变分推断

Variational Inference

① 背景

② 公式推导

③ SGVI

背景(Background)

$\boxed{\text{频率角度} \to \text{优化问题}}$

① 线性回归:

$f(w) = w^T x$, 模型

$\boxed{\begin{aligned} L(w) &= \sum_{i=1}^{N} \|w^T x_i - y_i\|^2 \\ \hat{w} &= \arg\min_w L(w) \end{aligned}}$ 策略.

解法(算法) $\begin{cases} \text{解析解}: \frac{\partial L}{\partial w} = 0 \\ \text{数值解}: \text{GD (Gradient Descent)} \end{cases}$

② SVM:

$\boxed{f(w) = \mathrm{sign}(w^T x + b)}$ 模型

$\boxed{\begin{aligned} & L(w) = \frac{1}{2} w^T w \\ & \begin{cases} \hat{w} = \arg\min_w L(w) \\ s.t. \quad y_i(w^T x_i + b) \geq 1 \end{cases} \end{aligned}}$ 策略

$\boxed{\begin{cases} \text{QP} \\ \text{Lagrange} + \text{对偶}. \end{cases}}$ 算法.

③ EM.

$$\hat{\theta} = \arg\max_\theta \log p(X|\theta)$$

$$\Rightarrow \theta^{(t+1)} = \arg\max_\theta \int p(z|x, \theta^{(t)}) \log p(x, z|\theta)\, dz$$

迭代.

⋮

贝叶斯角度 ⟶ 积分问题

先验 → 似然

$$p(\theta|X) = \frac{p(\theta)\,p(X|\theta)}{p(X)}$$

↓ 后验

$p(X) = \int p(X,\theta)\,d\theta$.

贝叶斯推断：求后验 $p(\theta|X)$.

贝叶斯决策：$p(\tilde{x}|X)$

$$= \int_\theta p(\tilde{x}|\theta)\,p(\theta|X)\,d\theta.$$

$$= E_{\theta\sim p(\theta|X)}\left[p(\tilde{x}|\theta)\right]$$

概率模型的中心任务就是求潜变量的后验 $p(z|x)$，以及关于该后验的期望。

Inference　　decision.

实际上，后验很难求．原因有：①潜在空间维度过高；②后验分布形式复杂

对于连续变量，积分可能没有解析解，而空间维度高和被积函数复杂又使其没有数值解．

对于离散变量，原则上总是可以计算，但隐状态的个数可能有指数多个，精确计算所需的代价过高。

所以精确推断一般是不可行的，需要使用近似推断

Inference
- 精确推断
- 近似推断
  - 随机近似：MCMC（MH，Gibbs）
    特点：给无限多的计算资源 ⟶ 精确结果
    实际中，采样方法需要的计算量很大，且判断一种采样方法是否生成了所需的概率分布的独立样本是很困难的．
  - 确定性近似：VI（变分推断）
    假设后验分布可以通过某一种形式分解，或具有一个具体的参数形式（高斯分布）.
    特点：永远无法生成精确的解.

公式推导 (Formula Deduction)

变分法：传统微积分中，我们讨论的是 x 值的微小波动对 $y(x)$ 的影响；变分法中，我们讨论的是函数 $y(x)$ 的变化对泛函 $F(y)$ 的影响。

从而，在变分法中，我们可以寻找一个 $y(x)$ 来最大化或最小化泛函 $F(y)$ (PRML附录D)

泛函：我们可以将泛函作为一个映射，他接受一个函数作为输入，返回泛函的值作为输出。一个典型的例子是熵 $H[p]$.

$$H[p] = \int -p(x)\log p(x)\,dx.$$

X: observed variable

Z: latent variable + parameter

和之前EM一样：(VI的目的也在于 max $p(X)$)

$$\log p(X) = \log p(X,Z) - \log p(Z|X)$$

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

$$= \underbrace{\int q(Z)\log\frac{p(X,Z)}{q(Z)}\,dZ}_{\mathcal{L}(q)} - \underbrace{\int q(Z)\log\frac{p(Z|X)}{q(Z)}\,dZ}_{KL(q\|p)}.$$

泛函，作为变分的对象.

即 $q(Z) = \arg\max\limits_{q(Z)} \int q(Z) \log \frac{p(X,Z)}{q(Z)} dZ$.

假设 Z 可以划分为 M 个互不相交的组 $\rightarrow Z = \{Z_i\},\ i=1,2,\dots,M$

则 $q(Z) = \prod_{i=1}^{M} q_i(Z_i)$ 来源于统计物理学的平均场理论.

于是，我们可以对 $\mathcal{L}(q)$ 关于所有 $q_i(Z_i)$ 都进行变分最优化。在变分推断中，我们对每个 $q_i(Z_i)$ 最优化 从而完成总体最优化的过程.

对于 $q_j(Z_j)$ 项，我们将其从公式中分解出来.

$$\mathcal{L}(q) = \int \prod_{i=1}^{M} q_i \left[\log p(X,Z) - \log \prod_{k=1}^{M} q_k\right] dZ$$ （用 q_i 代替 $q_i(Z_i)$）

$$= \int q_j \left[\prod_{i \neq j} q_i \log p(X,Z)\right] dZ - \int \prod_{i=1}^{M} q_i \sum_{k=1}^{M} \log q_k \, dZ$$ $= \int \prod_{i=1}^{M} q_i \sum_{k \neq j} \log q_j \, dZ + \int \prod_{i=1}^{M} q_i \log q_j \, dZ$

$$= \int q_j \left[\int \log p(X,Z) \prod_{i \neq j} q_i \, dZ_i\right] dZ_j - \int q_j \log q_j \, dZ_j + C_1.$$

$$= \int q_j \, E_{i \neq j}[\log p(X,Z)] \, dZ_j - \int q_j \log q_j \, dZ_j + C_1.$$

令 $\log \tilde{p}(X, Z_j) = E_{i \neq j}[\log p(X,Z)] + C_2$

则 $\mathcal{L}(q) = \int q_j \log \frac{\tilde{p}(X, Z_j)}{q_j} dZ_j + C_3$

这里不太明白，为什么能将期望看作是一个概率分布取对数？

PRML中，后续会进行归一化，C_2 就是那个归一化常数 但视频中作者没有关心 C_2 的计算

$\tilde{p}$ 的随机变量除了 Z_j 还有 X，为什么在这种情况下还可以写成 KL Divergence？

$$= -KL(q_j \,\|\, \tilde{p}(X, Z_j)) + C_3$$

显然，当 $\log q_j(z_j) = E_{i\neq j}[\log p(X,z)] + C_2$ 时，

$KL(q_j \| \tilde{p}(X,z_j)) = 0$。此时，$q_j$ 取得当前条件下的最优值 $q_j^*(z_j)$

则两边取指数

$$q_j^*(z_j) = \exp\{E_{i\neq j}[\log p(X,z)] + C_2\}$$

$$\int q_j^*(z_j)\,dz_j = \exp\{C_2\}\int \exp\{E_{i\neq j}[\log p(X,z)]\}\,dz_j$$

$$C_2 = \log \frac{1}{\int \exp\{E_{i\neq j}[\log p(X,z)]\}\,dz_j}$$

实际上，我们在大部分计算时不关心 C_2，只在必要时求解即可。

对于整体的求解，我们采用和坐标上升类似的方法，即

$$\begin{cases} \text{固定 } q_2, q_3, \cdots, q_m,\ \text{计算 } q_1^* \\ \text{固定 } q_1^*, q_3, \cdots, q_m,\ \text{计算 } q_2^* \\ \text{固定 } q_1^*, q_2^*, q_4, \cdots q_m,\ \text{计算 } q_3^* \\ \vdots \\ \text{固定 } q_1^*, q_2^*, \cdots, q_{m-1},\ \text{计算 } q_m^* \end{cases}$$

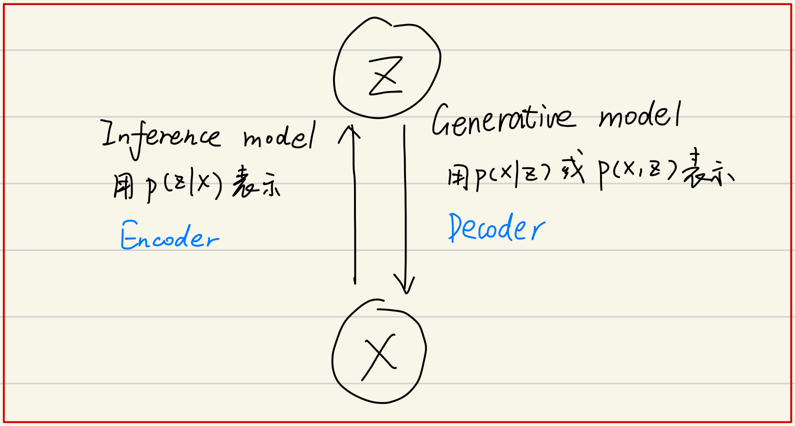
$\Rightarrow$ 以此为一轮循环，不断计算，直至收敛

算法保证收敛，因为下界关于每一个 q_i 都是凸函数。（PRML中有给出文献）

经典VI的问题：① 平均场理论的假设过强

② 期望在某些情况下仍然 intractable

SGVI



基于平均场理论的VI $\longrightarrow$ 坐标上升 (coordinate ascent)

本节介绍基于梯度上升的VI (SGVI)

我们将 $q(z)$ 看作是以ϕ为参数的分布，记为 $q_\phi(z)$

则 $$p(X) = \underbrace{\int q_\phi(z) \log \frac{p(X,z)}{q_\phi(z)} dz}_{L(\phi)} - \int q_\phi(z) \log \frac{p(z|X)}{q_\phi(z)} dz$$

则可以将目标写为 $\hat{\phi} = \underset{\phi}{\arg\max}\, L(\phi)$

$$L(\phi) = \int q_\phi(z) \log \frac{p(X,z)}{q_\phi(z)} dz$$

$$= \int q_\phi \log \frac{p(X,z)}{q_\phi} dz.$$

则 $\nabla_\phi L(\phi) = \nabla_\phi \int q_\phi \log \frac{p(x,z)}{q_\phi} dz$

$$= \int \nabla_\phi \left[q_\phi (\log p(x,z) - \log q_\phi) \right] dz$$

$$= \int \left[\nabla_\phi q_\phi (\log p(x,z) - \log q_\phi) - \nabla_\phi q_\phi \right] dz .$$

$$\int \nabla_\phi q_\phi dz = \nabla_\phi \int q_\phi dz = \nabla_\phi 1 = 0$$

$$= \int \nabla_\phi q_\phi (\log p(x,z) - \log q_\phi) dz .$$

$$\nabla_\phi q_\phi = q_\phi \nabla_\phi \log q_\phi$$

目的在于提一个q_ϕ出来.
从而能够写成期望的形式.

为什么要写成期望的形式？因为我们想让其结果可以通过蒙特卡罗的方式，通过采样可以得到.

$$= \int q_\phi \nabla_\phi \log q_\phi (\log p(x,z) - \log q_\phi) dz$$

$$= E_{q_\phi} \left[\nabla_\phi \log q_\phi (\log p(x,z) - \log q_\phi) \right] .$$

$z^{(l)} \sim q_\phi(z)$, $l = 1, 2, \cdots, L$

$$\approx \frac{1}{L} \sum_{i=1}^{L} \nabla_\phi \log q_\phi(z^{(l)}) \left(\log p(x,z) - \log q_\phi(z^{(l)}) \right)$$

但这会有一个弊端，当 $q_\phi(z^{(l)}) \approx 0$ 时，$\nabla_\phi \log q_\phi(z^{(l)})$ 的会很大.
从而导致整体的方差会很大（误差大），那就意味着需要很多采样点才能得到较为可靠的值，因为在实际中是很难实现的.

我们可以用重参数化技巧（reparameterization trick）解决该问题.

令 $z = g_\phi(\varepsilon, X)$

$\varepsilon \sim p(\varepsilon)$

从而我们将随机性转移到了 ε 上面

例子. 一维高斯分布：$p(x) = N(x|\mu, \sigma^2)$

我们其实可以用标准正态分布来生成 x，而不是直接从 $p(x)$ 中采样.

$x = \mu + \sigma \cdot \epsilon$ ，$\epsilon \sim N(0,1)$ 扩散模型中就用到了该技巧.

从而随机性从 x 转移到了 ϵ

在 VI 中.

$$\nabla_\phi L(\phi) = \nabla_\phi \int q_\phi(z) [\log p(X,z) - \log q_\phi(z)] dz$$

还有一个结论 $q_\phi(z) dz = p(\varepsilon) d\varepsilon$（没记明白，不知道怎么得出来的）.

$$= \nabla_\phi \int [\log p(X,z) - \log q_\phi(z)] p(\varepsilon) d\varepsilon.$$

$$= \int \nabla_z [\log p(X,z) - \log q_\phi(z)] \nabla_\phi z \, p(\varepsilon) d\varepsilon.$$

$$= E_{p(\varepsilon)} \{ \nabla_z [\log p(X,z) - \log q_\phi(z)] \nabla_\phi g_\phi(\varepsilon, X) \}.$$

从而，我们不需要从 $q_\phi(z)$ 中采样，只需要从已知的 $p(\varepsilon)$ 中采样即可.

$\varepsilon^{(n)} \sim p(\varepsilon)$ ，$n=1,2,\dots,N$ ，$\Rightarrow \nabla_\phi L(\phi) \approx \frac{1}{N} \sum_{i=1}^{N} \nabla_z [\log p(X,z) - \log q_\phi(z)] \nabla_\phi g_\phi(\varepsilon^{(n)}, X)$

则 $\phi^{(t+1)} = \phi^{(t)} + \lambda^{(t)} \nabla_\phi L(\phi)$

为什么从 $p(\varepsilon)$ 中采样可以避免 $q_\phi(z) = 0$？这里不太明白

同时PRML中VI一章中还介绍了关于VI的其他方法以及应用，可以看一下.