Supplementary Material

Variational Inference v. MCMC

전인태

1 Goal

아래와 같은 Bayes' theorem을 생각해보자:

$$p(z|x) = \frac{p(x|z)p(z)}{p(x)}.$$

$p(z|x)$는 posterior distribution (사후분포), $p(x|z)$는 likelihood, $p(z)$는 prior distribution (사전분포)이다. 분모의 $p(x)$를 구하기 위해, 분자의 $p(x|z)p(z) = p(x, z)$를 marginalize하여 z를 없애 $p(x)$를 얻을 수 있다. 예상되듯이 그러한 marginalization 과정은, 가능한 모든 z에 대해 summation을 구해줘야 하므로 prior가 어떤 형태의 분포를 가지느냐에 따라 이 과정은 매우 어려울 수 있다 (엄밀히 말하면 이산확률변수의 경우 summation, 연속확률변수의 경우 integration을 이용한다. 하지만 일반적으로 사용하는 Riemann-Stieltjes Integral은 결국 summation의 극한이므로 모두 크게 보면 summation이라 할 수 있다).
이때, 특정 prior와 likelihood 분포일 때 구한 posterior가 prior와 유사한 형태의 분포를 가지게 된다. 이러한 성질을 conjugacy라 하며, 그러한 특정 prior distribution들을 **conjugate prior**라 한다 (conjugate=켤레가 짝이라는 의미를 가짐을 생각해보자). 쉽게 말하면 prior가 A 분포를 따를 때 posterior는 A' 분포를 따르며, 이때 A' 분포의 parameter들이 얼마가 될지에 대한 공식이 이미 계산되어 있다는 것이다.
그러나 문제는 다음과 같은 상황들이다.

- 분모를 구하기 위한 marginalization이 너무 복잡한 경우
- prior나 likelihood의 형태가 너무 복잡한 경우

위와 같은 상황에서는 Bayes' theorem을 통해 posterior distribution을 구할 수 없을 것이고, 따라서 이러한 복잡한 상황을 근사를 통해 해결하고자 고안된 방법이 여럿 존재한다. 이 중 가장 대표적인 방법이 **Variational Inference**와 **MCMC**이다.

2 VI

VI의 기본적인 아이디어는 사후 분포 $p(z|x)$를 다루기 쉬운 확률분포 $q(z)$로 근사하는 것이다. 이 과정에서 Kullback-Leibler divergence를 이용한다. KLD는 두 확률분포 간 거리를 계산하는 데에 이용되는 일종의 measure이다. 예를 들어 P와 Q라는 두 이산확률분포가 있다고 하자. 이때 동일한 sample space $\mathcal{X}$에서의 두 분포 간 거리, 즉 KLD는

$$D_{KL}(P||Q) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}$$

로 정의된다. 즉, P와 Q의 로그의 차의 기댓값이다. 만약 P와 Q가 연속확률변수인 경우, measurable space $\mathcal{X}$에서 summation 대신 integration을 해주면 된다.

이제 p와 q에 적용하면, 다음을 얻을 수 있다.

$$\begin{aligned} D_{KL}(q(z)||p(z|x) &= \int q(z) \log \frac{q(z)}{p(z|x)} dz \\ &= \int q(z) \log \frac{q(z)p(x)}{p(x|z)p(z)} dz \\ &= \int q(z) \log \frac{q(z)}{p(z)} dz + \int q(z) \log p(x) dz - \int q(z) \log p(x|z) dz \\ &= D_{KL}(q(z)||p(z) + \log p(x) - E_{z \sim q(z)}[\log p(x|z)] \end{aligned}$$

우리의 목표는 KLD 값을 최대한 줄일 수 있는 q의 근사인 q^*를 찾는 것이다.
만약 우리가 prior $p(z)$와 likelihood $p(x|z)$를 알고 있다면, **Monte Carlo sampling**, **SGD** 등을 사용하여 VI를 진행할 수 있다.
이해를 위해 동전 던지기 예시를 생각해보자. 사전 분포로 베타분포를 사용하고 ($\theta \sim Beta(\alpha, \beta)$), 동전 던지기 상황이므로 동전 한 개를 던지는 사건은 앞면이 나올 확률이 θ인 베르누이 분포를 따를 것이다. 즉, $X_1, \cdots, X_n \sim Bernoulli(\theta)$이며 i.i.d.이다. 이때 사후 분포 역시 Beta 분포를 따른다는것이 알려져있다. 즉, conjugacy가 성립한다.

2.1 VI with Monte Carlo Sampling

이러한 상황에서 K개의 샘플을 통해 얻은 표본평균을 KLD의 평균에 대한 추정량으로 사용하자. 즉,

$$\int p(x)f(x)dx = E_{x \sim p(x)}[f(x)] \simeq \frac{1}{K} \sum_{i=0}^{K} [f(x_i)]_{x_i \sim p(x)}$$

이라 하자. 이러한 가정을 KLD에 적용하면 다음 식을 얻을 수 있다.

$$\begin{aligned} D_{KL}(q(z)||p(z|x) &= D_{KL}(q(z)||p(z)) + \log p(x) - E_{z\sim q(z)}[\log p(x|z)] \\ &= E_{z\sim q(z)}[\log \frac{q(z)}{p(z)}] + \log p(x) - E_{z\sim q(z)}[\log p(x|z)] \\ &\simeq \frac{1}{K}\sum_{i=0}^{K}\left[\log \frac{q(z_i)}{p(z_i)}\right]_{z_i\sim q(z)} + \log p(x) - \frac{1}{K}\sum_{i=0}^{K}[\log p(x|z_i)]_{z_i\sim q(z)} \\ &= \frac{1}{K}\sum_{i=0}^{K}[\log q(z_i) - \log p(z_i) - \log p(x|z_i)]_{z_i\sim q(z)} + \log p(x). \end{aligned}$$

여기에서 알 수 있는 것은 우리가 이 KLD 값을 축소시킬 수 있도록 하는 새로운 q^*를 찾는 것이고, 이때 q에 대한 어떠한 조건도 없으므로, 실제로는 q가 베타분포여야 하나, 어떤 분포든 근사로서 q^*가 될 수 있다. 즉, 누군가는 q^*로서 정규분포를 잡을 수 있고, 이때 정규분포의 모수인 평균과 분산을 조정해가면서 위 KLD 값을 가장 줄일 수 있을 때를 최적값 $q^* \sim N(\mu^*, \sigma^{2*})$으로 생각해주면 되는 것이다.

2.2 VI with SGD

앞선 상황과 동일한 동전 던지기 예시에서 $q^*(z)$로 정규분포를 사용했다고 가정하자. SGD를 사용하기 위해서는 미분이 가능한 함수식이 필요하므로, 미분이 불가능한 분포를 사용한다면 **SVI**(Stochastic Variational Inference)를 적용하기 어려울 것이다. $q^* \sim N(\theta^*, \sigma^2)$에서 KLD 값을 가장 작게 만들어주는 최적의 θ^*를 찾기 위해 KLD 식을 θ에 대해 미분해보자.

$$\begin{aligned} \frac{\partial}{\partial\theta}D_{KL}(q(z)||p(z|x) &= \frac{\partial}{\partial\theta}D_{KL}(q(z)||p(z) + \frac{\partial}{\partial\theta}\log p(x) - \frac{\partial}{\partial\theta}E_{z\sim q(z)}[\log p(x|z)] \\ &= \frac{\partial}{\partial\theta}E_{z\sim q(z)}[\log q(z) - \log p(z) - \log p(x|z)] \end{aligned}$$

가 되며, 위 식을 미분하려면 $\partial/\partial\theta$가 $E[]$ 안으로 들어가야 한다. 그런데 q는 θ에 의존하고 z는 q에서 뽑히기 때문에 불가능하다.
따라서 $z = \mu_q + \sigma_q\epsilon,\ \epsilon \sim N(0,1)$과 같은 노이즈 가정을하여 z가 θ에 의존하지 않게 변형하여 미분식을 $E[]$ 안으로 넣을 수 있다. 이러한 과정을 통해 미분을 완료하여 계산된 미분값(= Gradient)의 반대 방향으로 θ를 조금씩 업데이트해가며 결국 θ^*를 찾을 수 있을 것이다. 이 방법은 위의 Monte Carlo Sampling을 이용하는 방법보다 표준오차가 작아 더 유용하다고 알려져 있다 (증명 생략).

2.3 Variational Inference with EM

실전에서는 prior와 likelihood에 대한 정보가 없는 경우가 많다. 정확히는 likelihood의 파라미터와, prior를 무엇으로 둘지에 대한 확신이 없다. 그렇다면 우리는 posterior의 근사를 찾는 동시에, likelihood

$p(x|z)$의 파라미터 역시 추정해야 한다. EM algorithm은 $q(z)$의 파라미터 θ_q와 likelihood의 파라미터 θ_l을 찾는 다음의 과정을 수렴할 때까지 반복한다:

1. **Expectation**: $D_{KL}(q(z)||p(z|x)$를 줄이는 θ_q를 찾는다. (VI with MC, SVI 등을 이용하여)
2. **Maximization**: E-step에서 찾은 θ_q를 고정한 상태에서 $\log p(x)$의 하한을 최대화하는 $p(x|z)$의 파라미터 θ_l를 찾는다.

이때 prior $p(z)$의 파라미터는 임의로 설정해도 상관없다고 알려져 있다 (증명 생략). $p(z)$는 위 설명에서의 $p(x)$이다.
이때 KLD가 metric이므로 항상 0 이상이라는 점과, $D_{KL}(q(z)||p(z|x))$을 $logp(x)$에 대해 정리하면

$$\log p(x) = E_{z\sim q(z)}[\log p(x|z)] - D_{KL}(q(z)||p(z)) + D_{KL}(q(z)||p(z|x))$$

임을 고려하면,

$$\log p(x) \geq E_{z\sim q(z)}[\log p(x|z)] - D_{KL}(q(z)||p(z))$$

임을 유도할 수 있다. 이때 위 부등식의 우변을 **ELBO** (Evidence Lower Bound)라고 하며, ELBO를 최대화 시키는 방향으로 likelihood의 파라미터 θ_l을 업데이트하면 우리가 원하는 결과를 얻을 수 있다.

3 MCMC

MCMC(Markov Chain Monte Carlo)는
(1) target stationary distribution (목표 분포)가 알려져 있을 때,
(2) stationary distribution에 도달하기 위한
효율적인 transition rule (= transition probability)을 찾는 과정이다. 즉, Markov Chain이 수렴했을 때, (= stationary distribution)에 도달하면, Markov Chain에 의해 생성된 샘플은 target density의 샘플로 간주할 수 있다는 것이다. 다시말해 MCMC는 우리가 샘플을 얻고자 하는 목표 분포인 stationary distribution으로부터 랜덤 샘플을 얻는 방법이다.
이론 상 MCMC는 무한번 실행했을 때 항상 수렴한다는 장점이 있다. 그러나 Markov Chain은 독립 샘플이 아닌 종속적인 샘플을 생성한다. 이는 추정량의 분산을 증가시킨다는 단점이 있다.
MCMC 알고리즘에는 대표적으로 **Gibbs Sampler**와 **Metropolis-Hastings** 알고리즘이 존재한다.

3.1 Gibbs Sampler

Gibbs Sampler는 Metropolis-Hastings 알고리즘의 특수한 경우이다. 예를 들어 확률변수 x를 다음과 같이 d개의 요소로 분해할 수 있다고 가정하자:

$$x = \{x_1, x_2, \cdots, x_d\}.$$

이때, Gibbs Sampler에서 각 요소는 (i) randomly, or (ii) systematically 하게 선택되며, 각 샘플은 target distribution의 full conditional function에서 새로운 표본으로 업데이트된다. 예시와 함께 살펴보자.
정규분포의 모수 θ와 σ^2에 대하여, 반복적 시행에서 s번째 상태의 모수가 $\phi^{(s)} = \{\theta^{(s)}, \sigma^{2(s)}\}$로 주어져있다고 하자. 이때 Gibbs sampler는 다음과 같은 알고리즘을 통해 증명한다.

1. sample $\theta^{(s+1)} \sim p(\theta|\sigma^{2(s)}, x_1, \cdots, x_d)$
2. sample $\sigma^{2(s+1)} \sim p(\sigma^2|\theta^{(s+1)}, x_1, \cdots, x_d)$
3. $\phi^{(s+1)} := \{\theta^{(s+1)}, \sigma^{2(s+1)}\}$

결국 이 알고리즘은 S번 반복되었을 때 dependent한 파라미터 집합 수열인 $\{\phi^{(1)}, \phi^{(2)}, \cdots, \phi^{(S)}\}$를 생성할 것이다.
Gibbs Sampler는 결국 다음 상태를 뽑을 때 한꺼번에 모든 변수의 값을 정하는 것이 아니라, 다른 모든 변수는 현재값으로 고정되어 있다고 고정하고, 이 변수가 가질 수 있는 값의 조건부 분포로부터 하나를 샘플링하는 것이다. 이 과정을 모든 변수에 대해서 반복하면 하나의 새로운 상태가 만들어진다. 즉, 새로운 Markov Chain이 형성 되는 것이다. 최종적으로, 우리는 새롭게 형성된 마코프 체인을 통해서 sample을 얻게 된다.
Gibbs sampling의 장점은 우리가 찾고자 하는 target density(찾고자 하는 모수의 개수라 생각)가 가령 10차원일 때, 10차원의 분포로부터 샘플링을 하는 것은 매우 어려울 것이다. 그러나 우리는 Gibbs Sampler를 통해 10개의 1차원 문제로 바꾸어 풀 수 있어, 차원의 저주 문제를 해결할 수 있다.

3.2 Metropolis-Hastings Algorithm

Metropolis-Hastings Algorithm의 절차는 다음과 같다.
◦ $f(x)$: target density
◦ $T(x^*|x^{(s)})$: proposal density

1. $x^* \sim T(x^*|x^{(s)}$: $x^{(s)}$가 주어져 있으며, transition kernel $T(x^*|x^{(s)})$로부터 새로운 샘플 x^*을 생성.
2. MH ratio $\alpha = \frac{f(x^*)T(x^{(s)}|x^*)}{f(x^{(s)})T(x^*|x^{(s)})}$
3. $u \sim Unif(0,1)$이라 설정.
4. 만약 $u < p = \min\{\alpha, 1\}$이라면 새로운 샘플 x^*을 수용, $x^{(s+1)} := x^*$로 업데이트.
   $u \geq p$라면 x^*을 기각, $x^{(s+1)} := x^{(s)}$로 업데이트.

Metropolis-Hastings는 차원의 저주 문제를 가지고 있다. 고차원으로 갈수록 한 샘플에서의 확률값이 0으로 점점 가까워지기 때문에 Accept Ratio가 매우 낮아지므로 시간 상 비효율적이다. 즉, 고차원에서도 어떤 확률분포의 합은 결국 1이 되어야 하므로, 차원이 커질수록 이를 구성하는 분포도 차원이 커짐에 따라 넓게 퍼질 수밖에 없다.