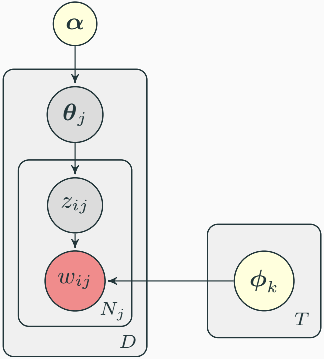
- Not a generative model: the corpus is modeled, but it is not clear how new documents can be modeled
- The set of parameters $p(w|z), p(d|z), p(z)$ to derive (learn) grows linearly with the size of the training set
- Lot of parameters: risk of overfitting

Latent Dirichlet Allocation

Latent Dirichlet Allocation

Definition

As for PLSA, a document is a mixture of latent classes (topics), defined as multinomial distributions of terms.



$$\begin{aligned}\boldsymbol{\theta}_j &\sim \mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha})\\ z_{ij} &\sim \mathrm{Mult}(z|\boldsymbol{\theta})\\ w_{ij} &\sim \mathrm{Mult}(w|\boldsymbol{\phi})\end{aligned}$$

Latent Dirichlet Allocation

Generative model

The model can be seen as a generative model, working as follows:

- For $j = 1$ to D
  - Sample a vector of classes proportions $\boldsymbol{\theta}_j \sim \text{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha})$ for document d_j
  - For $i = 1$ to N_j
    - Sample a class $z_{ij} \sim \text{Mult}(z|\boldsymbol{\theta}_j)$, $z_{ij} \in \{1, \dots, T\}$ for the l'i-th term in d_j
    - Sample the i-th term $w_{ij} \sim \text{Mult}(w|\boldsymbol{\phi}_{z_{ij}})$ in d_j

With respect to PLSA

$p(w|d)$ is not computed from an observed document, but rather sampled from a distribution: hence, d is a new document. The parameters to be determined are now $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_T)$ and $\boldsymbol{\phi}_{[1:T]} = [(\phi_{11}, \dots, \phi_{1N}), \dots, (\phi_{T1}, \dots, \phi_{TN})]$, with the last ones denoted in the following as a matrix $\boldsymbol{\Phi} \in \mathbb{R}^{T \times V}$, with $\boldsymbol{\phi}_i$ corresponding to row i of $\boldsymbol{\Phi}$

Latent Dirichlet Allocation

Parameters

- $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_T)$, parameter of the Dirichlet distributions related to the proportions of classes in a document
- $\boldsymbol{\Phi}$, where $\boldsymbol{\phi}_{ij}$ is the proportion of term $w_j \in V$ in the i-th class

Random variables

A random variable with Dirichlet distribution, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_T)$. Its instances $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_D)$ describe proportions of classes for each document d

$$p(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{i=1}^{T} \alpha_i)}{\prod_{i=1}^{T} \Gamma(\alpha_i)} \prod_{i=1}^{T} \theta_i^{\alpha_i - 1} = \frac{1}{\Delta(\boldsymbol{\alpha})} \prod_{i=1}^{T} \theta_i^{\alpha_i - 1}$$

Latent Dirichlet Allocation

Notes

- The value T (number of classes) is assumed known "a priori"
- The probabilities ϕ_{ki} are unknown but given "a priori" (and to be learned)
- The length $N_i$of a document is assumed known "a priori" and independent with respect to all the other variable or parameters of the model

Document length

The length of each document can also be modeled as a random variable, distributed according to some predefined distribution. For example,

$$N_i \sim \text{Poisson}(N|\xi)$$

However, learning ξ is assumed to be an independent task with respect to learning other parameters and latent variable values

Latent Dirichlet Allocation

Symbols

- $\mathbf{w} = (w_1, \ldots, w_N)$ multiset of terms in a document
- $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_D)$ set of terms in all documents in the corpus
- $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_V)$ distribution of terms in a class
- $\boldsymbol{\Phi} = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_T)^T$ set of distributions of all classes
- $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_T)$ distributions of classes for a document
- $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_D)^T$ set of distributions of classes in all documents
- $\mathbf{z} = (z_1, \ldots, z_N)$ latent variables, classes associated to terms in a document
- $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_D)$ latent variables, classes associated to terms in all documents
- $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_T)$ parameters of the Dirichlet generating the $\boldsymbol{\theta}$

Latent Dirichlet Allocation

Probabilities at term level

Joint

$$\begin{aligned} p(w, \boldsymbol{\theta}, z | \boldsymbol{\alpha}, \boldsymbol{\Phi}) &= p(w | z, \boldsymbol{\Phi}) p(z | \boldsymbol{\theta}) p(\boldsymbol{\theta} | \boldsymbol{\alpha}) \\ &= p(w | \boldsymbol{\phi}_z) p(z | \boldsymbol{\theta}) p(\boldsymbol{\theta} | \boldsymbol{\alpha}) \end{aligned}$$

Marginal

$$\begin{aligned} p(w | \boldsymbol{\alpha}, \boldsymbol{\Phi}) &= \int_{\boldsymbol{\theta}} p(\boldsymbol{\theta} | \boldsymbol{\alpha}) \sum_{z=1}^{T} p(w | z, \boldsymbol{\Phi}) p(z | \boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\theta}} p(\boldsymbol{\theta} | \boldsymbol{\alpha}) \sum_{z=1}^{T} p(w | \boldsymbol{\Phi}) p(z | \boldsymbol{\theta}) d\boldsymbol{\theta} \end{aligned}$$

Latent Dirichlet Allocation

Probabilities at document level

Joint

$$p(\mathbf{w}, \boldsymbol{\theta}, \mathbf{z}|\boldsymbol{\alpha}, \boldsymbol{\Phi}) = p(\boldsymbol{\theta}|\boldsymbol{\alpha}) \prod_{j=1}^{N} p(w_j|z_j, \boldsymbol{\Phi}) p(z_j|\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\boldsymbol{\alpha}) \prod_{j=1}^{N} p(w_j|\boldsymbol{\phi}_{z_j}) p(z_j|\boldsymbol{\theta})$$

Marginal

$$\begin{aligned} p(\mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\Phi}) = \prod_{j=1}^{N} p(w_j|\boldsymbol{\alpha}, \boldsymbol{\Phi}) &= \int_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\boldsymbol{\alpha}) \left(\prod_{j=1}^{N} \sum_{z_j=1}^{T} p(z_j|\boldsymbol{\theta}) p(w_j|z_j, \boldsymbol{\Phi}) \right) d\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\boldsymbol{\alpha}) \left(\prod_{j=1}^{N} \sum_{z_j=1}^{T} p(z_j|\boldsymbol{\theta}) p(w_j|\boldsymbol{\phi}_{z_j}) \right) d\boldsymbol{\theta} \end{aligned}$$

Latent Dirichlet Allocation

Probabilities at corpus level

Joint

$$
\begin{aligned}
p(\mathbf{W}, \boldsymbol{\Theta}, \mathbf{Z} | \boldsymbol{\alpha}, \boldsymbol{\Phi}) &= \prod_{i=1}^{D} p(\boldsymbol{\theta}_i | \boldsymbol{\alpha}) \prod_{j=1}^{N_i} p(w_{ij} | z_{ij}, \boldsymbol{\Phi}) p(z_{ij} | \boldsymbol{\theta}_i) \\
&= \prod_{i=1}^{D} p(\boldsymbol{\theta}_i | \boldsymbol{\alpha}) \prod_{j=1}^{N_i} p(w_{ij} | \boldsymbol{\phi}_{z_{ij}}) p(z_{ij} | \boldsymbol{\theta}_i)
\end{aligned}
$$

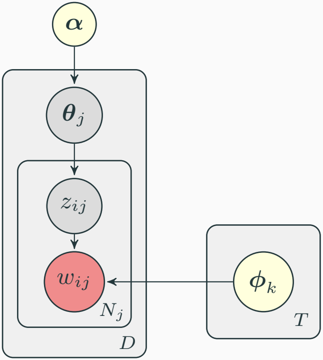
Marginal

$$
\begin{aligned}
p(\mathbf{W} | \boldsymbol{\alpha}, \boldsymbol{\Phi}) &= \prod_{i=1}^{D} p(\mathbf{w}_i | \boldsymbol{\alpha}, \boldsymbol{\Phi}) \\
&= \prod_{i=1}^{D} \int_{\boldsymbol{\theta}_i} p(\boldsymbol{\theta}_i | \boldsymbol{\alpha}) \left(\prod_{j=1}^{N_i} \sum_{z_{ij}=1}^{T} p(z_{ij} | \boldsymbol{\theta}_i) p(w_{ij} | z_{ij}, \boldsymbol{\Phi}) \right) d\boldsymbol{\theta}_i \\
&= \prod_{i=1}^{D} \int_{\boldsymbol{\theta}_i} p(\boldsymbol{\theta}_i | \boldsymbol{\alpha}) \left(\prod_{j=1}^{N_i} \sum_{z_{ij}=1}^{T} p(z_{ij} | \boldsymbol{\theta}_i) p(w_{ij} | \boldsymbol{\phi}_{z_{ij}}) \right) d\boldsymbol{\theta}_i
\end{aligned}
$$

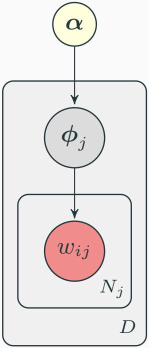
LDA and clustering

Different generative models

LDA



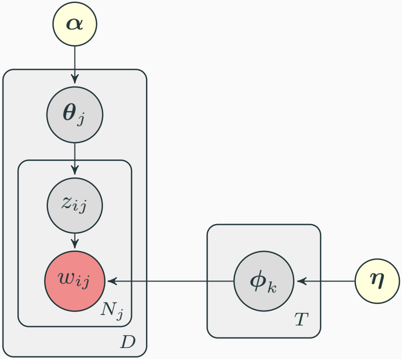
Clustering



Smoothed Latent Dirichlet Allocation

Definition

The $\boldsymbol{\phi}_k$ distributions are not considered as parameters, but rather as instances of a random variable with Dirichlet distribution



$$\boldsymbol{\phi}_k \sim \mathrm{Dir}(\boldsymbol{\phi}|\boldsymbol{\eta})$$
$$\boldsymbol{\theta}_j \sim \mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha})$$
$$z_{ij} \sim \mathrm{Mult}(z|\boldsymbol{\theta})$$
$$w_{ij} \sim \mathrm{Mult}(w|\boldsymbol{\phi})$$

Smoothed Latent Dirichlet Allocation

Generative model

- For $k = 1$ to T
    - Sample a multinomial $\boldsymbol{\phi}_k \sim$ Dir$(\boldsymbol{\phi}|\boldsymbol{\eta})$ for class k
- For $i = 1$ to D
    - Sample a vector of proportions of classes $\boldsymbol{\theta}_j \sim$ Dir$(\boldsymbol{\theta}|\boldsymbol{\alpha})$ for document d_j
    - For $i = 1$ to N_j
        - Sample a class $z_{ij} \sim$ Mult$(z|\boldsymbol{\theta})$ for the i-th term of d_j
        - Sample the i-term $w_{ij} \sim$ Mult$(w|\boldsymbol{\phi})$ of d_j

Smoothed Latent Dirichlet Allocation

Random variables

Two types of variables with Dirichlet distribution: $\boldsymbol{\theta}_i = (\theta_{i1}, \dots, \theta_{iT})$, distributions of classes for each document d_i

$$p(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{i=1}^{T} \alpha_i)}{\prod_{i=1}^{T} \Gamma(\alpha_i)} \prod_{i=1}^{T} \theta_i^{\alpha_i - 1} = \frac{1}{\Delta(\boldsymbol{\alpha})} \prod_{i=1}^{T} \theta_i^{\alpha_i - 1}$$

$\boldsymbol{\phi}_j = (\phi_{j1}, \dots, \phi_{jV})$, distributions of terms for each class j

$$p(\boldsymbol{\phi}|\boldsymbol{\eta}) = \frac{\Gamma(\sum_{i=1}^{V} \eta_i)}{\prod_{i=1}^{V} \Gamma(\eta_i)} \prod_{i=1}^{V} \phi_i^{\eta_i - 1} = \frac{1}{\Delta(\boldsymbol{\eta})} \prod_{i=1}^{V} \phi_i^{\eta_i - 1}$$

Smoothed Latent Dirichlet Allocation

Probability at term level

Joint

$$p(w, \boldsymbol{\theta}, \boldsymbol{\Phi}, z|\boldsymbol{\alpha}, \boldsymbol{\eta}) = p(w|z, \boldsymbol{\Phi})p(z|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\alpha})p(\boldsymbol{\Phi}|\boldsymbol{\eta})$$

Marginal

$$\begin{aligned} p(w|\boldsymbol{\alpha}, \boldsymbol{\eta}) &= \int_{\boldsymbol{\Phi}} \int_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\boldsymbol{\alpha})p(\boldsymbol{\Phi}|\boldsymbol{\eta}) \sum_{z=1}^{T} p(w|z, \boldsymbol{\Phi})p(z|\boldsymbol{\theta}) d\boldsymbol{\theta} d\boldsymbol{\Phi} \\ &= \int_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\boldsymbol{\alpha})p(\boldsymbol{\Phi}|\boldsymbol{\eta}) \sum_{z=1}^{T} p(w|\boldsymbol{\Phi})p(z|\boldsymbol{\theta}) d\boldsymbol{\theta} d\boldsymbol{\Phi} \end{aligned}$$

Smoothed Latent Dirichlet Allocation

Probability at document level

Joint

$$p(\mathbf{w}, \boldsymbol{\theta}, \mathbf{z} | \boldsymbol{\alpha}, \boldsymbol{\Phi}) = p(\boldsymbol{\theta}|\boldsymbol{\alpha}) \prod_{j=1}^{N} p(w_j | z_j, \boldsymbol{\Phi}) p(z_j|\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\boldsymbol{\alpha}) \prod_{j=1}^{N} p(w_j|\boldsymbol{\phi}_{z_j}) p(z_j|\boldsymbol{\theta})$$

Marginal

$$\begin{aligned} p(\mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\Phi}) &= \prod_{j=1}^{N} p(w_j|\boldsymbol{\alpha}, \boldsymbol{\Phi}) \\ &= \int_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\boldsymbol{\alpha}) \left(\prod_{j=1}^{N} \sum_{z_j=1}^{T} p(z_j|\boldsymbol{\theta}) p(w_j|z_j, \boldsymbol{\Phi}) \right) d\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\boldsymbol{\alpha}) \left(\prod_{j=1}^{N} \sum_{z_j=1}^{T} p(z_j|\boldsymbol{\theta}) p(w_j|\boldsymbol{\phi}_{z_j}) \right) d\boldsymbol{\theta} \end{aligned}$$

Smoothed Latent Dirichlet Allocation

Probability at corpus level

Joint

$$
\begin{aligned}
p(\mathbf{W}, \boldsymbol{\Theta}, \mathbf{Z} | \boldsymbol{\alpha}, \boldsymbol{\Phi}) &= \prod_{i=1}^{D} p(\boldsymbol{\theta}_i | \boldsymbol{\alpha}) \prod_{j=1}^{N_i} p(w_{ij} | z_{ij}, \boldsymbol{\Phi}) p(z_{ij} | \boldsymbol{\theta}_i) \\
&= \prod_{i=1}^{D} p(\boldsymbol{\theta}_i | \boldsymbol{\alpha}) \prod_{j=1}^{N_i} p(w_{ij} | \boldsymbol{\phi}_{z_{ij}}) p(z_{ij} | \boldsymbol{\theta}_i)
\end{aligned}
$$

Marginal

$$
\begin{aligned}
p(\mathbf{W} | \boldsymbol{\alpha}, \boldsymbol{\Phi}) &= \prod_{i=1}^{D} p(\mathbf{w}_i | \boldsymbol{\alpha}, \boldsymbol{\Phi}) \\
&= \prod_{i=1}^{D} \int_{\boldsymbol{\theta}_i} p(\boldsymbol{\theta}_i | \boldsymbol{\alpha}) \left(\prod_{j=1}^{N_i} \sum_{z_{ij}=1}^{T} p(z_{ij} | \boldsymbol{\theta}_i) p(w_{ij} | z_{ij}, \boldsymbol{\Phi}) \right) d\boldsymbol{\theta}_i \\
&= \prod_{i=1}^{D} \int_{\boldsymbol{\theta}_i} p(\boldsymbol{\theta}_i | \boldsymbol{\alpha}) \left(\prod_{j=1}^{N_i} \sum_{z_{ij}=1}^{T} p(z_{ij} | \boldsymbol{\theta}_i) p(w_{ij} | \boldsymbol{\phi}_{z_{ij}}) \right) d\boldsymbol{\theta}_i
\end{aligned}
$$

Smoothed Latent Dirichlet Allocation

Distribution

Probability of (generation of) a document $\mathbf{w}$ and of the entire corpus $\mathbf{W}$

$$p(\mathbf{w}|\boldsymbol{\alpha},\boldsymbol{\eta}) = \int_{\boldsymbol{\Phi}}\int_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\boldsymbol{\alpha})\left(\prod_{i=1}^{N}\sum_{z_i=1}^{T} p(z_i|\boldsymbol{\theta})p(w_i|z_i,\boldsymbol{\Phi})p(\boldsymbol{\Phi}|\boldsymbol{\eta})\right)d\boldsymbol{\theta}d\boldsymbol{\Phi}$$

$$p(\mathbf{W}|\boldsymbol{\alpha},\boldsymbol{\eta}) = \prod_{i=1}^{D}\int_{\boldsymbol{\Phi}}\int_{\boldsymbol{\theta}_i} p(\boldsymbol{\theta}_i|\boldsymbol{\alpha})\left(\prod_{j=1}^{N_i}\sum_{z_{ij}} p(z_{ij}|\boldsymbol{\theta}_i)p(w_{ij}|z_{ij},\boldsymbol{\Phi})p(\boldsymbol{\Phi}|\boldsymbol{\eta})\right)d\boldsymbol{\theta}_i d\boldsymbol{\Phi}$$

Inference in LDA

Idea

Two part process:

- Bayesian inference for the posterior distribution $p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\Phi})$
- estimation of the hyper-parameters $\boldsymbol{\alpha}, \boldsymbol{\Phi}$ by MLE

Problems

Determining the precise posterior is generally intractable. Need to use approximation methods.

- determining a different, but similar, distribution
- obtaining a collection of samples from the distribution

Dealing with posterior inference intractability

Approaches

Two different approaches

- Variational inference: applying variational (mean field) methods to EM. Faster, less precise
- Montecarlo sampling: applying MCMC (Markov Chain Montecarlo) sampling, in particular Gibbs sampling, of the posterior distribution. Slower, more precise

Posterior inference through variational methods

Idea

- Use of simpler distributions to approximate $p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\Phi})$
- Definition of a family of simpler distributions, parameterized by a set of variational parameters
- Search of the distribution in the family (and the corresponding variational parameters values) that best approximates $p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\Phi})$
- Approximation measured in terms of KL divergence

The inference problem (computing an integral) becomes now an optimization problem (maximization of a function)

Posterior inference through MCMC

- Sampling applied to estimate the distribution of $\mathbf{z}$
- $\mathbf{\Phi}, \boldsymbol{\theta}$ are then approximated from the posterior estimates of $\mathbf{z}$

- Gibbs sampling:
  - consider each word token w_{ij}
  - estimate the probability of assigning the word to each topic conditional on the topic assignments of all other words