Deep Learning 輪読会 2017

第19章　近似推論

2018.01.16

東京大学理学部天文学科

4年 鹿熊亮太

構成

19.1 最適化としての推論

19.2 期待値最大化

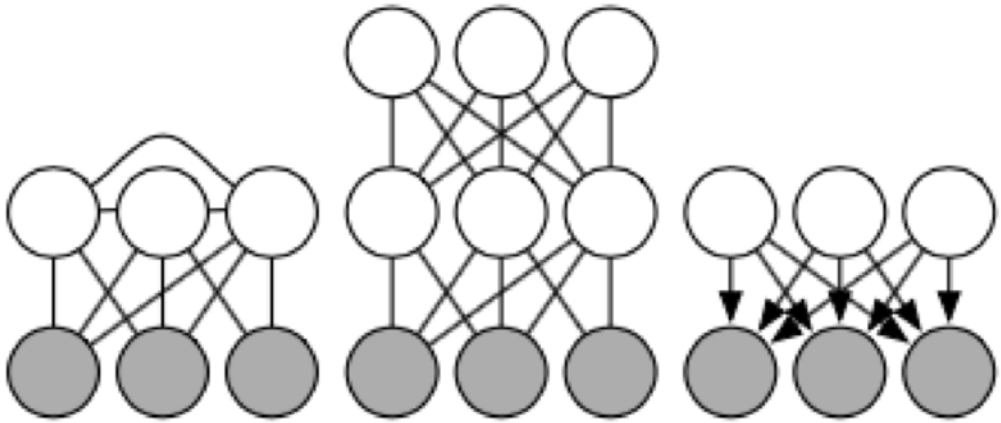
19.3 MAP推定とスパース符号化

19.4 変分推論と変分学習

19.5 学習による近似推論

はじめに

- 多くの確率モデルで学習が難しいのは、推論の実行が困難なため
  - p(h|v)を計算 or p(h|v)の期待値を取ること
- 計算困難な推論問題は、通常、構造化グラフィカルモデルにおける潜在変数間の相互作用から生じる



- 本章では、こうした問題に立ち向かう技術を紹介

19.1 最適化としての推論

- 困難な推論問題に立ち向かうためのアプローチの多くは、推論が最適化問題として記述できるという考えを利用
  - 近似推論アルゴリズムは、その根底にある最適化問題を近似することで得られる
- log p(v;θ) を計算する代わりに下界 L を計算し、これを最大にすることを考える
- エビデンス下界(evidence lower bound, ELBO)

$$\mathcal{L}(\boldsymbol{v}, \boldsymbol{\theta}, q) = \log p(\boldsymbol{v}; \boldsymbol{\theta}) - D_{\mathrm{KL}}\left(q(\boldsymbol{h} \mid \boldsymbol{v}) \| p(\boldsymbol{h} \mid \boldsymbol{v}; \boldsymbol{\theta})\right) \tag{19.1}$$

  - KLダイバージェンスは非負 ➔ Lは最大でもlog p(v;θ)
  - qがp(h|v)と等しいときに最大
- 推論は、『Lを最大にするqを見つける処理』と考えられる

19.2 期待値最大化

- 期待値最大化アルゴリズム（expectation maximization, EM）
  - 近似事後分布の学習のためのアプローチ
- 2つのステップ

① E-step

- ステップ開始時のパラメータ $\theta^{(0)}$でqを定義

$$q(\boldsymbol{h}^{(i)} \mid \boldsymbol{v}) = p(\boldsymbol{h}^{(i)} \mid \boldsymbol{v}^{(i)}; \boldsymbol{\theta}^{(0)})$$

- θ を変化させるとp(h|v,θ)も変化するが、qはp(h|v,$\theta^{(0)}$)のままとする
- q に関してLを最大化

② M-step

- 選択した最適化アルゴリズムを使って、θに関して $\sum_i \mathcal{L}(\boldsymbol{v}^{(i)}, \boldsymbol{\theta}, q)$ を完全または部分的に最大化

19.2 期待値最大化

- EMアルゴリズムに含まれる洞察
- モデルパラメータを更新するとき、すべての欠損変数（隠れ変数or潜在変数）は事後分布の推定値によって与えられる
- 異なるθの値に移動したあとでも、同じqの値を使い続けられる
  - 古典的な機械学習では、大幅なMステップ更新を導出するために用いられる
  - 深層学習では、モデルが複雑すぎるため、めったに使われない

19.3 MAP推定とスパース符号化

- 欠損変数の取りうる値における分布全体を推定するのではなく、
  欠損変数の最も可能性の高い値を計算する

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

  最大事後確率(maximum a posteriori, MAP)

- L(v,h,q)の最大化のためにMAP推定をqの値を手順として考える
  - 最適なqを提供しないため、近似推論と考えられる
- 主にスパース符号化モデルに利用されている

$$J(\boldsymbol{H}, \boldsymbol{W}) = \sum_{i,j} |H_{i,j}| + \sum_{i,j} \left(\boldsymbol{V} - \boldsymbol{H}\boldsymbol{W}^\top\right)^2_{i,j} \qquad (19.16)$$

19.4 変分推論と変分学習

- 制限された分布族qにおいてLを最大化できる
- $\mathbf{E}_q$ log p(h,v)が計算しやすいような族がいい
  - qが因子分解可能な分布とする

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

- 平均場近似法
  - 近似により、捉えたい相互作用の数を柔軟に決定
  - 統計力学の文脈で登場
- qの因子分解の方法を指定するだけでよく、事後分布を正確に近似できる特定のqの設計方法を推測する必要はない

19.4.1 離散潜在変数

- 離散の場合の変分法は、

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

  を収束基準を満たすまで hの異なる要素を繰り返し更新すること

- 二値スパース符号化モデルに適応する具体例が載っている

  …が、数学的にかなり細かいので割愛(ごめんなさい)

19.4.2 変分法

- そもそも変分法とはなんたるか
- 関数 f の関数は汎関数 J[f]
  - いうなれば、要素が無限個のベクトルの関数である
- 汎関数微分は、以下のようになる

  （gがfのみを引数とする場合 i.e. 導関数を引数としない）

$$\frac{\delta}{\delta f(\boldsymbol{x})} \int g\left(f(\boldsymbol{x}), \boldsymbol{x}\right) d\boldsymbol{x} = \frac{\partial}{\partial y} g(f(\boldsymbol{x}), \boldsymbol{x}). \tag{19.46}$$

  - 要素が無限個のベクトル微分のchain ruleみたいな

19.4.2. 変分法

- エントロピーを最大化する確率分布を変分法から求める
- 以下の条件を課す
  - 確率密度の積分が1　分散、平均値を固定

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

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

- 汎関数微分は、

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

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

- これを元の条件式に戻してみると、

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

  - 真の分布がわからない場合に正規分布を使う理由の１つ

19.4.2 変分法

- エントロピーの最小値に対応する臨界点はない
- 最小エントロピーを達成する特定の関数がないため
  - 最小の正の実数が存在しないのと同様
- ２つの点以外に正確に質量をゼロとする関数は積分が1とならないので、確率分布とならない
- ２つの点にのみ質量を置くように収束する確率分布は存在
  - これは、混合ディラック分布として記述できる
- こうした分布は、汎関数微分がゼロとなる特定の点について解く手法は存在せず、変分法による限界である。

19.4.3. 連続潜在変数

- 多くの場合、変分法自体の問題を解く必要はない
- 平均場近似を用いて、すべての j ≠ iについてq(h_j|v)を固定するなら、

$$\tilde{q}(h_i \mid \boldsymbol{v}) = \exp\left(\mathbb{E}_{\mathbf{h}_{-i} \sim q(\mathbf{h}_{-i} \mid \boldsymbol{v})} \log \tilde{p}(\boldsymbol{v}, \boldsymbol{h})\right)$$

  を正規化することで、q(h_i|v) の正しい関数形が得られる
- この式は、最適解が取る関数系も示す
- 深層学習における連続変数を用いた変分学習の実際の応用例については Goodfellow et al.+13d参照

19.4.4 学習と推論の相互作用

- 近似推論を学習アルゴリズムの一部として使うことは学習仮定に影響し、これが今度は推論アルゴリズムの精度に影響する
- 訓練アルゴリズムは、近似推論アルゴリズムの根底にある近似仮定がより新になるようにモデルを適応する傾向がある

19.5 学習による近似推論

- 不動点方程式や勾配に基づく最適化などの反復処理によって明示的に実行する最適化は、非常に高価で時間がかかることが多い
- 多くの推論のためのアプローチでは、近似推論の実行を学習することによって、このコストを回避する
  - 最適化処理が、入力 v を近似分布 $q^* = \arg\max_q \mathcal{L}(\boldsymbol{v}, q)$ に写像する関数fであると考えられる
- 近似推論を使うことによって、幅広い種類のモデルを訓練し、利用できるようになる。こうしたモデルの多くは次の章にて説明される

参考文献

- Deep Learning
  - Ian Goodfellow, Yoshua Bengio, Aaron Courville
  - 日本語版
    https://www.amazon.co.jp/%E6%B7%B1%E5%B1%A4%E5%AD%A6%E7%BF%92-Ian-Goodfellow/dp/4048930621