10-708: Probabilistic Graphical Models 10-708, Spring 2018

14 : Mean Field Assumption

Lecturer: Kayhan Batmanghelich *Scribes: Yao-Hung Hubert Tsai*

1 Inferential Problems

Can be categorized into three aspects:

- Marginalisation : $p(y) = \int p(y, \theta)d\theta$
- Expectation : $\mathbb{E}[f(y)|x] = \int f(y)p(y|x)dy$
- Prediction : $p(y_{t+1}) = \int p(y_{t+1}|y_t)p(y_t)dy_t$

We can use **Variational Methods** to approximate a complicated of-interest-density. In other words, the **Variational Principle** is to use a general family of methods to approximate complicated densities by a simpler class of densities.

2 Variational Calculus

Two types of derivations:

- *Variables* as input, output is a value. E.g., $\frac{df}{dx}$
- *Functions* as input, output is a value. E.g., $\frac{\delta F}{\delta f} \rightarrow maxH[p(x)]$ w.r.t. $p(x)$

Two basics on functional calculus:

- Functional derivative: $\frac{\delta f(x)}{\delta f(x')} = \delta(x - x')$
- Commutative rule: $\frac{\delta}{\delta f(x')} \frac{\partial f(x)}{\partial x} = \frac{\partial}{\partial x} \frac{\delta f(x)}{\delta f(x')}$

E.g., $\frac{\delta H[p(x)]}{\delta p(x)} = -1 - \log p(x)$. (See lecture slides for details or do it your self.)

3 Variational Methods

Note that, the notations used in the slides are inconsistent from slides to slides, and it results in lots of confusion, so I try to unify them.

Goal: Approximate a difficult distribution $p(z|D)$ with a new distribution $q(z)$

- $p(z|D)$ and $q(z)$ should be close
- Computation on $q(z)$ should be easy

Use Kullback-Leibler divergence (KL-divergence) to measure probability discrepancy between p and q, the loss function $J(q)$ (for unnormalized distribution) becomes

$$\begin{aligned}
&\sum_z q(z)\log\frac{q(z)}{\tilde{p}(z)}\\
=&\sum_z q(z)\log\frac{q(z)}{Z\cdot p(z)}(Z \text{ is the normalizer})\\
=&\sum_z q(z)\log\frac{q(z)}{p(z)}-\log Z\\
=&\mathcal{KL}(q||p)-\log Z
\end{aligned}$$

Since Z is constant, by minimizing $J(q)$, we force q to become close to p.

Therefore, if we want to minimize $\mathcal{KL}(q||p)$, we can actually maximize $-\sum_z q(z)\log\frac{q(z)}{\tilde{p}(z)}\leq\log Z$, which is also called *evidential lower bound (ELBO)*. In other words, $\log Z-\mathcal{KL}(q||p)=$ELBO$(q)$.

Alternative Interpretations:

- View 1: Minimize expected energy while maximizing entropy

$$J(p)=\mathbb{E}_q[\log q(z)]+\mathbb{E}_q[-\log\tilde{p}(z)]=-\mathbb{H}(z)+\mathbb{E}_q[E(z)]$$

  It is also called variational free energy or Helmholtz free energy

- View 2: Expected Evidence plus a penalty term that measures how far apart the two distributions are

$$J(p)=\mathcal{KL}(q||p)-\log Z=\mathcal{KL}(q||p)-\log p(D)$$

4 Forward or Reverse KL

- Information Projection:

$$\mathsf{KL}(q||p)=\sum_z q(z)\log\frac{q(z)}{p(z)}$$

  - We must ensure that if $p(x)=0$ then $q(x)=0$. (Infinite if $p(x)=0$ and $q(x)>0$)
  - *Zero Forcing:* q will *under-estimate* the support of p.

- Moment Projection:

$$\mathsf{KL}(p||q)=\sum_z p(z)\log\frac{p(z)}{q(z)}$$

  - Infinite if $q(x)=0$ and $p(x)>0$.
  - *Zero Avoiding:* q will *over-estimate* the support of p.

5 Interpreting Variational Lower Bound using Jensen Inequality

$$
\begin{aligned}
\log p(x) &= \log \int p(x|z)p(z)\frac{q(z|x)}{q(z|x)}dz \\
&\geq \int q(z|x)\log p(x|z)\frac{p(z)}{q(z|x)}dz \\
&= \mathbb{E}_{q(z|x)}[\log p(x|z)] - \mathcal{KL}(q(z|x)||p(z))\big(\text{Variational Lower Bound}\big) \\
&= \mathcal{F}(x, q)
\end{aligned}
$$

- x: data
- $q(z|x)$ is the approximate posterior for matching the true posterior $p(z|x)$
- $\mathbb{E}_{q(z|x)}[\log p(x|z)]$: Reconstruction
- $\mathcal{KL}(q(z|x)||p(z))$: Penalty Term
- Parameters for $q(z|x)$ is called variational parameters

Integration is now optimisation:

- Free form:

$$
\begin{gathered}
\frac{\delta \mathcal{F}(x, q)}{\delta q(z|x)} = 0 \text{ s.t. } \int q(z|x)dz = 1 \\
\rightarrow q(z|x) \propto p(z)\exp\left(\log p(x|z, \theta)\right)
\end{gathered}
$$

  - The optimal solution is the true posterior distribution.
  - **But** solving for the normalization is our original problem.

- Fixed form:

$$q_\phi(z|x) = f(z, x; \phi)$$

  - This is ideally a rich class of distributions.
  - ϕ: variational parameters

6 Naive Mean Field Approach

Assume the posterior is fully factorizable

$$q(z|x;\phi) = \prod_i q_i(z_i|x;\phi_i)$$

Goal:

$$\min_{q_1 \cdots q_D} \mathcal{KL}(q||p)$$

Instead, maximizing its variational lower bound:

$$L(q) = -J(q) = \sum_z q(z|x)\log \frac{\tilde{p}(z)}{q(z|x)}$$

7 Mean Field Updates

Let's focus on q_j (holding all the others constant)

$$\begin{aligned}
L(q_j) &= \sum_z \prod_i q_i(z|x)\Big[\log \tilde{p}(z) - \sum_k \log q_k(z_k|x)\Big] \\
&= \sum_{z_j} \sum_{z_{-j}} q_j(z_j|x) \prod_{i\neq j} q_i(z_i|x)\Big[\log \tilde{p}(z) - \sum_k \log q_k(z_k|x)\Big] \\
&= \sum_{z_j} q_j(z_j|x) \sum_{z_{-j}} \prod_{i\neq j} q_i(z_i|x) \log \tilde{p}(z) - \sum_{z_j} q_j(z_j|x) \sum_{z_{-j}} \prod_{i\neq j} q_i(z_i|x)\Big[\sum_{k\neq j} \log q_k(z_k|x) + \log q_j(z_j|x)\Big] \\
&= \sum_{z_j} q_j(z_j|x) \log f_j(z_j, x) - \sum_{z_j} q_j(z_j|x) \log q_j(z_j|x) + \text{const.} \\
&\text{with } \log f_j(z_j, x) = \sum_{z_{-j}} \prod_{i\neq j} q_i(z_i|x) \log \tilde{p}(z)
\end{aligned}$$

To sum up,

$$L(q_j) = \mathbb{E}_{q_j}\Big[\mathbb{E}_{q_{-j}}\big[\log \tilde{p}(z)\big]\Big] + \mathcal{H}(q_j)$$

Solving $\frac{\delta L(q_j)}{\delta q_j} = 0$, we get

$$\frac{\delta L(q_j)}{\delta q_j} = \mathbb{E}_{q_{-j}}\big[\log \tilde{p}(z)\big] - \log q_j - 1 = 0$$

In short,

$$q_j^* \propto \exp\Big(\mathbb{E}_{q_{-j}}\big[\log \tilde{p}(z)\big]\Big)$$

8 Wrapping Up

Advantages:

- Applicable to almost *all probabilistic models*: non-linear, non-conjugate, high-dimensional, directed and undirected.
- Can be *faster to converge* than competing methods.
- Easy *convergence assessment.*
- *Numerically stable.*
- Can be used on *modern computing architectures* (CPUs and GPUs).
- Principled and scalable approach for *model selection.*

Disadvantages:

- An *approximate posterior* only - not always guaranteed to find exact posterior.
- *Difficulty in optimisation* can get stuck in local minima.
- Typically *under-estimates the variance* of the posterior and can bias maximum likelihood parameter estimates.

- *Limited theory* and guarantees for variational methods.

Mean filed v.s. LBP

- LBP minimizes the *Bethe* energy while MF maximizes the *ELBO*.
- LBP is *exact* for trees whereas MF is not, suggesting LBP will in general.
- LBP optimizes over *node and edge marginals*, whereas naive MF only optimizes over *node marginals*, again suggesting LBP will be more accurate.
- MF objective has many more local optima than the LBP objective, so optimizing the MF objective seems to be harder.
- MF tends to be more *overconfident* than BP.
- The advantage of MF is that it gives a lower bound on the partition function while for LBP we don't know the relationship.
- MF is *easier* to extend to other distributions besides discrete and Gaussian.