Latent Dirichlet Allocation

1. Introduction

text corpora 모델링의 목표는 통계적인 관계는 보존하면서 큰 용량의 collections의 짧은 description를 찾는 것이다. 이를 위해 IR(Information Retrieval) 분야에서는 많은 시도가 이루어졌는데, 그 중 하나가 tf-idf이다. tf-idf는 term frequency / document frequency을 통해서 문서 내에서 뿐만 아니라 문서 간에 단어가 가지고 있는 중요도를 수치화한다. 하지만 tf-idf는 description를 많이 해주지 못하고 문서 내, 문서 간 통계적 구조를 거의 밝히지 못한다. 이러한 문제를 해결하기 위해 IR 연구자들은 Latent Semantic Indexing(LSI)를 제안하였다.

LSI는 데이터의 가장 많은 variance을 가지는 tf-idf 변수의 선형 공간을 알아내기 위해 matrix X에 대해 SVD를 수행하므로 위의 tf-idf의 단점을 극복할 수 있다고 연구자들이 말했다.

generative model과 관련해서 LSI의 대안으로 나온 것이 probabilistic LSI(pLSI)이다. pLSI는 문서 내의 각 단어를 mixture model로부터의 sample로 보는데, 그 mixture components들은 'topics'로 볼 수 있는 multinomial random variables이다. 따라서 각 단어는 single topic으로부터 생성되고 문서 내의 다른 단어들은 다른 topic으로부터 생성 될 것이다.

pLSI가 확률에 기반한 텍스트 모델링에서 중요한 발전이긴 하지만, 문서에 대한 확률 모델링을 제공하지 않는다는 점에서 불완전하다고 할 수 있다. pLSI에서 각 문서는 mixing proportions for topics(list of numbers)로 제시되고 이 숫자에 대한 generative probabilistic model은 존재하지 않는다. 이러한 점은 다음과 같은 문제를 야기한다 (1) corpus의 크기가 증가할수록 모델의 모수가 선형적으로 증가하고 이는 overfitting의 문제로 이어진다. (2) training set 밖에서 문서에 확률을 어떻게 정할지 clear하지 않다.

pLSI의 단점을 어떻게 극복하기 위해, LSI나 pLIS를 포함하는, 차원축소 방법을 기반인 fundamental probabilistic assumptions을 살펴보자. 이러한 방법들은 bag-of-words 가정에 기반하는데, 이는 문서 내에서 단어의 순서가 무시될 수 있다는 것이다. probability theory에서는 이는 exchangeability 가정이다.

Finetti는 이를 theorem으로 정리하였는데, exchangeable random variables가 mixture distribution(보통 infinite)을 가진다고 말하였다. 이를 이해하기 위해 exchangeable random variables의 정의부터 살펴보자.

Definition Formally, an exchangeable sequence of random variables is a finite or infinite

sequence $X_1, X_2, X_3, \cdots$ of random variables such that for any finite permutation σ^2 of the indices 1, 2, 3, ..., (the permutation acts on only finitely many indices, with the rest fixed), the joint probability distribution of the permuted sequence, $X_{\sigma(1)}, X_{\sigma(2)}, X_{\sigma(3)}, \cdots$ is the same as the joint probability distribution of the original sequence.[1]

exchangeable random variables는 iid random variables와 밀접하게 연관되어 있다. iid 확률변수는 어떤 분포 형태를 조건으로 하면, 곧 exchangeable하다. $f(\boldsymbol{x}) = f(x_1, \cdots, x_n) = f(x_1) \times \cdots \times f(x_n) = f(x_{\sigma(1)}) \times \cdots \times f(x_{\sigma(n)})$. 또한 그 역도 성립한다. 예시는 bivariate normal distribution이 있다.
이러한 exchangeability에 기반한 de Finetti의 theorem을 살펴보면 아래와 같다.

Theorem De Finetti's theorem states that the probability distribution of any infinite exchangeable sequence of Bernoulli random variables is a "mixture" of the probability distributions of independent and identically distributed sequences of Bernoulli random variables. "Mixture", in this sense, means a weighted average, but this need not mean a finite or countably infinite (i.e., discrete) weighted average: it can be an integral rather than a sum.[2]

위의 논의를 종합해보면, 기본적으로 bags-of-words 가정은 문서내에서 단어들의 순서가 무시될 수 있으며 이를 probability thoery의 정의로는 exchangeability라고 한다. 그런데, exchangeable random variables는 De Finetti's에 의하면 mixture distribution으로 표현될 수 있다. 이 논리가 바로 저자들이 LDA 모델을 생각하게 된 경위이다. 중요한 점은 exchangeable r.v가 iid r.v와 같은 가정이 아니라는 것이다. 오히려, 어떤 내재하는 latent parameter of a probability distribution을 조건으로 하는 'conditionally independent and identically distributed'라는 표현이 맞을 것이다(De Finetti's theorem에 나와있다.) 따라서 exchangeability라는 가정이 있으면, 텍스트 모델링에서 단어들의 joint 분포가 어떤 분포를 조건으로 한 상태에서 iid이므로 marginal product로 표현될 수 있을 것이다.

2. Notation and terminology

- A word is the basic unit of discrete data, defined to be an item from a vocabulary indexed by $\{1, \cdots, V\}$. We represent words using unit-basis vectors that have a

[1] https://en.wikipedia.org/wiki/Exchangeable_random_variables
[2] https://en.wikipedia.org/wiki/De_Finetti%27s_theorem

single component equal to one and all other components equal to zero. Thus, using superscripts to denote components, the vth 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 $\boldsymbol{w} = (w_1, \cdots, w_N)$, where w_n is the nth word in the sequence.
- A corpus is a collection of M documents denoted by $D = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_M\}$.

3. Latent Dirichlet Allocation

기본 아이디어는 문서들이 latent topics의 random mixtures로 표현된다는 것인데, 각각의 topic은 latent multinomial variables로 간주된다. LDA는 corpus D에서 각각의 문서 $\boldsymbol{w}$에 대해 다음의 generative process을 가정한다.

1. Choose $N \sim Poisson(\xi)$.
2. Choose $\theta \sim Dir(\alpha)$.
3. For each of the N words w_n:
   (a) Choose a topic $z_n \sim Multinomial(\theta)$.
   (b) Choose a word w_n from $p(w_n \mid z_n, \beta)$, a multinomial probability conditioned on the topic z_n

여기서 word probabilities는 $k \times V$ matrix인 β으로 parameterized 되는데 $\beta_{ij} = p(w^j = 1 \mid z^i = 1)$이다. 또한 Dirichlet 분포의 차원인 k가 고정되어 있다고 가정한다.
α, β의 모수가 주어진 상태에서 topic mixture인 θ, N개의 topic인 $\boldsymbol{z}$, N개의 단어인 $\boldsymbol{w}$의 joint 분포는 아래와 같다.

$$\begin{aligned} p(\theta, \boldsymbol{z}, \boldsymbol{w} \mid \alpha, \beta) &= p(\boldsymbol{z}, \boldsymbol{w} \mid \theta, \alpha, \beta) p(\theta \mid \alpha, \beta) \\ &= p(\theta \mid \alpha) p(\boldsymbol{w} \mid \boldsymbol{z}, \theta, \alpha, \beta) p(\boldsymbol{z} \mid \theta, \alpha, \beta) \\ &= p(\theta \mid \alpha) p(\boldsymbol{w} \mid \boldsymbol{z}, \beta) p(\boldsymbol{z} \mid \theta) \\ &= p(\theta \mid \alpha) \prod_{n=1}^{n} p(z_n \mid \theta) p(w_n \mid z_n, \beta) \end{aligned}$$

이를 $\boldsymbol{z}$에 대해서 모두 더하고, θ에 대해서 적분을 하면 문서 $\boldsymbol{w}$에 대한 marginal 분포를 얻는다.

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

마지막으로 문서들의 확률을 곱한다면 corpus의 확률을 얻는다.

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

이러한 LDA의 가장 큰 장점은 하나의 문서가 여러개의 topics으로 구성될 여지가 있다는 것이다. 하나의 문서는 여러개의 단어로 구성되고 각 단어들은 Dirichlet 분포에서 매번 topic을 뽑기 때문에, 여러개의 단어로 구성된 하나의 문서가 여러개의 topics으로 구성된다.

3.1 LDA and exchangeability

LDA에서는 단어들이 topics에 의해서 생성되고 ($p(w_n \mid z_n, \beta)$에서 생성) 그 topics들은 문서 내에서 infinitely exchangeable하다 (더 생각). De Finetti's의 theorem에 따르면 exchangeable observations은 어떤 latent variable가 주어졌을 때 conditionally independent하다. 만약 이 latent variable을 topics에 대한 multinomial의 random parameter로 한다면, 단어들은 topics을 조건으로 할 때, conditionally independent하고 topics도 θ을 조건으로 할 때, conditionally indenpendent하다. 따라서 아래와 같이 단어와 topics의 분포를 유도할 수 있다.

$$\begin{aligned}
p(\boldsymbol{w}, \boldsymbol{z}) &= \int p(\boldsymbol{w}, \boldsymbol{z}, \theta) d\theta \\
&= \int p(\theta) p(\boldsymbol{w}, \boldsymbol{z} \mid \theta) d\theta \\
&= \int p(\theta) p(\boldsymbol{z} \mid \theta) p(\boldsymbol{w} \mid \boldsymbol{z}, \theta) d\theta \\
&= \int p(\theta) \left(\prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n) \right) d\theta \ \because De\ Finetti's\ Thm
\end{aligned}$$

3.2 A continuous mixture of unigram

LDA는 hidden topic variable인 z에 대해서 marginalize함으로써 two-level model로 이해할

수 있다. 단어가 $p(w \mid \theta, \beta)$로 분포되어 있다고 하자.

$$p(w \mid \theta, \beta) = \sum_z p(w \mid z, \beta) p(z \mid \theta)$$

이제 문서 $\boldsymbol{w}$에 대해서 다음의 generatvie process을 정의한다.

1. Choose $\theta \sim Dir(\alpha)$.
2. For each of the N words w_n:
   (a) Choose a word w_n from $p(w_n \mid \theta, \beta)$.

이러한 과정은 문서의 marginal 분포를 continuous mixture 분포로 정의한다.

$$p(\boldsymbol{w} \mid \alpha, \beta) = \int p(\theta \mid \alpha) \left(\prod_{n=1}^{N} p(w_n \mid \theta, \beta) \right) d\theta$$

여기서 $p(w_n \mid \theta, \beta)$는 mixture components이고 $p(\theta \mid \alpha)$는 mixture weights이다.

4. Relationship with other latent variable models

4.1 Unigram model

unigram model에서는 모든 단어들이 하나의 multinomial 분포의 표본이다.

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

4.2 Mixture of unigrams

unigram model의 가정을 확장하여 latent variable인 discrete random topic variable z을 가정한다. 이를 통해 mixture of unigrams model로 확장한다. 이 모델 하에서는 topic z을 우선 뽑은 후에, 그 topic을 조건으로 하는 conditional multinomial 분포인 $p(w \mid z)$에서 N개의 단어를 뽑는다.

$$p(\boldsymbol{w}) = \sum_z p(z) \prod_{n=1}^{N} p(w_n \mid z)$$

4.3 Probabilistic latent semantic indexing

pLSI에서는, topic z가 주어지면 문서 라벨 d와 단어 w_n가 conditionally independent하다.

$$\begin{aligned} p(d, w_n) &= \sum_z p(d, w_n, z) \\ &= \sum_z p(d)p(w_n, z \mid d) \\ &= \sum_z p(d)p(w_n \mid z, d)p(z \mid d) \\ &= p(d) \sum_z p(w_n \mid z)p(z \mid d) \end{aligned}$$

여기서 $p(z \mid d)$가 특정 문서 d에 대해서 mixture weights of the topics의 역할을 하므로 한 문서가 여러 topics을 포함할 수 있다. 그러나 d는 training set의 문서 라벨이라는 점이 중요하다. 따라서 모델이 새로운 문서에 대해서 확률을 할당할 방법이 마땅하지 않다.
하지만 LDA에서는 이 mixture weights을 k-parameter hidden random variable(multinomial 분포)로 간주하여 randomness을 부여한다. 따라서 pLSI처럼 training set에만 연관되어 있지 않고 overfitting으로부터 자유롭다.

5. Inference and Parameter Estimation

5.1 Inference

LDA에서 핵심이 되는 문제는 hidden variables의 posterior 분포의 계산이다.

$$p(\theta, \boldsymbol{z} \mid \boldsymbol{w}, \alpha, \beta) = \frac{p(\theta, \boldsymbol{z}, \boldsymbol{w} \mid \alpha, \beta)}{p(\boldsymbol{w} \mid \alpha, \beta)}$$

게다가 hidden variable에 대해서 marginalize하면 아래와 같은 분포를 얻는데, 여기서 θ와 β가 곱해져 있는 부분 때문에 다루기 어렵다.

$$\begin{aligned} p(\boldsymbol{w} \mid \alpha, \beta) &= \int p(\boldsymbol{w}, \theta \mid \alpha, \beta) d\theta \\ &= \int p(\theta \mid \alpha, \beta)p(\boldsymbol{w} \mid \theta, \alpha, \beta) d\theta \\ &= \int \frac{1}{B(\alpha)} \left(\prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \right) \left(\prod_{n=1}^{N} \sum_{i=1}^{k} \prod_{j=1}^{V} (\theta_i \beta_{ij})^{w_n^j} \right) d\theta \end{aligned}$$

정확한 추론을 위해서 정확한 posterior 분포를 유도하는 것은 어렵지만, 여러 근사 방법들이 제안되었다. 해당 논문에서는 Variational inference을 통한 근사를 살펴본다.

5.2 Variational Inference

기본 아이디어는 Jensen's inequality을 이용해서 log likelihood의 lower bound을 얻는 것인데, 이 lower bound는 variational parameters로 index된다. variational parameters은 가능한 타이트하게 lower bound을 찾는 optimization에 의해서 선택된다.

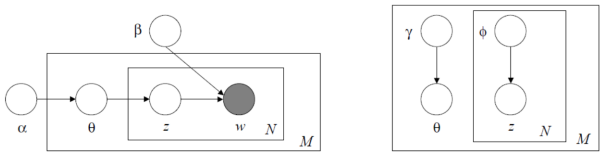

Figure 5: (Left) Graphical model representation of LDA. (Right) Graphical model representation of the variational distribution used to approximate the posterior in LDA.

왼쪽 그림에서 $\theta, \boldsymbol{z}, \boldsymbol{w}$ 사이에 있는 edges로 인해서 β, θ간의 coupling 문제가 발생한다. 따라서 이러한 edges와 $\boldsymbol{w}$ 노드를 없애고 free variational parameters을 부여함으로써 왼쪽 그림과 같이 latent variables에 대한 분포를 얻는다. 이 family는 variational 분포에 의해서 특징된다.

$$\begin{aligned} q(\theta, \boldsymbol{z} \mid \gamma, \phi) &= q(\theta \mid \gamma) q(\boldsymbol{z} \mid \phi) \ (\theta, \boldsymbol{z} \ are \ indep \ in \ variational \ distribution) \\ &= q(\theta \mid \gamma) \prod_{n=1}^{N} q(z_n \mid \phi_n) \end{aligned}$$

여기서 γ 는 Dirichlet에 대한 모수이고 ϕ는 multinomial 모수인데, 모두 free variational 모수이다.

이제 모수 γ, ϕ을 추정하기 위해 optimization 문제가 남아있다. 이는 곧, log likelihood의 tight lower bound을 찾는 것이고 이는 다시 아래 optimization 문제로 바꿔서 표현할 수 있다(Appendix A 참조)

$$(\gamma^*, \phi^*) = argmin_{(\gamma,\phi)} D(q(\theta, \boldsymbol{z} \mid \gamma, \phi) \mid\mid p(\theta, \boldsymbol{z} \mid \boldsymbol{w}, \alpha, \beta)).$$

즉, variational parameters을 찾는 것은 variational distribution $q(\theta, \boldsymbol{z} \mid \gamma, \phi)$과 true posterior $p(\theta, \boldsymbol{z} \mid \boldsymbol{w}, \alpha, \beta)$간의 KL divergence을 최소화하는 문제이다. 이 최소화 문제는 iterative fixed-point method을 통해서 달성될 수 있다. 특히, Appendix A.3을 통해서 KL divergence

의 미분 값을 계산하고 0으로 두면, 아래와 같은 업데이트를 얻을 수 있음을 알 수 있다.

$$\phi_{mi} \propto \beta_{iw_n} exp\left\{E_q\left[log(\theta_i) \mid \gamma\right]\right\}$$

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

Appendix A.1에서 알 수 있듯이, multinomial에 대한 기댓값은 아래와 같이 업데이트 된다.

$$E_q\left[log(\theta_i) \mid \gamma\right] = \Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{k} \gamma_j\right)$$

여기서 Ψ는 $log\Gamma$의 일차 도함수이다.

5.3 Parameter estimation

해당 단원에서는 empirical Bayes 방법을 사용한 LDA 모델의 모수 추정을 제시한다. 특히, corpus $D = \{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_M\}$가 주어질 때, 데이터의 log likelihood을 최대로 하는 모수 α, β를 찾고 싶다.

$$l(\alpha, \beta) = \sum_{d=1}^{M} logp(\boldsymbol{w}_d \mid \alpha, \beta)$$

위에서 설명했듯이, $p(\boldsymbol{w}_d \mid \alpha, \beta)$는 θ, β의 coupling으로 인해 계산하기 쉽지 않다. 하지만 variational inference은 α, β에 대해서 최대화할 수 있는 log likelihood의 lower bound을 제공한다. 따라서 empirical Bayes 추정치의 근사값을 variational EM을 통해 얻을 수 있는데, 이는 lower bound을 variational 모수인 γ, ϕ에 대해서 최대화하고, 고정된 γ, ϕ에 대해서 lower bound을 최대화하는 α, β을 찾는다. 자세한 variational EM은 Appendix A.4에 있다.