19. 근사 추론

2021084348 인공지능학과 서하은

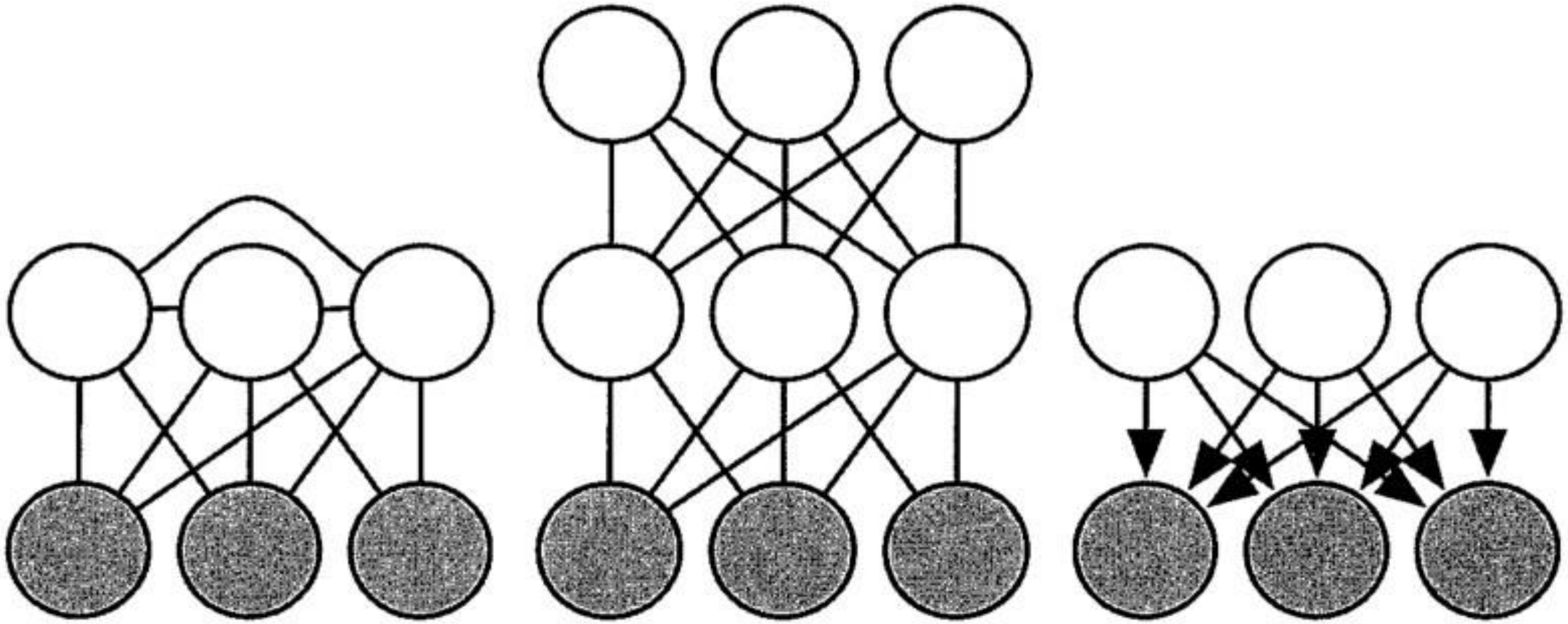
19. 근사 추론

p(h|v) 나 그 수량들에 대한 기댓값을 구하기 어려운 경우가 많음.
(v는 가시 변수들의 집합, h는 잠재변수들의 집합)

은닉층이 여러 개인 대부분의 그래프 모형에서는 계산이 처리 불가능함.

-> 근사 추론을 통해 처리 불가능한 추론 문제들을 해결함.

19. 근사 추론



처리 불가능한 추론 문제들은 주로 잠재변수들 사이의 상호작용으로 인해 발생함.

19.1 최적화로서의 추론

추론 => 최적화 문제로 변환(근사)

모형은 관측변수 v와 잠재변수 h로 구성됨.

로그가능도 $\log p(\boldsymbol{v};\boldsymbol{\theta})$ 를 계산해야 하는데 비용이 많이 듦.

로그가능도 $\log p(\boldsymbol{v};\boldsymbol{\theta})$ 대신 구하기 쉬운 하계인 $\mathcal{L}(\boldsymbol{v},\boldsymbol{\theta},q)$ 를 구함.

19.1 최적화로서의 추론

$$\begin{aligned}\mathcal{L}(\boldsymbol{v},\boldsymbol{\theta},q) &= \log p(\boldsymbol{v};\boldsymbol{\theta}) - D_{\mathrm{KL}}(q(\boldsymbol{h}|\boldsymbol{v}) \,\|\, p(\boldsymbol{h}|\boldsymbol{v};\boldsymbol{\theta}))\\
&= \log p(\boldsymbol{v};\boldsymbol{\theta}) - \mathbb{E}_{\mathbf{h}\sim q}\log\frac{q(\boldsymbol{h}|\boldsymbol{v})}{p(\boldsymbol{h}|\boldsymbol{v})}\\
&= \log p(\boldsymbol{v};\boldsymbol{\theta}) - \mathbb{E}_{\mathbf{h}\sim q}\log\frac{q(\boldsymbol{h}|\boldsymbol{v})}{\frac{p(\boldsymbol{h},\boldsymbol{v};\boldsymbol{\theta})}{p(\boldsymbol{v};\boldsymbol{\theta})}}\\
&= \log p(\boldsymbol{v};\boldsymbol{\theta}) - \mathbb{E}_{\mathbf{h}\sim q}[\log q(\boldsymbol{h}|\boldsymbol{v}) - \log p(\boldsymbol{h},\boldsymbol{v};\boldsymbol{\theta}) + \log p(\boldsymbol{v};\boldsymbol{\theta})]\\
&= -\mathbb{E}_{\mathbf{h}\sim q}[\log q(\boldsymbol{h}|\boldsymbol{v}) - \log p(\boldsymbol{h},\boldsymbol{v};\boldsymbol{\theta})].\end{aligned}$$

$\mathcal{L}$를 최대화하는 q를 찾는 최적화 문제로 변환됨.

불완전한 최적화 절차나 제한된 q분포에 적용해서 계산 비용 줄임.

19.2 기댓값 최대화

기댓값 최대화 알고리즘 (EM 알고리즘)

잠재변수가 있는 모형 훈련에 주로 쓰임.

근사 사후분포를 이용한 학습에 대한 접근 방식임.
(대부분의 알고리즘은 근사 추론을 이용함.)

E-단계와 M-단계를 수렴에 도달할 때까지 번갈아 수행함.

19.2 기댓값 최대화

기댓값 단계(E-단계):

단계 시작에서 매개변수들을 $\boldsymbol{\theta}^{(0)}$ 으로 표기함.

훈련에 사용할 견본 $\boldsymbol{v}^{(i)}$들의 모든 색인 i에 대해 $q(\boldsymbol{h}^{(i)} \mid \boldsymbol{v}) = p(\boldsymbol{h}^{(i)} \mid \boldsymbol{v}^{(i)}; \boldsymbol{\theta}^{(0)})$ 으로 설정함.

최대화 단계(M-단계):

선택한 최적화 알고리즘을 이용해서

$\sum_i \mathcal{L}(\boldsymbol{v}^{(i)}, \boldsymbol{\theta}, q)$ 를 $\boldsymbol{\theta}$ 에 대해 완전히 또는 부분적으로 최대화함.

19.2 기댓값 최대화

EM 알고리즘의 특징

1. 학습과정의 기본 구조가 존재함.
(매개변수들을 갱신->자료 가능도 개선, 결측변수를 사후분포 추정값으로 설정.)

2. $\mathcal{L}$ 를 최대화하는 좌표 상승법 알고리즘으로 볼 수 있음.
(한 단계에서는 $\mathcal{L}$ 를 q에 대해 최대화, 다른 단계에서는 $\mathcal{L}$ 를 $\boldsymbol{\theta}$ 에 대해 최대화.)

3. $\boldsymbol{\theta}$ 가 다른 값으로 이동해도 같은 q값을 계속 사용할 수 있음.

19.3 MAP 추론과 희소 부호화

일반적인 추론은 결측변수들의 모든 가능한 값에 대한 전체 분포를 추론하는 것

(=변수 집합이 주어졌을 때 다른 변수 집합의 조건부 확률분포를 계산하는 것.)

최대 사후확률 추론 (MAP 추론)

결측변수들이 가질 가능성이 가장 많은 값을 계산하는 것.

ex) 잠재변수모형의 경우 다음을 계산함.

$$\boldsymbol{h}^* = \arg\max_{\boldsymbol{h}} p(\boldsymbol{h} | \boldsymbol{v})$$

19.3 MAP 추론과 희소 부호화

MAP 추론은 기본적으로 희소 부호화 모형에 쓰인다.

희소 부호화: 희소성을 유발하는 사전분포를 은닉 단위들에 가하는 선형 인자 모형.

$$p(\boldsymbol{x} \mid \boldsymbol{h}) = \mathcal{N}(\boldsymbol{v}; \boldsymbol{W}\boldsymbol{h} + \boldsymbol{b}, \beta^{-1}\boldsymbol{I}).$$

19.3 MAP 추론과 희소 부호화

p(h|v)는 계산과 표현이 어려움 => 로그가능도와 기울기의 계산 불가능해짐.

MAP 추론으로 h를 추정하고,

추정값에 관한 분포로 정의되는 증거 하계를 최대화함으로써 매개변수를 학습함.

훈련 집합의 h벡터를 모아서 행렬 H, v벡터를 모아 행렬 V를 형성하여

다음 식을 최소화하는 것이 희소 부호화 학습과정임.

$$J(\boldsymbol{H}, \boldsymbol{W}) = \sum_{i,j} |H_{i,j}| + \sum_{i,j} \left(\boldsymbol{V} - \boldsymbol{H}\boldsymbol{W}^\top\right)^2_{i,j}.$$

19.4 변분 추론과 변분 학습

$\mathcal{L}$ 를 제한된 분포 q들의 모임에 대해 최대화 할 수 있다는 개념.

=> $\mathbb{E}_q \log p(\boldsymbol{h}, \boldsymbol{v})$ 의 계산이 쉬운 모임을 선택해야 함.

평균장 접근 방식

q가 반드시 다음과 같은 인수곱 분포여야 한다는 제약을 가하는 접근 방식

$$q(\boldsymbol{h} \mid \boldsymbol{v}) = \prod_i q(h_i \mid \boldsymbol{v})$$

19.4 변분 추론과 변분 학습

변분 접근 방식의 장점

- q의 구체적인 매개변수 형식을 지정할 필요 X.
- q의 분포가 어떤 인수들의 곱 형태인지만 지정하면 됨.
- 이산 잠재변수인지 연속 잠재변수인지에 따라 최적화 방법이 달라짐.

19.4.1 이산 잠재변수

분포 q가 개별 이진 변수 $\hat{h}_i$ 들에 관해 인수분해된다고 가정함.

=> q를 확률값으로 이루어진 벡터 $\hat{\boldsymbol{h}}$ 로 매개변수화함.

=> 매개변수화된 값들을 최적화하여 q를 선택함.

최적화 속도를 높이기 위해 고정점 방정식을 반복적으로 풂.

=> 수렴 판정기준이 충족될 때까지 아래 식을 $\hat{\boldsymbol{h}}$ 에 대해 풂.

$$\frac{\partial}{\partial \hat{h}_i} \mathcal{L} = 0$$

19.4.1 이산 잠재변수

Ex) 이진 희소 부호화 모형

최대가능도로 모형을 훈련하려면 $p(h|v)$의 기댓값을 계산해야 하는데 매우 복잡함.

=> 평균장 근사를 도입한 변분 추론과 변분 학습을 이용하여 모형을 훈련함.

19.4.2 변분법

함수 f가 입력인 함수를 범함수 $J[f]$ 라고 함.

범함수의 경우 구체적인 x값에서의 함수 f(x)의 개별 값들에 대한 편미분 가능.

미분 가능이고 도함수가 연속함수인 f(x)와 g(y,x)에 대해 다음 식이 성립함.

$$\frac{\delta}{\delta f(\boldsymbol{x})} \int g(f(\boldsymbol{x}), \boldsymbol{x}) d\boldsymbol{x} = \frac{\partial}{\partial y} g(f(\boldsymbol{x}), \boldsymbol{x})$$

19.4.2 변분법

함수를 하나의 벡터에 대해 최적화할 때

=> 벡터에 대한 기울기를 취하고 그 기울기의 모든 성분이 0인 점을 찾음.

범함수를 최적화할 때

=> 모든 점에서 범함수 미분이 0인 함수를 찾음.

19.4.2 변분법

Ex) 확률분포 함수 중 미분 엔트로피가 최대인 확률분포 함수를 구하는 문제

확률분포 p(x)의 엔트로피 정의 => $H[p] = -\mathbb{E}_x \log p(x)$

연속값들의 경우 엔트로피의 기댓값 => $H[p] = -\int p(x) \log p(x) dx$

19.4.2 변분법

P1) H[p]를 p(x)에 대해 최대화할 수 없음.

-> 라그랑주 승수들을 이용해 p(x)의 적분이 1이어야 한다는 제약을 추가함.

P2) 분산이 증가함에 따라 엔트로피가 무한히 증가하면 '최대' 엔트로피 구할 수 없음.

->고정된 분산 σ^2 에 대해 엔트로피가 최대인 분포를 찾는 것으로 최적화 문제 수정.

P3) 분산이 변해도 엔트로피가 변하지 않으면 해가 무수히 많은 과소결정 문제가 됨.

->해의 유일성을 위해 분산의 평균이 반드시 m이어야 한다는 제약을 추가함.

19.4.2 변분법

수정된 라그랑주 범함수는 다음과 같음.

$$\mathcal{L}[p] = \lambda_1\left(\int p(x)dx - 1\right) + \lambda_2(\mathbb{E}[x] - \mu) + \lambda_3(\mathbb{E}[(x-\mu)^2] - \sigma^2) + H[p] \quad (19.50)$$

$$= \int \left(\lambda_1 p(x) + \lambda_2 p(x)x + \lambda_3 p(x)(x-\mu)^2 - p(x)\log p(x)\right)dx - \lambda_1 - \mu\lambda_2 - \sigma^2\lambda_3.$$

p에 대해 최소화하기 위해, 범함수 미분들을 0으로 둠.

$$\forall x, \frac{\delta}{\delta p(x)}\mathcal{L} = \lambda_1 + \lambda_2 x + \lambda_3(x-\mu)^2 - 1 - \log p(x) = 0.$$

대수 법칙들을 이용해 정리하면

$$p(x) = \exp\left(\lambda_1 + \lambda_2 x + \lambda_3(x-\mu)^2 - 1\right)$$

19.4.2 변분법

앞선 조건을 충족하는 λ 값들을 선택하면 다음과 같은 분포가 나옴.

$$p(x) = \mathcal{N}(x; \mu, \sigma^2)$$

=> 진 분포를 모를 때는 정규분포를 사용해도 된다는 의미임.

19.4.3 연속 잠재변수

그래프 모형에 연속 잠재변수들이 있는 경우

=> 변분법을 이용해 $\mathcal{L}$ 을 $q(\boldsymbol{h}|\boldsymbol{v})$ 에 대해 최대화함.

19.4.3 연속 잠재변수

평균장 근사가 $q(\boldsymbol{h}|\boldsymbol{v}) = \prod_i q(h_i|\boldsymbol{v})$ 일 때

비정규화 분포 $\tilde{q}(h_i|\boldsymbol{v}) = \exp(\mathbb{E}_{\mathbf{h}_{-i} \sim q(\mathbf{h}_{-i}|\boldsymbol{v})} \log \tilde{p}(\boldsymbol{v}, \boldsymbol{h}))$ 를

정규화해서 최적의 $q(h_i|\boldsymbol{v})$ 값을 구함.

수렴이 일어날 때까지 i의 각 값에 대해 고정점 방정식을 반복해 최적의 값을 찾음.

19.4.4 학습과 추론의 상호작용

학습 알고리즘으로 근사 추론을 사용하면 학습 과정이 달라지고,
이는 다시 추론 알고리즘에 영향을 미침으로써 상호작용함.

=> 근사 추론을 포함한 훈련은 근사 가정이 실제로 성립할 가능성이 커지는 쪽으로 모형을 적응시킴.

19.4.4 학습과 추론의 상호작용

매개변수들을 훈련할 때 변분학습은 $\mathbb{E}_{\mathbf{h}\sim q}\log p(\boldsymbol{v},\boldsymbol{h})$ 를 증가시킴.

=> 특정한 v에 대해, h의 값 중 q(h|v) 하에서 확률이 높은 값들에 대한 p(h|v)가 증가하고, 확률이 낮은 값들에 대한 p(h|v)는 감소함.

19.4.4 학습과 추론의 상호작용

변분 근사가 끼치는 피해의 양은 $\log p(\boldsymbol{v};\boldsymbol{\theta})$ 와 $\mathcal{L}(\boldsymbol{v},\boldsymbol{\theta},q)$ 의 차이값을 통해 계산함.

=> 대부분의 경우 변분 근사가 학습과정에 큰 피해를 끼치지 않음.

더 정확한 피해 값을 구하려면 $\boldsymbol{\theta}^* = \max_{\boldsymbol{\theta}} \log p(\boldsymbol{v};\boldsymbol{\theta})$ 값을 이용함.

19.5 학습된 근사 추론

고정점 방정식이나 기울기 최적화 같은 반복적인 절차로 추론을 수행하면 시간이 오래 걸려 비용 문제가 발생함.

=> 추론 과정을 입력 v를 근사 분포 $q^* = \operatorname{argmax}_q \mathcal{L}(\boldsymbol{v}, q)$ 로 사상하는 하나의 함수로 간주하고 하나의 신경망을 이용해서 근사함.

19.5.1 각성-수면 알고리즘

v에서 h를 추론하도록 훈련할 때 사용할 지도 학습 자료 집합이 없음.

=>모형 분포에서 h와 v 모두의 표본을 추출함으로써 문제 해결함.

모형 분포하에서 확률이 높은 v의 값들에 대해서만 추론 신경망을 훈련할 수 있다는 단점이 있음.

19.5.1 각성-수면 알고리즘

인간과 동물의 꿈 수면 가설

=> 주어진 v로부터 h를 예측하는 훈련에 사용하는 p(h,v)의 표본들을 꿈이 제공한다는 가설.

=> 생물학적 꿈의 역할이 q를 예측하는 신경망을 훈련하는 것이라고 가정함.

19.5.2 그 밖의 학습된 추론 접근 방식

학습된 추론 신경망

=> 평균장 고정점 방정식을 반복하는 것보다 빠르게 추론 수행 가능함.

예측 희소 분해 모형

=> 얕은 부호기 신경망 훈련해서 입력의 희소 부호를 예측함.

변분 자동부호기

=> 추론 신경망을 위해 명시적인 목푯값을 만들 필요 없이 $\mathcal{L}$ 만 정의하면 됨.

감사합니다

Q & A