Insight campus SeSAC

토픽 모델링 (Topic Modeling)

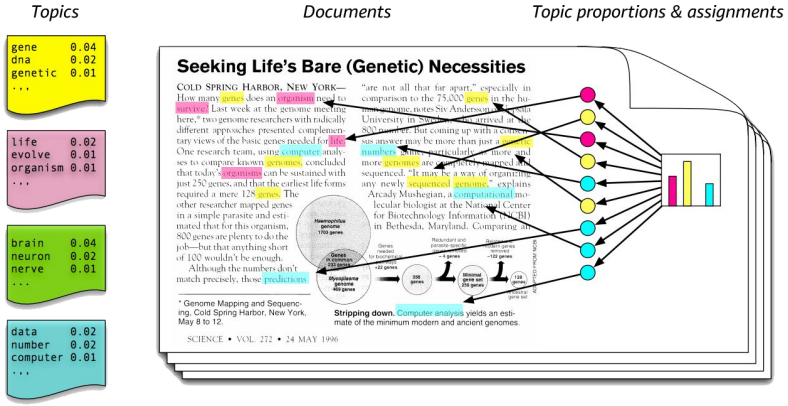
실무형 인공지능 자연어처리

3 잠재 디리클레 할당 (LDA)

Latent Dirichlet Allocation

잠재 디리클레 할당 (LDA) (1)

- 잠재 디리클레 할당(LDA, Latent Dirichlet Allocation)란 주어진 문서에 대해 어떤 주제가 존재하는지에 대한 확률모형 (토픽모델링)
- LDA는 토픽별 단어의 분포, 문서별 토픽의 분포를 추정
- 문서가 생성될 확률인 사후분포에 기반한 변수를 추론하여 텍스트 내에 숨겨져 있는 주제를 찾아내는 방식
- 결과적으로 전체 텍스트 문서 집합의 주제(토픽)들, 각 텍스트 문서별 주제의 확률, 각 단어들이 각 주제에 포함될 확률을 도출(디리클레:확률분포명칭)

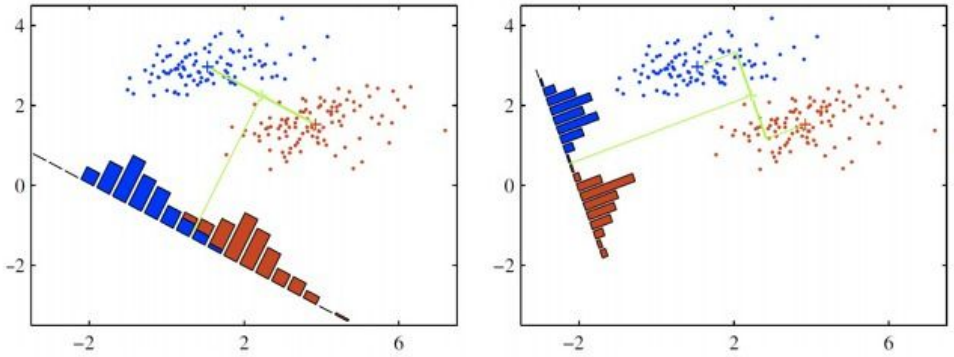


잠재 디리클레 할당 (LDA) (2)

- 잠재(Latent) : 사전적인 의미는 “잠재적인, 숨어 있는”. 우리가 직접 관찰할 수 있는 것은 문서 내용뿐. α, β, θ, z는 모두 감춰진 파라미터

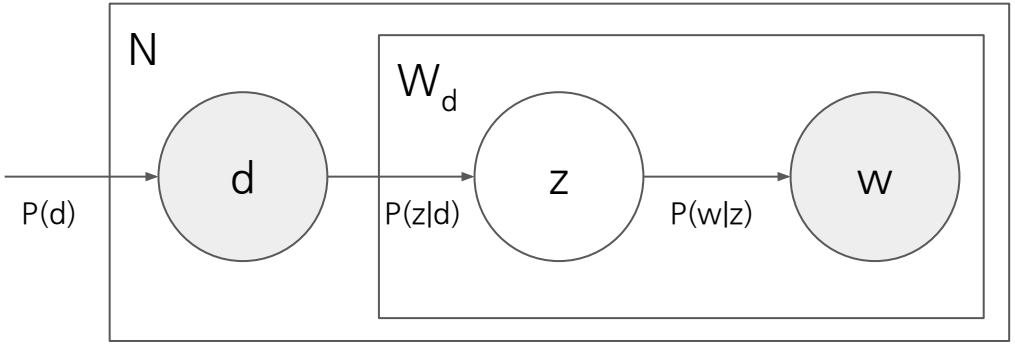
- 디리클레(Dirichlet) : 19세기 독일 수학자의 이름. 디리클레 분포(Dirichlet Distribution)를 사용하고 있음.
(θ를 결정할 때 α를 파라미터로 디리클레 분포를 사용)

- 할당(Allocation) : ‘할당’. 각 단어를 결정할 때, θ에 대한 다항 분포(Multinomial Distribution)로 주제를 ‘할당’한 뒤 그 주제로부터 단어를 추출.



확률적 잠재 의미 분석 (pLSA)

- 토픽모델링을 위해 잠재 의미 분석에서 사용하는 **특이값 분해가 아닌 확률적 방법을 사용**
- 토픽 모델링 가정 **"문서는 여러 주제로 구성되어 있고, 각 주제는 단어 집합으로 구성된다."**



N : 문서집합

W_d : 문서내 단어집합

d : 문서

z : 주제

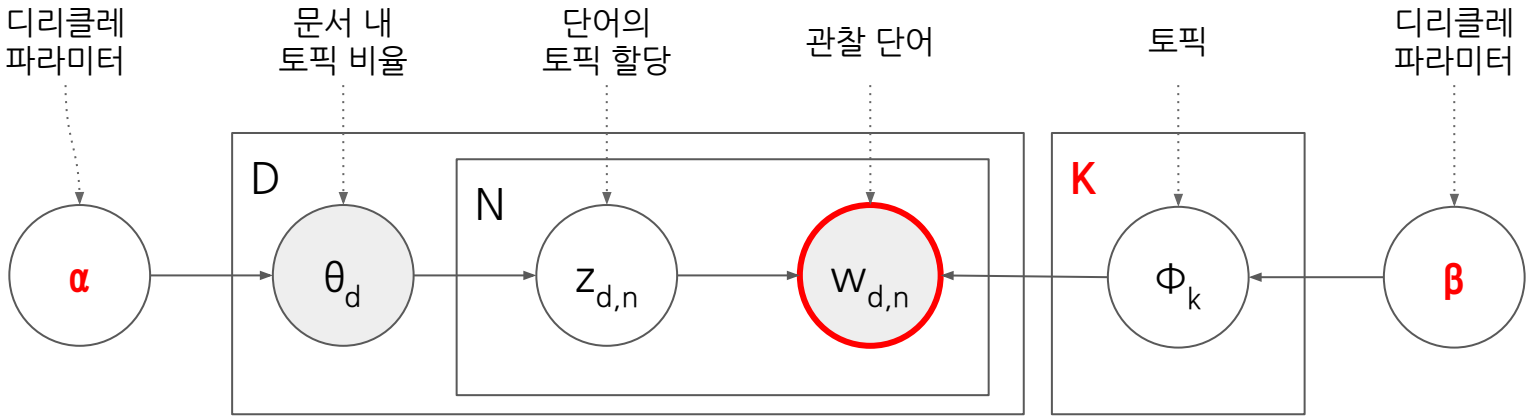
w : 단어

문서(d)가 주어지면 주제(z)는 확률 P(z|d)로 문서 내에 존재한다.

주제(z)가 주어지면 단어(w)는 확률 P(w|z)로 주제 내에 존재한다.

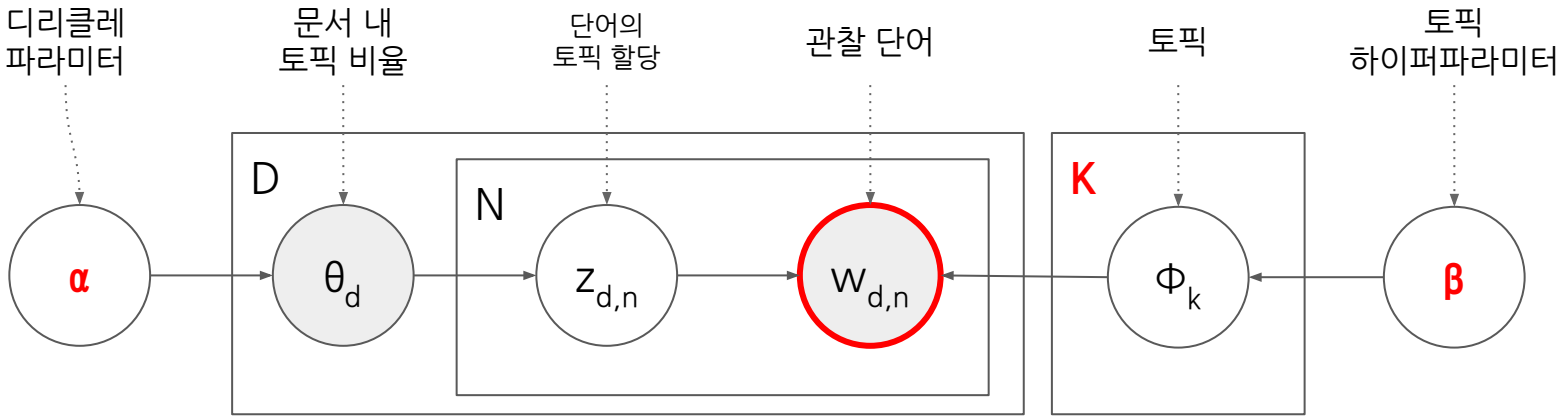
잠재 디리클레 할당 모델 (1)

α	디리클레 파라미터 (보통 0.1)	D	전체 문서 갯수
θ_d	문서 내 토픽 비율	Φ_k	토픽
$z_{d,n}$	단어의 토픽 할당	K	토픽수
$w_{d,n}$	관찰단어	β	토픽 하이퍼파라미터 (보통 0.001)
N	N은 d번째 문서의 단어 수		

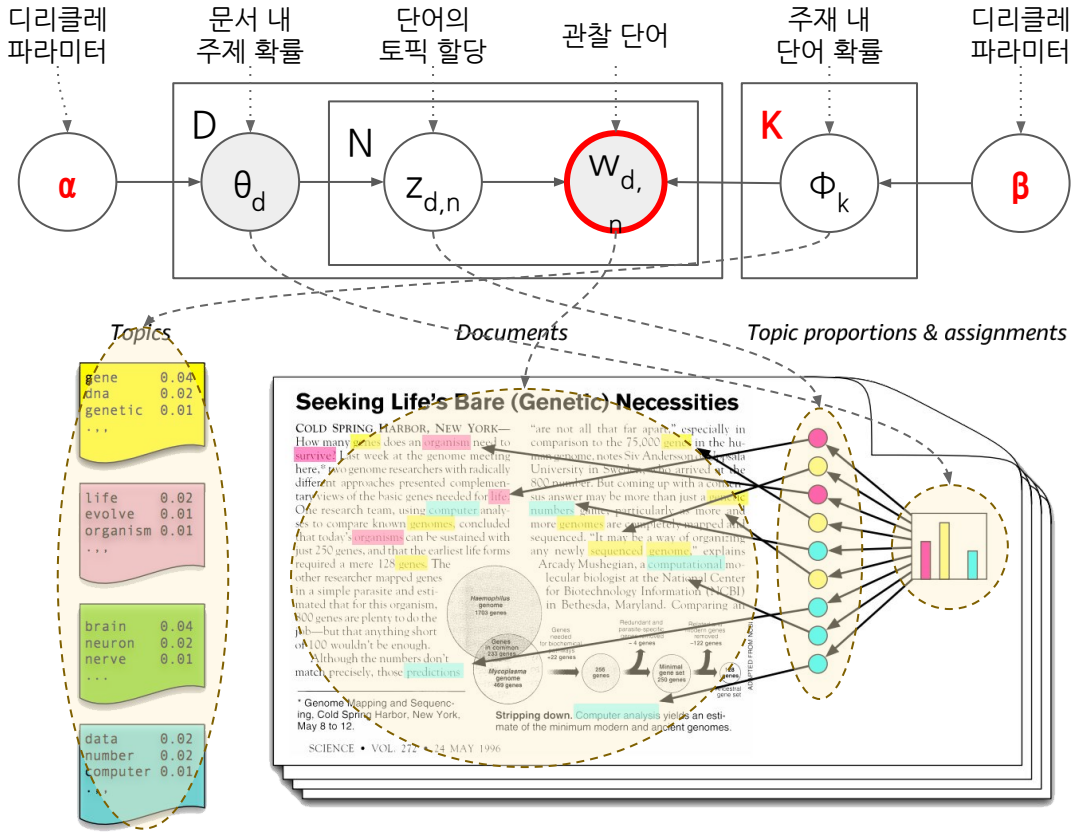


잠재 디리클레 할당 모델 (2)

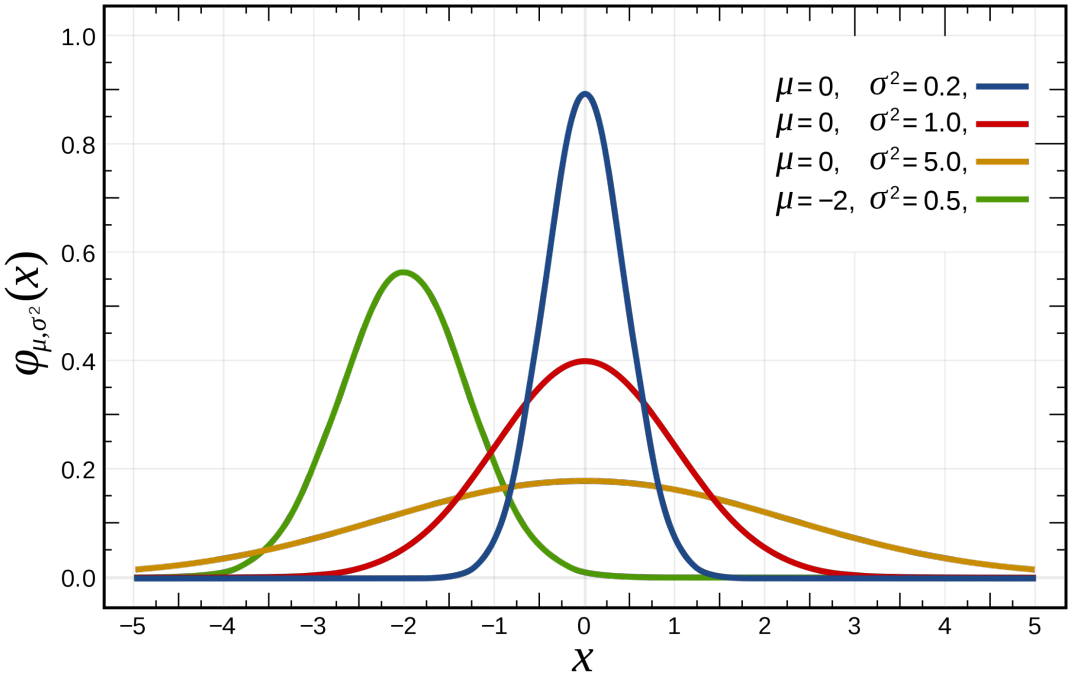
- 관찰 가능한 변수는 d번째 문서에 등장한 n번째 단어 $w_{d,n}$가 유일
- 이 정보를 가지고 하이퍼파라미터(사용자 지정) α,β를 제외한 모든 잠재 변수를 추정
- 사전에 결정해주어야할 값은 α, β, K값
  - 보통 **α은 0.1, β는 0.001로 사용**



잠재 디리클레 할당 모델 (2)



확률 분포의 파라미터 (예. 정규분포)



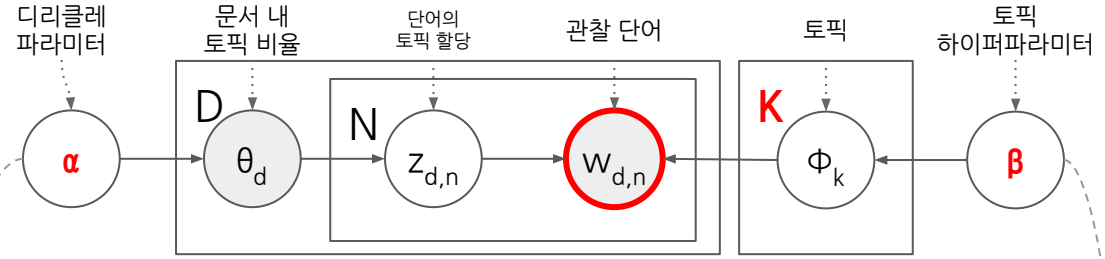
https://ko.wikipedia.org/wiki/%EC%A0%95%EA%B7%9C_%EB%B6%84%ED%8F%AC

확률론과 통계학에서, 정규 분포(normal distribution) 또는 가우스 분포(Gaussian distribution)는 연속 확률 분포의 하나이다. 정규분포는 수집된 자료의 분포를 근사하는 데에 자주 사용되며, 이것은 중심극한정리에 의하여 독립적인 확률변수들의 평균은 정규분포에 가까워지는 성질이 있기 때문이다.

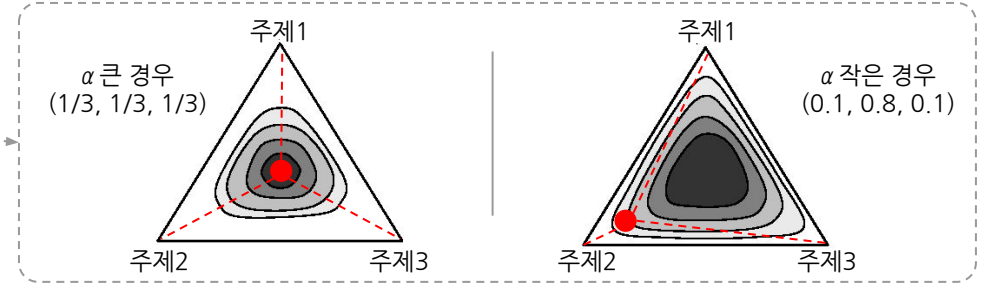
정규분포는 2개의 매개 변수 평균 u 과 표준편차 sigma 에 대해 모양이 결정되고, 이때의 분포를 N(u ,sigma ^{2})로 표기한다. 특히, 평균이 0이고 표준편차가 1인 정규분포 N(0,1)을 표준 정규 분포 (standard normal distribution)라고 한다.

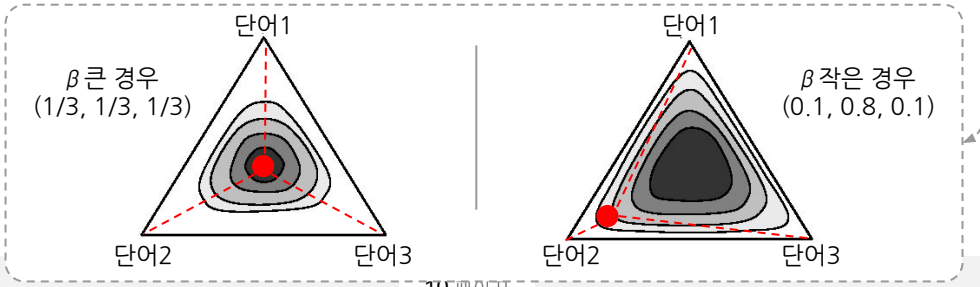
파라미터에 따라 확률 분포가 달라짐

잠재 디리클레 할당 모델 (3)



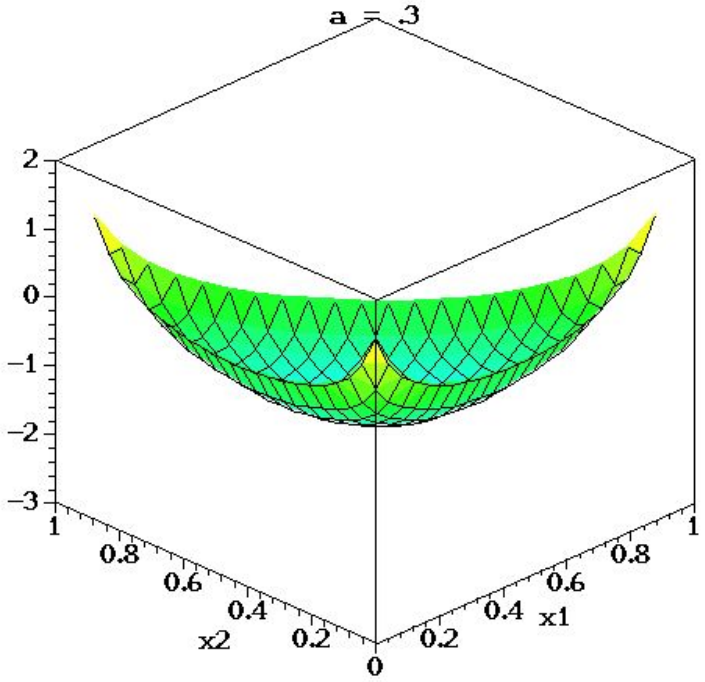
문서 내 여러 주제가 존재할 확률이 높음





주제 내 단어는 특정 주제 하나에 할당될 확률이 높음

잠재 디리클레 할당 모델 (3)



잠재 디리클레 할당 가정

- 잠재 디리클레 할당 가정 : **단어 교환성(exchangeability)**
  - BoW라고 표현하기도 함. 단어 교환성은 단어들 순서를 고려하지 않고 유무만 중요하다는 가정
  - '나는 양념 치킨을 좋아해 하지만 후라이드 치킨을 싫어해', '나는 후라이드 치킨을 좋아해 하지만 양념 치킨을 싫어해' 간에 차이가 없다고 생각
  - 단어 빈도수만으로 표현이 가능. 이를 기반으로 교환성을 포함하는 모형을 제시한 것이 바로 LDA

- 단순히 단어 하나를 단위로 생각하는 것이 아니라 단어 묶음을 한 단위로 생각하는 **n-gram방식으로 LDA의 교환성 가정을 확장**시킬 수도 있음

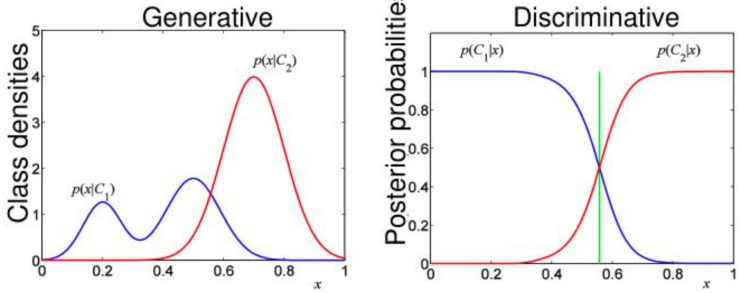
Generative / Discriminative Model

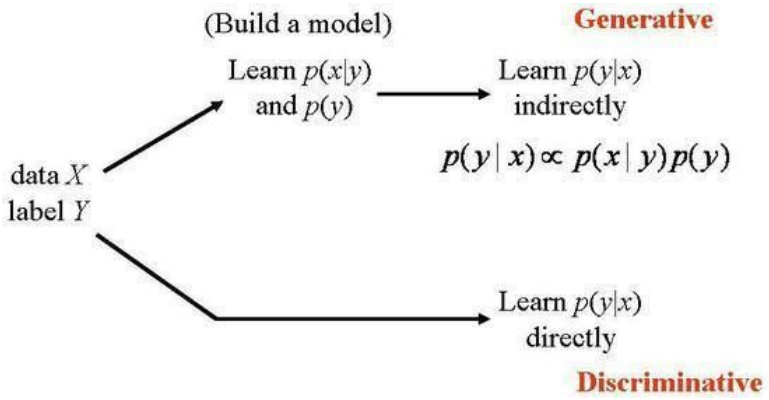
- Discriminative model이란 데이터 X와 레이블 Y가 주어졌을 때 사후확률 p(Y|X)을 **직접적**으로 도출하는 모델
  - 지도학습(supervised learning) : 레이블을 사용하여 결정경계(decision boundary)를 학습
  - 학습데이터 양이 충분하다면 좋은 성능을 냄

- Generative model이란 두 개의 확률모형 p(Y), p(X|Y)으로 정의하고, 베이즈룰을 사용사여 사후확률 p(Y|X)를 **간접적**으로 도출하는 모델
  - 비지도학습(unsupervised learning) : 레이블 정보가 필요없음
  - **토픽 모델링이 대표적 사례**
  - Discriminative model에 비교하여 가정이 많음. 가정이 잘 구축된다면 이상치에도 강건하고 학습데이터가 적어도 좋은 예측성능을 보임
  - 학습데이터가 많을수록 Discriminative model과 비슷한 성능으로 수렴

Generative / Discriminative Model

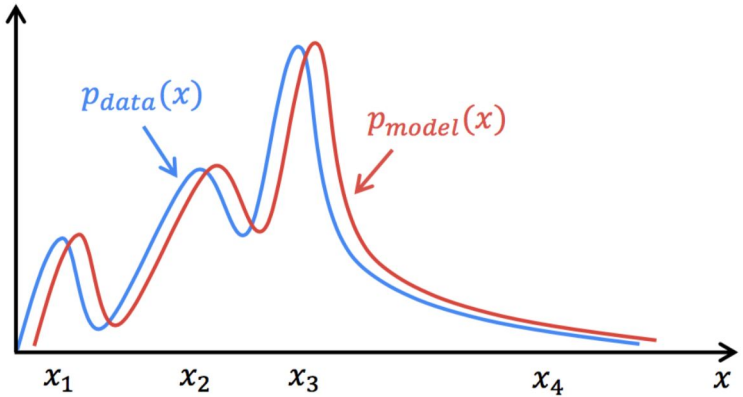
- generative model은 사후확률을 **간접적**으로, disciriminative model은 **직접적**으로 도출
- generative model은 **데이터 범주의 분포**를, disciriminative model은 **결정경계를 학습**





Generative Model

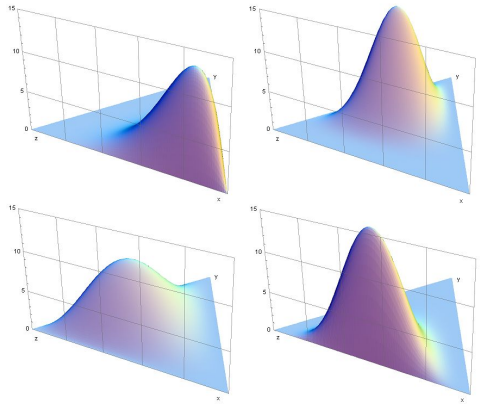
- 확률 분포와 파라미터를 안다고 할때, 랜덤 프로세스에 따라 데이터를 생성하는 모델
- 토픽 모델링에서 문서의 주제 분포와 각 주제별로 특정 단어를 생성할 확률을 알고 있으면, 특정 문서가 만들어질 확률을 계산할 수 있음
- Generative model의 목적 가운데 하나는 **데이터 분포를 학습**하는 것
- 구축한 모델에 데이터를 넣으면 실제 데이터 확률에 가깝게 값을 반환



디리클레-다항분포(Dirichlet-Multinomial Distribution)

- 디리클레 분포(Dirichlet distribution)는 연속 확률분포의 하나로, k차원의 실수 벡터 중 벡터의 요소가 양수이며 모든 요소를 더한 값이 1인 경우에 대해 확률값이 정의되는 분포이다.

- 디리클레 분포는 베이즈 통계학에서 **다항 분포에 대한 사전 켤레확률**이다. 이 성질을 이용하기 위해, 디리클레 분포는 베이즈 통계학에서의 사전 확률로 자주 사용된다.



사전 켤레확률 [편집]

디리클레 분포 $\theta \sim \mathrm{Dir}(\alpha)$와 그에 대한 다항 분포 $X|\theta \sim \mathrm{Multinomial}(\theta)$에 대하여, X가 주어졌을 때 θ의 사후 확률 $\theta|X$는 다음과 같이 나타낼 수 있다.

$$\theta|X \sim \mathrm{Dir}(\alpha + X)$$

즉, 디리클레 분포는 다항 분포에 대한 사전 켤레확률인 성질을 가지며, 사후 확률 분포는 α 벡터에 덧셈하는 것으로 계산이 가능하다.

https://ko.wikipedia.org/wiki/%EB%94%94%EB%A6%AC%ED%81%B4%EB%A0%88_%EB%B6%84%ED%8F%AC

켤레 사전 분포

- 사후확률 분포 p(θ|x)와 사전확률 분포 p(θ)와 같은 군으로 묶일 때 그 사후확률/사전확률을 모두 묶어 켤레분포(conjugate distributions)라 함 => 사후분포와 짝이되는 사전분포
- 그 사전확률 분포를 켤레사전분포(Conjugate prior distribution)라고 함
- **사전확률과 사후확률이 동일한 분포를 따른다면 계산이 매우 편해**지기 때문에 많이 사용

가능도			켤레 사전 분포
가짓수 = 2	Bernoulli A, B 둘 중 하나만 일어나는 사건	Binomial A, B가 여러 번 일어나는 사건	Beta
가짓수 > 2	Categorical A, B, … Z 중 하나만 일어나는 사건	**Multinomial** **A, B, … Z가 여러 번 일어나는 사건**	**Dirichlet**

깁스 샘플링(Gibbs sampling) (1)

- n차의 데이터에서 확률 분포 계산이 어려움. n차 데이터를 1차의 n개 데이터로 가정. 한차원에 대한 데이터를 샘플링. 각 데이터를 샘플링하여 합치면 n 차원 데이터를 샘플링 한것에 근사한다는 것이 깁스샘플링의 아이디어
- LDA 는 n차 벡터에 대한 확률 분포 계산이기 때문에 깁스샘플링을 적용하여 1차원의 데이터를 샘플링하는 것으로 전체 확률 계산 할 수 있음.
  - 1차원의 데이터 = 단어 1개
  - 단어를 1개씩 보아가면서 전체 확률을 추론.
- C_{mj}는 단어 m이 주제 j에 속하는 횟수이고, C_{dj}는 문헌 d의 단어들이 주제 j에 속하는 횟수
  - 문헌 d에 속하는 어떤 단어 m이 주제 j에 속할 확률은 주제 j에 속하는 모든 단어 중에서 단어 m이 차지하는 비중과 문헌 d에 속하는 모든 주제 중 주제 j가 차지하는 비중의 곱에 비례

$$P(Z|W) \propto \frac{C_{mj}^{WT} + \beta}{\sum_{m'} C_{m'j}^{WT} + V\beta} \cdot \frac{C_{dj}^{DT} + \alpha}{\sum_{j'} C_{dj'}^{DT} + T\alpha}$$

깁스 샘플링(Gibbs sampling) (2)

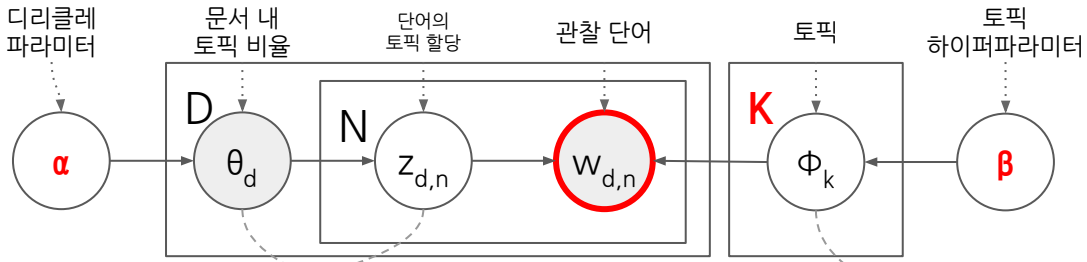
문서 d_i 에서 단어 w_i 가 토픽 j로 할당될 확률

토픽j 에 할당된 단어수

문서d_i 내 토픽j 에 할당된 단어수

$$P(z = j \mid w_i, d_i) \propto \frac{C_{w_i j}^{WT} + \beta}{\sum_{w=1}^{W} C_{wj}^{WT} + W\beta} \cdot \frac{C_{d_i j}^{DT} + \alpha}{\sum_{t=1}^{T} C_{d_i t}^{DT} + T\alpha}$$

깁스 샘플링(Gibbs sampling) (3)



문서 d_i 에서 단어 w_i 가 토픽 j로 할당될 확률

토픽j 에 할당된 단어수

문서d_i 내 토픽j 에 할당된 단어수

$$P(z = j \mid w_i, d_i) \propto \frac{C_{w_i j}^{WT} + \beta}{\sum_{w=1}^{W} C_{wj}^{WT} + W\beta} \cdot \frac{C_{d_i j}^{DT} + \alpha}{\sum_{t=1}^{T} C_{d_i t}^{DT} + T\alpha}$$

잠재 디리클레 할당 한계

- LDA 분석방법 샘플링을 이용하기 때문에 실행시마다 결과가 달라질 수 있음
  - 문서 수가 적고 단어가 희소 할 수록 결과가 달라질 수 있음
- 단어의 분포만을 가지고 주제를 그룹핑 하기 때문에 사람이 인지하는 주제와 얼마나 일치할까에 대한 문제
- 파라미터 설정의 어려움
  - 토픽의 수 K값을 얼마로 두는게 적절한지 모름
  - 적절한 K값을 설정하고 그에 따르는 α, β값을 잘 튜닝해야 좋은 결과를 얻을 수 있음

잠재 디리클레 할당 절차

토픽 개수 k 설정

모든 단어에
k개 토픽을 임의 할당

재할당 반복

- D개의 전체 문서에 k개 토픽이 분포되어있다고 가정

- 모든 단어에 k개 토픽 중 하나를 임의 할당
  - 각 문서는 토픽을 가짐
  - 토픽은 단어 분포를 가짐

- 임의 할당 했지만 올바르게 할당되었다고 가정
- 다음 과정을 반복하여 토픽을 재할당
  - p(z | d) : 문서 d의 단어들 중 토픽 z에 해당하는 단어의 비율
  - p(w | z) : 단어 w를 갖고 있는 모든 문서에서 토픽 z가 할당된 비율
  - p(w | z) * p(z | d) 에 따라 토픽 z를 할당
- 안정적인 상태(결과가 수렴)까지 반복

LDA 계산 절차 예제 (1)

A: Cute kitty
B: Eat rice or cake
C: Kitty and hamster
D: Eat bread
E: Rice, bread and cake
F: Cute hamster eats bread and cake

Insight campus SeSAC

LDA 계산 절차 예제 (2)

cute	kit	eat	rice	cake	kit	ham	eat	bre	rice	bre	cake	cute	ham	eat	bre	cake

w	cute	kit	eat	rice	cake	kit	ham	eat	bre	rice	bre	cake	cute	ham	eat	bre	cake
z	#1	#2	#1	#2	#1	#2	#2	#1	#1	#1	#2	#1	#2	#2	#2	#1	#1

θ	A	B	C	D	E	F
#1	1.1	2.1	0.1	2.1	2.1	2.1
#2	1.1	1.1	2.1	0.1	1.1	3.1

Φ	cute	kit	eat	rice	cate	ham	bre
#1	1.001	0.001	2.001	1.001	3.001	0.001	2.001
#2	1.001	2.001	1.001	1.001	0.001	2.001	1.001

LDA 계산 절차 예제 (3)

토픽할당

w	cute	kit	eat	rice	cake	kit	ham	eat	bre	rice	bre	cake	cute	ham	eat	bre	cake
z	?	#2	#1	#2	#1	#2	#2	#1	#1	#1	#2	#1	#2	#2	#2	#1	#1

문서 내 토픽 분포

θ	A	B	C	D	E	F
#1	0.1	2.1	0.1	2.1	2.1	2.1
#2	1.1	1.1	2.1	0.1	1.1	3.1
	1.2	3.2	2.2	2.2	3.2	5.2

토픽 내 단어 분포

	cute	kit	eat	rice	cake	ham	bre	합계
#1	0.001	0.001	2.001	1.001	3.001	0.001	2.001	8.007
#2	1.001	2.001	1.001	1.001	0.001	2.001	1.001	8.007

(A 내에 #1이 있을 확률) = 0.1 / (0.1+1.1) = 0.083...

(#1 내의 단어가 cute일 확률) = 0.001 / 8.007 = 0.00012...

(A 안의 cute가 #1일 확률) = 0.083 * 0.00012 ... = 0.00008...

(A 내에 #2이 있을 확률) = 1.1 / (0.1+1.1) = 0.916...

(#2 내의 단어가 cute일 확률) = 1.001 / 8.007 = 0.125...

(A 안의 cute가 #2일 확률) = 0.916 * 0.125 ... = 0.114...

LDA 계산 절차 예제 (4)

$$P(\text{토픽 1} \mid \text{문서A}) = \frac{0.1}{0.1+1.1} = 0.08333$$

$$P(cute \mid \text{토픽1}) = \frac{0.001}{8.007} = 0.00012$$

$$P(\text{토픽1} \mid cute, \text{문서A}) \propto 0.08333 \cdot 0.00012 = 0.00001$$

$$P(\text{토픽2} \mid \text{문서A}) = \frac{1.1}{0.1+1.1} = 0.91667$$

$$P(cute \mid \text{토픽2}) = \frac{1.001}{8.007} = 0.12501$$

$$P(\text{토픽2} \mid cute, \text{문서A}) \propto 0.91667 \cdot 0.12501 = 0.11459$$

LDA 계산 절차 예제 (4)

w	cute	kit	eat	rice	cake	kit	ham	eat	bre	rice	bre	cake	cute	ham	eat	bre	cake
z	#2	#2	#1	#2	#1	#2	#2	#1	#1	#1	#2	#1	#2	#2	#2	#1	#1

w	cute	kit	eat	rice	cake	kit	ham	eat	bre	rice	bre	cake	cute	ham	eat	bre	cake
z	#2	#2	#1	#2	#1	#2	#2	#1	#1	#1	#2	#1	#2	#2	#2	#1	#1

w	cute	kit	eat	rice	cake	kit	ham	eat	bre	rice	bre	cake	cute	ham	eat	bre	cake
z	#2	#2	#1	#2	#1	#2	#2	#1	#1	#1	#2	#1	#2	#2	#2	#1	#1

w	cute	kit	eat	rice	cake	kit	ham	eat	bre	rice	bre	cake	cute	ham	eat	bre	cake
z	#2	#2	#1	#2	#1	#2	#2	#1	#1	#1	#2	#1	#2	#2	#2	#1	#1

Insight campus SeSAC

LDA 계산 절차 예제 (5)

토픽할당

w	cute	kit	eat	rice	cake	kit	ham	eat	bre	rice	bre	cake	cute	ham	eat	bre	cake
z	#2	#2	#1	#1	#1	#2	#2	#1	#1	#1	#1	#1	#2	#2	#1	#1	#1

문서 내 토픽 분포

θ	A	B	C	D	E	F
#1	0.1	3.1	0.1	2.1	3.1	3.1
#2	2.1	0.1	2.1	0.1	0.1	2.1
	1.1	3.2	2.2	2.2	3.2	5.2

토픽 내 단어 분포

Φ	cute	kit	eat	rice	cake	ham	bre	합계
#1	0.001	0.001	3.001	2.001	3.001	0.001	3.001	11.007
#2	2.001	2.001	0.001	0.001	0.001	2.001	0.001	6.007

#1: eat(0.272), cake(0.272), bread(0.272), rice(0.181), cute(0.001), kitty(0.001) hamster(0.001)

#2: cute(0.333), kitty(0.333), hamster(0.333), eat(0.001), rice(0.001), cake(0.001), bread(0.001)

감사합니다.

Insight campus SeSAC

daniel@fins.ai