Advanced Deep Learning

Approximate Inference-2

U Kang
Seoul National University

Outline

- ☑ Inference as Optimization
- ☑ Expectation Maximization
- ☐ **MAP Inference**
- ☐ Variational Inference and Learning

MAP Inference

- If we wish to develop a learning process based on maximizing $L(v, h, q)$, then it is helpful to think of MAP inference as a procedure that provides a value of q.

- What is MAP inference?
  - Finds the most likely value of a variable

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

MAP Inference

- Exact inference consists of maximizing

$$\mathcal{L}(\boldsymbol{v}, \boldsymbol{\theta}, q) = \mathbb{E}_{\mathbf{h} \sim q}\left[\log p(\boldsymbol{h}, \boldsymbol{v})\right] + H(q)$$

  with respect to q over an unrestricted family of probability distributions, using an exact optimization algorithm.
- We restrict the family of distributions of q to take on a Dirac distribution $q(\boldsymbol{h} \mid \boldsymbol{v}) = \delta(\boldsymbol{h} - \boldsymbol{\mu})$

$$\delta(x) = \begin{cases} +\infty, & x = 0 \\ 0, & x \neq 0 \end{cases} \qquad \int_{-\infty}^{\infty} \delta(x)\, dx = 1$$

MAP Inference

- We can now **control** q entirely via μ. Dropping terms of L that do not vary with μ, we are left with the optimization problem

$$\boldsymbol{\mu}^* = \arg\max_{\boldsymbol{\mu}} \log p(\boldsymbol{h} = \boldsymbol{\mu}, \boldsymbol{v})$$

which is equivalent to the MAP inference problem

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

ELBO:

$$\mathcal{L}(\boldsymbol{v}, \boldsymbol{\theta}, q) = \mathbb{E}_{\mathbf{h} \sim q}\left[\log p(\boldsymbol{h}, \boldsymbol{v})\right] + H(q)$$

MAP Inference

- Thus, we can think of the following algorithm for maximizing ELBO, which is similar to EM
  - Alternate the following two steps
    - Perform MAP inference to infer h* while fixing θ
    - Update θ to increase log p(h*, v)

ELBO:

$$\mathcal{L}(\boldsymbol{v}, \boldsymbol{\theta}, q) = \mathbb{E}_{\mathbf{h}\sim q}\left[\log p(\boldsymbol{h}, \boldsymbol{v})\right] + H(q)$$

Outline

- ☑ Inference as Optimization
- ☑ Expectation Maximization
- ☑ MAP Inference
- ☐ **Variational Inference and Learning**
  - **Discrete Latent Variables**
  - Calculus of Variations
  - Continuous Latent Variables

Previously

- Lower bound L
  - Inference = max L   w.r.t. q
  - Learning = max L w.r.t. θ
  - EM algorithm -> allows us to make large learning steps with fixed q
  - MAP inference enable us to learn using a point of estimate rather than inferring the entire distribution

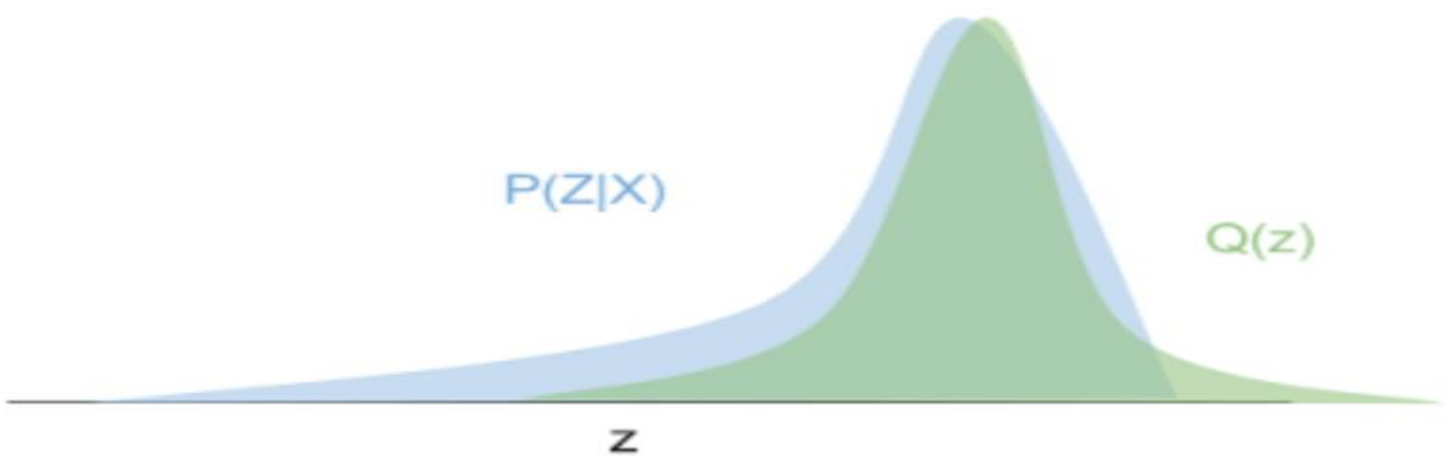
Variational Methods

- We want to do the following:
  - *Given this surveillance footage* X*, did the suspect show up in it?*
  - *Given this twitter feed* X*, is the author depressed?*

- Problem: cannot compute $P(z|x)$

Variational Methods

- Idea:
  - Allow us to re-write *statistical inference* problems as *optimization problems*
    - *Statistical inference = infer value of RV given another*
    - *Optimization problem = find the parameter values that minimize cost function*



Variational Learning

- Key idea: We can maximize L over a restricted family of distributions q
  - L is the lower bound of $\log p(v;\theta)$
  - Chose family such that $\mathbb{E}_{\mathbf{h}\sim q}[\log p(\boldsymbol{h}, \boldsymbol{v})]$ is easy to compute.
  - Typically: impose that q is a factorial distribution

The Mean field approach

- q is a factorial distribution $q(\boldsymbol{h} \mid \boldsymbol{v}) = \prod_i q(h_i \mid \boldsymbol{v})$
  - where h_i are independent (and thus q(h|v) cannot match the true distribution p(h|v))
- Advantage: no need to specify parametric form for q.
  - The optimization problem determines the optimal probability under the constraints

Discrete Latent Variable

- The goal is to maximize ELBO:

$$\mathcal{L}(\boldsymbol{v}, \boldsymbol{\theta}, q) = \mathbb{E}_{\mathbf{h}\sim q}\left[\log p(\boldsymbol{h}, \boldsymbol{v})\right] + H(q)$$

- In the mean field approach, we assume q is a factorial distribution

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

- We can parameterize q with a vector $\hat{h}$ whose entries are probabilities; then $q(h_i = 1|v) = \hat{h_i}$
- Then, we simply optimize the parameters $\hat{h}$ by any standard optimization technique (e.g., gradient descent)

Outline

- ☑ Inference as Optimization
- ☑ Expectation Maximization
- ☑ MAP Inference and Sparse Coding
- ☐ **Variational Inference and Learning**
  - Discrete Latent Variables
  - **Calculus of Variations**
  - Continuous Latent Variables

Recap

- Inference is to compute $p(h|v) = \frac{p(v|h)p(h)}{p(v)}$
- However, exact inference requires an exponential amount of time in these models
  - Computing $p(v)$ is intractable
- Approximate inference is needed

Recap

- We compute Evidence Lower Bound (ELBO) instead of $p(v)$.
- ELBO: $\mathcal{L}(\boldsymbol{v}, \boldsymbol{\theta}, q) = \log p(\boldsymbol{v}; \boldsymbol{\theta}) - D_{\mathrm{KL}}\left(q(\boldsymbol{h} \mid \boldsymbol{v}) \| p(\boldsymbol{h} \mid \boldsymbol{v}; \boldsymbol{\theta})\right)$
  - After rearranging the equation,

$$\mathcal{L}(\boldsymbol{v}, \boldsymbol{\theta}, q) = \mathbb{E}_{\mathbf{h} \sim q}\left[\log p(\boldsymbol{h}, \boldsymbol{v})\right] + H(q)$$

  - For any choice of q, L provides a lower bound.
  - If we take q equals to p, we can get $p(v)$ exactly.

Calculus of Variations

- In machine learning, minimizing a function $J(\theta)$ by finding the input vector $\theta \in R^n$ is the purpose.
- This can be accomplished by solving for the critical points where $\nabla_\theta J(\theta) = 0$.

Calculus of Variations

- But in some cases, we actually want to solve for a function $f(x)$.
- Calculus of variations is a method of finding the critical points w.r.t $f(x)$.
- A function of a function f is **functional** $J[f]$.
- We can take **functional derivatives** (a.k.a. **variational derivatives**), of $J[f]$ with respect to individual values of the function $f(x)$ at any specific value of x.
- **Functional derivatives** is denoted $\frac{\delta}{\delta f(x)} J$.

Calculus of Variations

- Euler-Lagrange equation (simplified form)
  - Consider a functional $J(f) = \int g(f(x), x)\, dx$. Extreme point of J is given by the condition $\frac{\partial}{\partial f(x)} g(f(x), x) = 0$
  - (Proof) Assume we change f by the amount of $\epsilon \cdot \eta(x)$ by an arbitrary function $\eta(x)$. Then, $J\big(f(x) + \epsilon\eta(x)\big) = \int g(f(x) + \epsilon\eta(x), x)\, dx = \int [g(f(x), x) + \frac{\partial g}{\partial f}\epsilon\eta(x)]\, dx = J + \epsilon \int \frac{\partial g}{\partial f}\eta(x)\, dx$
    - =: Use $y(x_1 + \epsilon_1, \ldots, x_D + \epsilon_D) = y(x_1, \ldots, x_D) + \sum_{i=1}^{D} \frac{\partial y}{\partial x_i}\epsilon_i + O(\epsilon^2)$
  - Note that at the extreme point, $J\big(f + \epsilon\eta(x)\big) = J$, which implies $\int \frac{\partial g}{\partial f}\eta(x) = 0$. Since this is true for any $\eta(x)$, $\frac{\partial g}{\partial f} = 0$.

Calculus of Variations

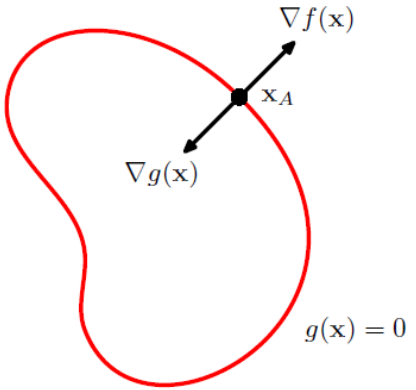
- $\mathcal{L}(\boldsymbol{v}, \boldsymbol{\theta}, q) = \mathbb{E}_{\mathbf{h}\sim q}\left[\log p(\boldsymbol{h}, \boldsymbol{v})\right] + H(q)$
  - $H(q) = -\int q(x) \log q(x)\, dx$
- We want to maximize L.
- So we have to find the q which becomes a critical point of L
- Find $\frac{\delta}{\delta q(x)} L$=0

Example

- Consider the problem of finding the probability distribution function over $x \in R$ that has the maximal differential entropy $H(p) = -\int p(x) \log p(x)\, dx$ among the distribution with $E(x) = \mu$ and $Var(x) = \sigma^2$
  - I.e., the problem is to find $p(x)$ to maximize $H(p) = -\int p(x) \log p(x)\, dx$, such that
    - p(x) integrates to 1
    - $E(x) = \mu$
    - $Var(x) = \sigma^2$

Lagrange Multipliers

- How to maximize (or minimize) a function with equality constraint?
- Lagrange multipliers
  - Problem: maximize f(x) when g(x)=0
  - Solution
    - Maximize $L(x, \lambda) = f(x) + \lambda \cdot g(x)$
    - λ is called Lagrange multiplier
    - We find x and λ s.t. $\nabla_x L = 0$ and $\frac{\partial L}{\partial \lambda} = 0$
      - $\nabla_x f(x)$ and $\nabla_x g(x)$ are orthogonal to the surface g(x)=0; thus $\nabla_x f(x) = -\lambda \nabla_x g(x)$ for some λ
      - $\frac{\partial L}{\partial \lambda} = 0$ leads to g(x)=0



Calculus of Variations

- Goal: find $p(x)$ which maximizes $H(p) = -\int p(x) \log p(x)\, dx$, s.t. $\int p(x)\, dx = 1$, $E(x) = \mu$, and $Var(x) = \sigma^2$
- Using Lagrange multiplier, we maximize

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

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

Calculus of Variations

- We set the functional derivatives equal to 0:

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

- We obtain $p(x) = \exp\left(\lambda_1 + \lambda_2 x + \lambda_3 (x-\mu)^2 - 1\right).$
- We are free to choose Lagrange multipliers as long as $\int p(x)\,dx = 1$, $E(x) = \mu$, and $Var(x) = \sigma^2$
- We may set the followings.

$$\lambda_1 = 1 - \log \sigma\sqrt{2\pi} \qquad \lambda_2 = 0 \qquad \lambda_3 = -\frac{1}{2\sigma^2}$$

- *Then, we obtain* $p(x) = \mathcal{N}(x; \mu, \sigma^2).$
- This is one reason for using the normal distribution when we do not know the true distribution.

Outline

- ☑ Inference as Optimization
- ☑ Expectation Maximization
- ☑ MAP Inference and Sparse Coding
- ☐ **Variational Inference and Learning**
  - Discrete Latent Variables
  - Calculus of Variations
  - **Continuous Latent Variables**

Continuous Latent Variables

- When our model contains continuous latent variables, we can perform variational inference by maximizing L using calculus of variations.
- If we make the mean field approximation, $q(h|v) = \prod_i q(h_i|v)$, and fix $q(h_i|v)$ for all $i \neq j$, then the optimal $q(h_j|v)$ can be obtained by normalizing the unnormalized distribution

$$\tilde{q}(h_j|v) = \exp(E_{h_{-j} \sim q(h_{-j}|v)} \log p\,(h, v))$$

- Thus, we apply the above equation iteratively for each value of j until convergence

Proof of Mean Field Approximation

- (claim) Assuming $q(h|v) = \prod_i q(h_i|v)$, the optimal $q\left(h_j|v\right)$ is given by normalizing the unnormalized distribution $\tilde{q}\left(h_j|v\right) = \exp(E_{h_{-j}\sim q\left(h_{-j}|v\right)} \log p\,(h,v))$
- (proof) Note that ELBO = $E_{h\sim q}[\log p(h,v)] + H(q)$, and $q(h|v) = \prod_i q_i(h_i|v)$.
  - Thus, ELBO = $\int \prod_i q_i\,(\log p(h,v))\,dh - \int (\prod_i q_i)(\log \prod_i q_i)\,dh = \int q_j\,\{\int \log p(h,v)\,(\prod_{i\neq j} q_i\,dh_i)\}\,dh_j - \int (\prod_i q_i)(\sum_i \log q_i)dh$
  - If we take out terms related to q_j, then ELBO becomes

    $\int q_j\,E_{h_{-j}}(\log p(h,v))dh_j - \int q_j \log q_j\,dh_j + const$

    $= \int q_j\ \log p^*(h_j,v)\,dh_j - \int q_j \log q_j\,dh_j + const \qquad (1)$
    - where $p^*(h_j,v)$ is a prob. distribution and $\log p^*(h_j,v) = E_{h_{-j}}(\log p(h,v)) + const$
  - Note that (1) is negative KL divergence -$D_{KL}(q_j||p^*(h_j,v))$; thus, the best q_j maximizing ELBO is given by $q_j = p^*(h_j,v)$.
  - In that case, $\log q_j = \log p^*(h_j,v) = E_{h_{-j}}(\log p(h,v)) + const$. Thus, $q_j \propto \exp\left(E_{h_{-j}}(\log p(h,v))\right) = \exp(E_{h_{-j}\sim q\left(h_{-j}|v\right)} \log p\,(h,v))$

What you need to know

- Inference as Optimization
- Expectation Maximization
- MAP Inference and Sparse Coding
- Variational Inference and Learning

Questions?