LDA and Topic Modeling

Sunday, 2 April 2023 7:22 PM

Latent Dirichlet Allocation is a generative model.

In ML it is used for the task of Topic Modeling. Topic Modeling is clustering of documents into topics which are latent

Some Preliminaries

Multinomial Distribution

- Generalization of binomial distribution

Suppose we have vector x in R^K such that only one coordinate is 1 rest are 0.

If we denote $P(x_k = 1) = \mu_k$ distribution of x is given as

$$P(x|\mu) = \prod_{k=1}^{K} \mu_k^{x_k}$$

where $\mu = (\mu_1, \dots \mu_k)$, $\mu_k \geq 0$ & $\sum_k \mu_k = 1$

Now consider dataset D of N independent observations $x_1 \dots x_N$

Likelihood

$$p(D/\mu) = \prod_{i=1}^{N} \prod_{k=1}^{K} \mu_k^{x_{nk}}$$

$$= \prod_{i=1}^{K} \mu_k^{(\sum_n x_{nk})}$$

$$= \prod_{k=1}^{K} \mu_k^{m_k} \qquad \left(m_k = \sum_n x_{nk}\right)$$

(no. of observations of $x_k = 1$)

Now the Maximum Likelihood solution for μ is

$$\mu_k^{ML} = \frac{m_k}{N}$$

We consider the joint distribution of quantities $m_1 \dots m_K$ conditioned on μ and N

$$Mult(m_1, m_2, \dots m_k / \mu, N) = \binom{N}{m_1 m_2 \dots m_K} \prod_{k=1}^{K} \mu_k^{m_k}$$

where

$$\binom{N}{m_1 m_2 \dots m_K} = \frac{N!}{m_1! \, m_2! \dots m_k!}$$

& $\sum_{k=1}^{K} m_k = N$

This is the multinomial distribution

Dirichlet Distribution :-

Conjugate Prior for Multinomial distribution

$p(\mu|\alpha)$ varies as $\prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$

where $0 \leq \mu_k \leq 1$

$\sum_k \mu_k = 1$

$\alpha_1, \dots \alpha_K$ are the parameters of the distribution

$$\alpha = (\alpha_1, \dots \alpha_K)$$

The Dirichlet distribution is given as

$$Dir(\mu|\alpha) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \dots \Gamma(\alpha_k)} \prod_{i=1}^{K} \mu_k^{\alpha_k - 1}$$

$$\alpha_0 = \sum_{k=1}^{K} \alpha_k$$

$$\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du$$

(extension of factorial function for reals)

$$\Gamma(x+1) = x\Gamma(x)$$

$$\Gamma(1) = 1$$

Now

$p(\mu|D, \alpha)$ (alpha)

$\propto$ (varies symbol) $\; p(D|\mu) \, p(\mu|\alpha)$

$\propto \prod_{k=1}^{K} \mu_k^{\alpha_k + m_k - 1}$

Thus

$$p(\mu|D, \alpha) = Dir(\mu|\alpha + m)$$

$$= \frac{\Gamma(\alpha_0 + N)}{\Gamma(\alpha_1 + m_1) \dots \Gamma(\alpha_K + m_k)} \prod_{k=1}^{K} \mu_k^{\alpha_k + m_k - 1}$$

$$m = (m_1, \dots m_K)$$

Topic Modeling & LDA

Given a corpus of documents, we consider that each word in each document has a latent assignment into one of K topics. LDA is standard algo. for the task

Each document is a mixture of topics.

Each topic is a probability distribution over words.

High Level Idea as Generative Model :-

Each document is generated by choosing a topic

Each word is generated using particular topic

Notations & Algorithm :- (From a blog by Aaron Mora)

D - Documents
V - Vocabulary / Dictionary

We represent a document d with a vector $w_d \in N^{n_d}$ where n_d is number of terms in document d.

$$w_d = (w_{d1}, w_{d2}, \dots w_{dn_d})$$

where w_{d_i} stands for index of word in position i of document in the V.

Let K - number of latent topics in corpus

θ_d a vector in K-dimensional simplex Δ^{K-1} i.e. $\theta_{dk} \in [0,1]$ → Share of topic k in doc. d.

$\beta_k \in \Delta^{V-1}$ vector of term probabilities in topic k

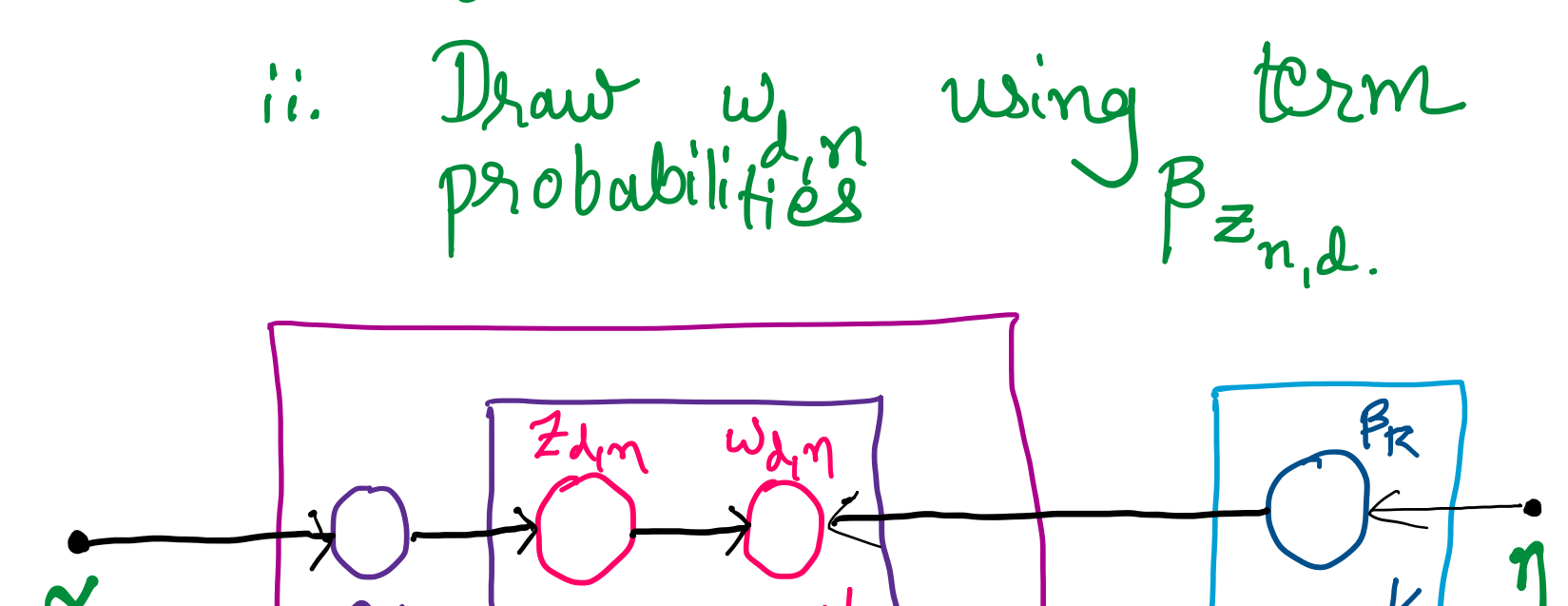
β_{kv} is the probability of observing word v in topic k

Consider the latent variable

$z_{d,n} \in (1, \dots K)$ is topic assigned to n^{th} word in document d.

The Algorithm

1. Draw K vectors $\beta_k \in \Delta^{V-1}$ from $Dir(\beta_k|\eta)$
2. Draw D vectors $\theta_d \in \Delta^{k-1}$ from $Dir(\theta_d|\alpha)$
3. Each word $w_{d,n}$ in doc d is generated in two steps:
   - i. Draw $z_{n,d} \in (1, \dots, K)$ according to topic probabilities θ_d.
   - ii. Draw $w_{d,n}$ using term probabilities $\beta_{z_{n,d}}$.



The posterior for observed corpus $w_{d,n}$ given η, α :-

$$p(z, \theta, \beta | w, \alpha, \eta) = \left[\prod_{d,n} p(w_{d,n}|z_{d,n}, \beta)\right] \left[\prod_{d,n} p(z_{d,n}|\theta_d)\right] \times \left[\prod_d p(\theta_d|\alpha)\right] \left[\prod_k p(\beta_k|\eta)\right]$$

β is $V \times K$ matrix of term probabilities with column k given by β_k

Θ is $K \times D$ matrix of topic memberships with column d given by θ_d

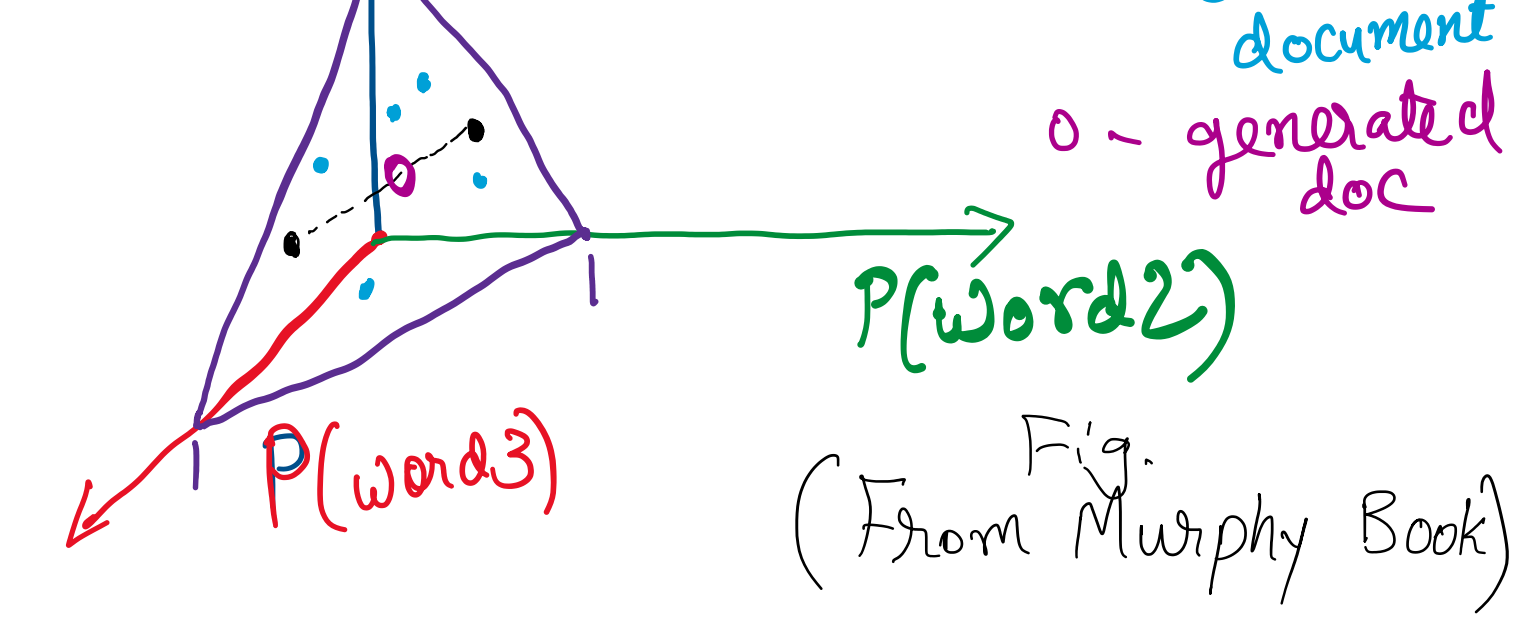


Fig. (From Murphy Book)