Perplexity

Reference

- https://shuyo.wordpress.com/2011/05/24/collapsed-gibbs-sampling-estimation-for-latent-dirichlet-allocation-1/

NLP에서 Language Modeling이나 topic modeling 등, 모델의 성능을 평가하기 위해서 쓰이는 지표 중 하나가 Perplexity이다. 한글로 직역하면 당혹이라는 뜻인데 직관적으로 무슨 뜻을 의미하는지 잘 와닿지는 않는다. perplexity는 모형이 얼마나 좋은지에 대한 지표이다. 머신러닝에서 따로 held-out 된 테스트 데이터에 대해서 정확도를 통해 모형을 비교하는 것처럼 NLP에서도 따로 held-out된 코퍼스에 대해서 perplexity을 계산한다. 따라서 앞으로 특별한 얘기가 없는한, '문서'는 모형을 훈련할 때 쓰인 문서가 아니라 따로 held-out된 문서이다.

우선 perplexity의 공식부터 살펴보자.

$$Perplexity(D_{test}) = exp\left(-\frac{log\prod_{d=1}^{M} p(\mathbf{w}_d)}{\sum_{d=1}^{M} N_d}\right)$$

테스트 문서는 M개가 있고, N_d는 d번째 문서의 단어이다. 즉, $\sum_{d=1}^{M} N_d$는 전체 문서의 단어 개수이다. 분자를 살펴보면, 이는 log likelihood임을 알 수 있다. 즉 테스트 코퍼스를 구성하는 M개의 문서에 대한 likelihood는

$$\prod_{d=1}^{M} p(\mathbf{w}_d)$$

이를 직접적으로 maximize하면 참 좋겠지만 $p(\mathbf{w})$는 intractable 하다.

$$\begin{aligned} p(\mathbf{w} \mid \alpha, \beta) &= \sum_{z} p(\mathbf{w}, z \mid \alpha, \beta) \\ &= \sum_{z} \int\int p(\mathbf{w}, \theta, \phi, z \mid \alpha, \beta) d\theta d\phi \end{aligned}$$

따라서 코퍼스의 정확한 likelihood인 $p(\mathbf{w} \mid \alpha, \beta)$을 maximize하지 않고 다른 방법의 최적화 방법으로, 크게 두 흐름이 있다. variational inference와 gibbs sampling이 그것이다.

Perplexity Via Variational Inference

blei 논문에 따르면 variational distribution q에 대해서

$$log\, p(\mathbf{w} \mid \alpha, \beta) \geq E_q\left[log\, p(\theta, z, \mathbf{w} \mid \alpha, \beta)\right] - E_q\left[log\, q(\theta, z)\right]$$

즉, $log\, p(\mathbf{w} \mid \alpha, \beta)$의 lower bound, 일명 ELBO를 maximize함으로써 코퍼스의 likelihood를 maximize 하는 것과 동일한 결과를 얻는다. 따라서 VI에 의해 쓰여진 논문은 perplexity도 다른 방식으로 구한다. 결국 코퍼스의 likelihood에 대한 lower bound을 estimates로 사용하므로, perplexity을 계산할 때에도 이를 이용한다. 즉,

$$Perplexity(D_{test}) = exp\left(-\frac{log\prod_{d=1}^{M} p(\mathbf{w}_d)}{\sum_{d=1}^{M} N_d}\right) \leq exp\left(-\frac{E_q\left[log\, p(\theta, z, \mathbf{w} \mid \alpha, \beta)\right] - E_q\left[log\, q(\theta, z)\right]}{\sum_{d=1}^{M} N_d}\right)$$

이러한 접근법은 scikit learn이나 gensim에서 구현된 perplexity 계산법과 일치한다. (github에서 source code를 확인할 수 있다.)

Perplexity Via Gibbs Sampling

VI 말고도 코퍼스의 likelihood을 간접적으로 최대로 만드는 방법이 있는데 바로 gibbs sampling 이다. VI는 likelihood의 lower bound을 최대로 만듦으로써 likelihood 자체를 최대로 만드는 것을 노리는 반면, 깁스 샘플링은 사후 분포를 근사하는데, 이를 full conditional 분포로 근사한다. LDA에서 사후 분포는 아래와 같이 적을 수 있다.

$$p(unobserved \mid observed) = p(\theta, \phi, z \mid \mathbf{w}, \alpha, \beta)$$

이를 아래와 같은 full condition 분포로 iterative하게 사후 분포를 근사한다.

$$p(\theta \mid \phi, z, \mathbf{w}, \alpha, \beta)$$

$$p(\phi \mid \theta, z, \mathbf{w}, \alpha, \beta)$$

$$p(z \mid \theta, \phi, \mathbf{w}, \alpha, \beta)$$

그런데 $p(z_i \mid \mathbf{z}_{-i}, \alpha, \beta)$을 구한다면, ϕ, θ에 대한 분포도 유도할 수 있음이 알려져 있다. 따라서 효율적인 계산을 위해 ϕ, θ를 integrate out하고 collapsed gibbs sampling을 하는 것이다. 이러한 깁스 샘플링에서는 $p(\mathbf{w} \mid \alpha, \beta)$을 어떻게 근사하는 것일까? 사실 깁스 샘플링에서는 perplexity의 정확한 형태를 유도할 수 있다.

$$\begin{aligned} logp(\mathbf{w} \mid \alpha, \beta) &= \sum_{z} \int\int p(\mathbf{w}, \theta, \phi, z \mid \alpha, \beta) d\theta d\phi \\ &= \sum_{d}\sum_{n} log\left(\sum_{z} \theta_{d,z}\phi_{n,z}\right) \end{aligned}$$

따라서

$$Perplexity(D_{test}) = exp\left(-\frac{\sum_d \sum_n log\left(\sum_z \theta_{d,z}\phi_{n,z}\right)}{\sum_{d=1}^{M} N_d}\right)$$