变分推断 Variational Inference

一、背景

频率角度 → 优化问题

- 回归：$f(w)=w^Tx+b$，loss function：$L(w)=\sum_{i=1}^{N}\|w^Tx_i-y_i\|^2$ 无约束
  - $\hat{w}=\arg\max L(w)$
  - ① 解析解：$\hat{w}=(X^TX)^{-1}X^TY$
  - ② 数值解：可用 GD，SGD 求解
- SVM（分类）：$f(w)=\text{sign}(w^Tx+b)$
  - loss function：$\min \frac{1}{2}w^Tw$，s.t. $y_i(w^Tx_i+b)\geq 1$，$i=1,\dots,N$ 有约束
  - 可用 Lagrange 乘子法或 QP 求解
- EM：$\hat{\theta}=\arg\max \log P(X|\theta)$
  - $\theta^{(t+1)}=\arg\max \int_z \log P(X,Z|\theta)\,P(Z|X,\theta^{(t)})\,dz$

贝叶斯角度 → 积分问题

$$P(\theta|X)=\frac{P(X|\theta)\,P(\theta)}{P(X)}=\frac{P(X|\theta)\,P(\theta)}{\int_\theta P(X|\theta)P(\theta)\,d\theta}$$

（似然；$P(\theta)$ → 先验）

贝叶斯推断：指在贝叶斯框架中求出后验概率。

贝叶斯决策：可理解为预测

$X\to N$ 个样本，

对于新样本 $\tilde{x}$，有 $P(\tilde{x}|X)=\int_\theta P(\tilde{x},\theta|X)\,d\theta$

$=\int_\theta P(\tilde{x}|\theta)P(\theta|X)\,d\theta$（$P(\theta|X)$：后验）

$=E_{\theta|X}[P(\tilde{x}|\theta)]$

Inference 求出后验，就可以做决策了。

二、变分推断一公式推导.

X：观测变量

Z：隐变量+参数.

(X,Z)：完整数据.

（将参数看作隐变量，somehow不写出了.）

$$\log P(X) = \log P(X,Z) - \log P(Z|X)$$
$$= \log \frac{P(X,Z)}{q(Z)} - \log \frac{P(Z|X)}{q(Z)}$$

等式两边对Z求期望：

$$\text{左边} = \int_Z \log P(X)\, q(Z)\, dZ = \log P(X) \int_Z q(Z)\, dZ = \log P(X)$$

$$\text{右边} = \underbrace{\int_Z q(Z) \cdot \log \frac{P(X,Z)}{q(Z)}}_{ELBO} \underbrace{- \int_Z q(Z) \cdot \log \frac{P(Z|X)}{q(Z)}\, dZ}_{KL(q\|p)}$$
$$= \underbrace{\mathcal{L}(q)}_{\text{变分}} + \underbrace{KL(q\|p)}_{\geq 0}$$

由于$P(X)$与q无关，因此等式左边$=\log P(X)$一定时，$\mathcal{L}(q)+KL(q\|p)$为固定值，q无论如何取值，最大又能达到与$\log P(X)$一样大。

我们想找到一个$q(Z)$，使$q(Z)$近似等于$p(Z|X)$，因此，我们希望找到一个$q(Z)$使$\mathcal{L}(q)$能够达到最大.

则 $\tilde{q}(Z) = \arg\max_{q(Z)} \mathcal{L}(q) \Longrightarrow \tilde{q}(Z) = p(Z|X)$

假设$q(Z)$可以划分为M个划分，这M个划分相互独立。 由平均场理论得来.

$$q(Z) = \prod_{i=1}^{M} q_i(Z_i)$$

求解q_j，固定$q_1, q_2, \cdots q_{j-1}, q_{j+1}, \cdots q_M$.

$$\mathcal{L}(q) = \underbrace{\int_Z q(Z) \log P(X,Z)\, dZ}_{①} - \underbrace{\int_Z q(Z) \log q(Z)\, dZ}_{②}$$

$$① = \int_Z \prod_{i=1}^{M} q_i(Z_i) \log P(X,Z)\, dZ_1 dZ_2 \cdots dZ_M$$
$$= \int_{Z_j} q_j(Z_j) \left(\int_{Z_1 \cdots Z_M (i \neq j)} \prod_{i \neq j}^{M} q_i(Z_i) \log P(X,Z)\, dZ_{1,\cdots (i\neq j)} \cdots dZ_M \right) dZ_j$$
$$= \int_{Z_j} q_j(Z_j) \underbrace{\left(\int_{Z_1 \cdots Z_M (i \neq j)} \log P(X,Z) \prod_{i \neq j}^{M} q_i(Z_i)\, dZ_i \right)}_{E_{\prod_{i\neq j} q_i(Z_i)}[\log P(X,Z)]} dZ_j$$

$$= \int_{z_j} q_j(z_j) \cdot E_{\prod_{i \neq j} q_i(z_i)}[\log P(x,z)] \cdot dz_j$$

$$② = \int_z q(z) \log q(z)\, dz$$
$$= \int_z \prod_{i=1}^{M} q_i(z_i) \sum_{i=1}^{M} \log q_i(z_i)\, dz$$
$$= \int_z \prod_{i=1}^{M} q_i(z_i) [\log q_1(z_1) + \log q_2(z_2) + \cdots + \log q_M(z_M)] dz$$

可写为M个积分相加.

其中一项：

$$\int_z \prod_{i=1}^{M} q_i(z_i) \log q_1(z_1)\, dz = \int_z q_1 q_2 \cdots q_M \cdot \log q_1\, dz$$
$$= \int_{z_1 z_2 \cdots z_M} q_1 q_2 \cdots q_M \log q_1\, dz_1 dz_2 \cdots dz_M$$
$$= \int_{z_1} q_1 \log q_1\, dz_1 \cdot \int_{z_2} q_2(z_2) dz_2 \cdots \int_{z_M} q_M(z_M) dz_M$$
$$= \int_{z_1} q_1 \log q_1\, dz_1$$

$$= \sum_{i=1}^{M} \int_{z_i} q_i(z_i) \log q_i(z_i)\, dz_i$$
$$= \int_{z_j} q_j(z_j) \log q_j(z_j)\, dz_j + C \quad (\text{因为现在只关心 } q_j)$$

①式：$\int_{z_j} q_j(z_j) \cdot \underbrace{E_{\prod_{i \neq j} q_i(z_i)}[\log P(x,z)]}_{\text{可写为 log 的形式，设为 } \log \hat{P}(x,z_j)} \cdot dz_j$

则 $L(q) = ① - ② = \int_{z_j} q_j(z_j) \log \frac{\hat{P}(x,z_j)}{q_j(z_j)}\, dz_j$

$$= -KL(q_j \| \hat{P}(x,z_j))$$

$\therefore q_j(z_j) = \hat{P}(x,z_j)$ 或 $\log q_j(z_j) = \log \hat{P}(x,z_j)$

三、变分推断再回首

对上一节进行澄清：

X：观测变量，$X=\{x^{(i)}\}_{i=1}^N$

Z：隐变量，$Z=\{z^{(i)}\}_{i=1}^N$

上一节中：$\log_\theta P(X) = \underbrace{ELBO}_{\mathcal{L}(q)} + \underbrace{KL(q\|p)}_{\geq 0} \geq \mathcal{L}(q)$

对此项，由于是似然，$\log_\theta P(X)=\log\prod_{i=1}^N P_\theta(x^{(i)})=\sum_{i=1}^N \log P_\theta(x^{(i)})$，对于MLE，由于样本独立同分布，因此MLE使$\log_\theta P(X)$达到最大，实际上是使每一项都达到最大。

改为：$\log P_\theta(x^{(i)}) = ELBO + KL(q\|p) \geq \mathcal{L}(q)$

目标函数：$\hat{q}=\arg\min_q KL(q\|p)=\arg\max_q \mathcal{L}(q)$

由上一节：

$$\log q_j(z_j) = E_{\prod_{i\neq j} q_i(z_i)}[\log P(x^{(i)},z)]dz_j + Const$$
$$= \int_{z_1}\int_{z_2}\cdots\int_{z_{j-1}}\int_{z_{j+1}}\cdots\int_{z_m} q_1\cdots q_{j-1}q_{j+1}\cdots q_m \cdot \log P(x^{(i)},z)\, dz_1\cdots dz_{j-1}dz_{j+1}\cdots dz_m$$

它实际上是一个迭代式，

$$\hat{q}_1(z_1)=\int_{q_2}\cdots\int_{q_m} q_2\cdots q_m[\log P_\theta(x^{(i)},z)]\,dz_2\cdots dz_m$$
$$\hat{q}_2(z_2)=\int_{q_1}\int_{q_3}\cdots\int_{q_m}\hat{q}_1 q_3\cdots q_m[\log P_\theta(x^{(i)},z)]\,dz_1dz_3\cdots dz_m$$
$$\vdots$$
$$\hat{q}_m(z_m)=\int_{\hat{q}_1}\int_{\hat{q}_2}\cdots\int_{q_{m-1}}\hat{q}_1\hat{q}_2\cdots\hat{q}_{m-1}[\log P_\theta(x^{(i)},z)]\,dz_1dz_2\cdots dz_{m-1}$$

坐标上升的思想。

Classical VI存在的问题：

① 假设太强（可分成M份相互独立的变量）

② 求其期望也会无法计算出来

四. 随机梯度变分推断 (SGVI)

前面用坐标上升的方法求 $q(z)$，很容易想到是否可以用梯度的方法求 $q(z)$

设 $q(z)=q_\phi(z)$，ϕ 是 $q(z)$ 分布的参数.

$$\hat{q} = \arg\max_q \mathcal{L}(q)$$ 应用梯度上升求 $\mathcal{L}(q)$ 对 q 的梯度

于是求解 q 变为求解 ϕ.

$\arg\max_q \mathcal{L}(q) \longrightarrow \arg\max_\phi \mathcal{L}(\phi)$

$\hat{\phi} = \arg\max_\phi \mathcal{L}(\phi)$

其中 $ELBO = E_{q_\phi(z)}\left[\log \frac{P_\theta(x^{(i)},z)}{q_\phi(z)}\right] = E_{q_\phi(z)}[\log P_\theta(x^{(i)},z) - \log q_\phi(z)] = \mathcal{L}(\phi)$

求解：

$$\begin{aligned}\nabla_\phi \mathcal{L}(\phi) &= \nabla_\phi E_{q_\phi}[\log P_\theta(x^{(i)},z) - \log q_\phi] \\ &= \nabla_\phi \int q_\phi [\log P_\theta(x^{(i)},z) - \log q_\phi]\, dz \\ &= \underbrace{\int \nabla_\phi q_\phi [\log P_\theta(x^{(i)},z) - \log q_\phi]\, dz}_{①} + \underbrace{\int q_\phi \nabla_\phi [\log P_\theta(x^{(i)},z) - \log q_\phi]\, dz}_{②}\end{aligned}$$

（$\log P_\theta(x^{(i)},z)$ 与 ϕ 无关）

$$\begin{aligned}② &= \int q_\phi \nabla_\phi [\log P_\theta(x^{(i)},z) - \log q_\phi]\, dz \\ &= -\int q_\phi \nabla_\phi \log q_\phi\, dz \\ &= -\int q_\phi \cdot \frac{1}{q_\phi} \nabla q_\phi\, dz \\ &= -\int \nabla_\phi q_\phi\, dz \\ &= -\nabla_\phi \underbrace{\int q_\phi\, dz}_{1} \\ &= 0\end{aligned}$$

因此，$\nabla_\phi \mathcal{L}(\phi) = ① = \int \nabla_\phi q_\phi [\log P_\theta(x^{(i)},z) - \log q_\phi]\, dz$

$$\begin{aligned}&= \int q_\phi \nabla_\phi \log q_\phi \cdot [\log P_\theta(x^{(i)},z) - \log q_\phi]\, dz \\ &= E_{q_\phi}[\nabla_\phi \log q_\phi \cdot (\log P_\theta(x^{(i)},z) - \log q_\phi)]\end{aligned}$$

这个期望可以通过 MC 采样来近似，从而得到梯度，然后利用梯度上升的方法来得到参数：

设采样样本为 $z^{(l)} \sim q_\phi(z)$，$l = 1, 2, \cdots, L$.

则 $E_{q_\phi}[\nabla_\phi \log q_\phi \cdot (\log P_\theta(x^{(i)},z) - \log q_\phi)] \approx \frac{1}{L} \sum_{l=1}^{L} \nabla_\phi \log q_\phi(z^{(l)}) (\log P_\theta(x^{(i)},z^{(l)}) - \log q_\phi(z^{(l)}))$

但由于存在对数项，如果采样得到的 $q_\phi(z^{(l)})$ 接近于0，log 的值会无穷小，打个比方说，如果两个采样点差一点点，log 的差距都会非常大，会导致期望与括号中式子的方差会非常大，就需要大量的样本才能得到较好的近似，L 如果过大，一般认为是不可行的。

因此采用重参数化技巧

重参数化技巧 (Reparameterization Trick)

由于 $\nabla_\phi \mathcal{L}(\phi) = \nabla_\phi E_{q_\phi}[\log P_\theta(x^{(i)}, z) - \log q_\phi]$，是对 q_ϕ 的分布求期望，如果我们可以对于一个不确定的分布求期望，计算会变得很容易。

假设 $z = g_\phi(\varepsilon, x^{(i)})$，$\varepsilon \sim p(\varepsilon)$

将 z 的随机性转移给 ε

$z \sim q_\phi(z|x^{(i)})$

$\varepsilon \sim p(\varepsilon)$

$\Longrightarrow |q_\phi(z|x^{(i)})\,dz| = |p(\varepsilon)\,d\varepsilon|$

则
$$\begin{aligned}
\nabla_\phi \mathcal{L}(\phi) &= \nabla_\phi E_{q_\phi}[\log P_\theta(x^{(i)}, z) - \log q_\phi] \\
&= \nabla_\phi \int [\log P_\theta(x^{(i)}, z) - \log q_\phi]\, q_\phi\, dz \\
&= \nabla_\phi \int [\log P_\theta(x^{(i)}, z) - \log q_\phi]\cdot p(\varepsilon)\, d\varepsilon \\
&= \nabla_\phi E_{p(\varepsilon)}[\log P_\theta(x^{(i)}, z) - \log q_\phi] \\
&= E_{p(\varepsilon)}[\nabla_\phi(\log P_\theta(x^{(i)}, z) - \log q_\phi)] \\
&= E_{p(\varepsilon)}[\nabla_z(\log P_\theta(x^{(i)}, z) - \log q_\phi)\cdot \nabla_\phi z] \\
&= E_{p(\varepsilon)}[\nabla_z(\log P_\theta(x^{(i)}, z) - \log q_\phi)\cdot \nabla_\phi g_\phi(\varepsilon, x^{(i)})]
\end{aligned}$$

链式法则 $\frac{\partial L}{\partial \phi} = \frac{\partial L}{\partial z}\cdot\frac{\partial z}{\partial \phi}$

进行 MC 采样，

$\varepsilon^{(l)} \sim p(\varepsilon)$

$l = 1, 2, \cdots, L$

$$\nabla_\phi \mathcal{L}(\phi) \approx \frac{1}{L}\sum_{l=1}^{L} \nabla_z[\log P_\theta(x^{(i)}, z) - \log q_\phi(z|x^{(i)})]\cdot \nabla_\phi g_\phi(\varepsilon^{(l)}, x^{(i)})$$

$z = g_\phi(\varepsilon^{(l)}, x^{(i)})$

SGVI：

$$\phi^{(t+1)} \leftarrow \phi^{(t)} + \lambda^{(t)}\cdot \nabla_\phi \mathcal{L}(\phi)$$