白板 (12)

变分推断 - 背景

Variational Inference

频率派 → 优化问题

- 回归：① $f(w)=w^T \cdot x$ (model)，② loss function 策略 $\begin{cases} L(w)=\sum_{i=1}^{N}\|w^T x_i - y_i\|^2 \\ \hat{w}=\arg\min L(w) \end{cases}$
  - ③ 算法
    - 解析解 $\frac{\partial L(w)}{\partial w}=0 \Rightarrow w^* = \underbrace{(X^T X)^{-1} X^T}_{\text{MP逆矩阵}} Y$
    - 数值解：Gradient Descend / SGD
- SVM (分类问题)
  - ① $f(w)=\mathrm{sign}(w^T x + b)$
  - ② loss function：
    $\begin{cases} \min \frac{1}{2} w^T w \\ s.t.\ y_i(w^T x_i + b) \geq 1,\ i=1\cdots N \end{cases}$ 有约束的优化问题
  - ③ QP、拉格朗日.对偶.
- EM：$\hat{\theta}=\arg\max_{\theta} \cancel{P(x|\theta)} \log p(x|\theta)$.

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

贝叶斯角度 — 积分问题 $\underbrace{p(\theta|x)}_{\text{后验}}=\frac{\overbrace{p(x|\theta)}^{\text{似然}}\cdot\overbrace{p(\theta)}^{\text{先验}}}{p(x)}$，$p(x)$ 积分 $=\int_\theta p(x|\theta)\cdot p(\theta)\,d\theta$

贝叶斯 inference：求后验 →
- 精确推断
- 近似推断
  - 确定性近似 VI
  - 随机性近似 MCMC / MH / Gibbs.

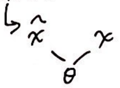
贝叶斯 决策(预测) $\tilde{x}$

给新样本 $\tilde{x}$，求 $p(\tilde{x}|x)$

$p(\tilde{x}|x)=\int_\theta p(\tilde{x},\theta|x)\cdot d\theta=\int_\theta p(\tilde{x}|\theta)\cdot\underbrace{p(\theta|x)}_{\text{后验}}d\theta$ = 可理解成关于后验求期望(积分)

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

引入 θ



白板（12）　　变分推断—公式推导

x: observed data

z: latent variable + parameter
[隐变量 + 参数]

(x, z): complete data

θ也看作 z，合并在 z 里面.

做一个 EM.

$$\log P(x) = \log P(x,z) - \log P(z|x)$$

引入 $q(z)$ $= \log \frac{P(x,z)}{q(z)} - \log \frac{P(z|x)}{q(z)}$

对 $q(z)$ 求期望.

左边 $= \int_z \log P(x) \cdot q(z) dz = \log P(x) \cdot \underbrace{\int_z q(z) dz}_{=1}$ （与 z 无关）

$\therefore$ 左 $= \log P(x)$

右边 $= \underbrace{\int_z q(z) \cdot \log \frac{P(x,z)}{q(z)} dz}_{\text{ELBO}} - \underbrace{\int_z q(z) \log \frac{P(z|x)}{q(z)} dz}_{KL(q \| P) \geq 0}$

$= \mathcal{L}(q) + \underset{\geq 0}{KL(q \| p)}$

$\therefore$ 找一个 $\mathcal{L}(q)$，让其接近 $\log P(x)$. 那么 KL 就接近 0 $\Rightarrow q(z) \approx p(z|x)$

$\because \tilde{q}(z) = \underset{q(z)}{\arg\max}\, \mathcal{L}(q) \Rightarrow \tilde{q}(z) \approx p(z|x)$ ▲

假设 $q(z)$ 可以分成 M 个，且其相互独立. $\Rightarrow$ 平均场理论 mean theory

$$q(z) = \prod_{i=1}^{M} q_i(z_i)$$

右边 $\mathcal{L}(q) = \underbrace{\int_z q(z) \log P(x,z) dz}_{①} - \underbrace{\int_z q(z) \log q(z) \cdot dz}_{②}$

刚刚设的 $L(q)$

拆成 $z_1, z_2 \cdots z_m$

$$① = \int_z \prod_{i=1}^{M} q_i(z_i) \cdot \log P(x,z) dz_1 dz_2 \cdots dz_m$$

解析 $q_j(z_j)$，其它固定 ▲

$$= \int_z q_j(z_j) \underbrace{\left(\int_{\substack{z_1 \cdots z_m \\ i \neq j}} \prod_{i \neq j}^{M} q_i(z_i) \log P(x,z) dz_1 \cdots dz_m \right)}_{\underbrace{\int_{z_1 \cdots z_m (i \neq j)} P(x,z) \prod_{i \neq j}^{M} q_i(z_i) dz_i}_{= E_{\prod_{i \neq j}^{M} q_i(z_i)}[\log P(x,z)]}} \cdot dz_j \quad (i \neq j)$$

最终 ① $= \int_{z_j} q_j(z_j) \cdot \left\{ \boxed{E_{\prod_{i=1}^{N} q_i(z_i)}[\log P(x,z)]} \right\} dz_j$

令其为 $\log \hat{p}(x, z_j)$

$$② = \int_z q(z) \log q(z) \cdot dz = \int_z \prod_{i=1}^{M} q_i(z_i) \cdot \log \prod_{i=1}^{M} q_i(z_i) \cdot dz$$

$q(z)$ 省略写 $q(z)$

$$= \int_z \prod_{i=1}^{M} q_i(z_i) \cdot \sum_{i=1}^{M} \log q_i(z_i) \cdot dz$$

$$= \int_z \prod_{i=1}^{M} q_i(z_i) \left[\log q_1(z_1) + \log q_2(z_2) + \cdots + \log q_m(z_m) \right] \cdot dz$$

$\to$ 其中某一项 $\int_z \prod_{i=1}^{M} q_i(z_i) \cdot \log q_1(z_1)$ 为例

$= \int_z q_1 q_2 \cdots q_m \cdot \log q_1 dz$

$= \int_{z_1 z_2 \cdots z_m} q_1 q_2 \cdots q_m \log q_1 dz_1 dz_2 \cdots dz_m$

$= \int_{z_1} q_1 \log q_1 dz_1 \cdot \underbrace{\int_{z_2} q_2 dz_2}_{=1} \cdots \underbrace{\int_{z_m} q_m dz_m}_{=1}$

$= \int_{z_1} q_1 \log q_1 dz_1$

$$= \text{看成每项推导} = \sum_{i=1}^{M} \int_{z_i} q_i(z_i) \cdot \log q_i(z_i) \cdot dz_i$$

把 $q_j(z_j)$ 之外看作已知（固定）▲ $= \int_{z_j} q_j(z_j) \log q_j(z_j) \cdot dz_j + C \to$ 常数

那么 ① − ② $= \int_{z_j} q_j(z_j) \cdot \log \frac{\hat{p}(x, z_j)}{q_j(z_j)} \cdot dz_j + \cancel{C}$ → 常数略掉

$= -KL(\underset{\downarrow\, q_j(z_j)}{q_j} \| \hat{p}(x, z_j)) \leq 0$. 取等号时. $q_j(z_j) = \hat{p}(x, z_j)$

变分推断（VI）

x: observed variable $\rightarrow X = \{x_i\}_{i=1}^N = \{x^{(i)}\}_{i=1}^N$

z: latent variable $\rightarrow Z = \{z_i\}_{i=1}^N = \{z^{(i)}\}_{i=1}^N$

(x, z): complete data

θ: model parameter

$$ELBO = E_{q(z)}\left[\log \frac{P(X,Z|\theta)}{q(z)}\right]$$

$$= E_{q(z)}[\log P(X,Z|\theta)] + H[q(z)]$$

$$KL(q\|p) = \int_Z q(z) \cdot \log \frac{q(z)}{p(z|x,\theta)} \cdot dz$$

可假定θ已求出，或者是一个常数，$\Rightarrow \mathcal{L}(q)$

$$\log P(X|\theta) = \underbrace{ELBO}_{\mathcal{L}(q,\theta)} + \underbrace{KL(q\|p)}_{\geq 0} \geq \mathcal{L}(q,\theta)$$

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

假定θ已知

$$\log P_\theta(x^{(i)}) = \underbrace{ELBO}_{\mathcal{L}(q)} + \underbrace{KL(q\|p)}_{\geq 0} \geq \mathcal{L}(q)$$

视频在此进行了变量写法的轻微改变，样本写入 $x^{(i)}$ 右上角

$$P_\theta(x^{(i)}, z) = P(x, z|\theta)$$

新 旧

VI：

前提假设 Assumption：$q(z) = \prod_{i=1}^{M} q_i(z_i)$ → 平均场理论 ▲

$$\log q_j(z_j) = E_{\prod_{i \neq j} q_i(z_i)}[\log P(X,Z|\theta)] + C$$

C：常数

$$= \int_{q_1}\int_{q_2}\cdots\int_{q_{j-1}}\int_{q_{j+1}}\cdots\int_{q_m} q_1 q_2 \cdots q_{j-1} q_{j+1} \cdots q_m [\log P_\theta(x^{(i)}, z)] \, dq_1 dq_2 \cdots dq_{j-1} dq_{j+1} \cdots dq_m$$

▲迭代优化过程

$$\hat{q}_1(z_1) = \int_{q_2}\cdots\int_{q_m} q_2 \cdots q_m [\log P_\theta(x^{(i)}, z)] \, dq_2 \cdots dq_m$$

使用 $\hat{q}_1$

$$\hat{q}_2(z_2) = \int_{\hat{q}_1}\int_{q_3}\cdots\int_{q_m} \hat{q}_1 q_3 \cdots q_m [\log P_\theta(x^{(i)}, z)] \, d\hat{q}_1 \cdots dq_m$$

$\vdots$

$$\hat{q}_m(z_m) = \int_{\hat{q}_1}\int_{\hat{q}_2}\cdots\int_{\hat{q}_{m-1}} \hat{q}_1 \hat{q}_2 \cdots \hat{q}_{m-1} [\log P_\theta(x^{(i)}, z)] \, d\hat{q}_1 \cdots d\hat{q}_{m-1}$$

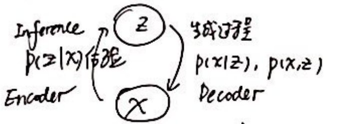
▲ 坐标~~下降~~（上升）法 coordinate Ascend.

判停问题：$\mathcal{L}^{(t+1)} \leq \mathcal{L}^{(t)}$ 时停止。

以上就是经典的基于平均场理论的变分推断问题。

存在问题 ① 一些模型不存在（不满足）平均场理论，如神经网络

② 此处求积分并不是都可以求出来 intractable 例如 $p(z|x)$ 后验，求积分不可求



随机梯度变分推断（SGVI）

$\theta^{(t+1)} \leftarrow \theta^{(t)} + \lambda \cdot$ 梯度方向，求 $\mathcal{L}$ 梯度

$q(z)$ 其实是关于 x 的概率分布 $q(z|x)$。

$q(z)$ 是一个分布，假定其参数为 ϕ，最好的 ϕ 求出来，那么最好 q 也求出来了。

$q(z) = q_\phi(z)$，求 ϕ。那么 ELBO 写成相减形式 $= E_{q_\phi(z)}[\log P_\theta(x^{(i)}, z) - \log q_\phi(z)] = \mathcal{L}(\phi)$

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

梯度 $\nabla_\phi \mathcal{L}(\phi) = \nabla_\phi E_{q_\phi}[\log P_\theta(x^{(i)}, z) - \log q_\phi]$（把 z 省略不写了）

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

—— 接入下一张纸

白板(12)　　随机梯度变分推断(SGVI)

承接上张笔记，交换 $\int$ 与 ∇ $\cancel{= \int \nabla_\phi q_\phi (\log P_\theta(x^{(i)}, z) - \log q_\phi) dz +}$

$\cancel{\int q_\phi \nabla_\phi}$

$$= \int \nabla_\phi q_\phi \cdot (\log P_\theta(x^{(i)}, z) - \log q_\phi) dz + \int q_\phi \nabla_\phi [\underbrace{\log P_\theta(x^{(i)}, z)}_{与\phi无关} - \log q_\phi] dz$$

①　②

$= ① + ②$

$$② = -\int q_\phi \nabla_\phi \log q_\phi dz = -\int q_\phi \frac{1}{q_\phi} \nabla_\phi q_\phi \cdot dz = -\int \nabla_\phi q_\phi dz = -\nabla_\phi \underbrace{\int q_\phi dz}_{=1} = 0$$

对常数求梯度

$$① = \underbrace{\int q_\phi \nabla_\phi \log q_\phi}(\log P_\theta(x^{(i)}, z) - \log q_\phi) dz$$

$\hookrightarrow = \int \nabla_\phi q_\phi$ 把 $\log q_\phi$ 增加出来，② 使用过.

$$= E_{q_\phi}[\nabla_\phi \log q_\phi (\log P_\theta(x^{(i)}, z) - \log q_\phi)]$$ (梯度用期望表示出来)

$q_\phi(z)$

$z^{(l)} \sim q_\phi(z)$

$l = 1, 2 \cdots L$

∇_ϕ 梯度可以表示成 $\cancel{E_{q_\phi}}$ q_ϕ 的期望. 采用 mcmc，蒙特卡罗方式去估计

采样

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

$z^{(l)} \sim q_\phi(z)$　$l = 1, 2 \cdots L$ ~ mcmc 采样 $\approx \frac{1}{L}\sum_{l=1}^{L} \nabla_\phi \log q_\phi(z^{(l)})(\log P_\theta(x^{(i)}, z^{(l)}) - \log q_\phi(z^{(l)}))$

上面的 $\log q_\phi$ 项，

可能会 $-\infty$，很大.

导致 high variance. 方差↑，要求样本↑，但是 L 个是

不可行的.

[还有部分没看]……