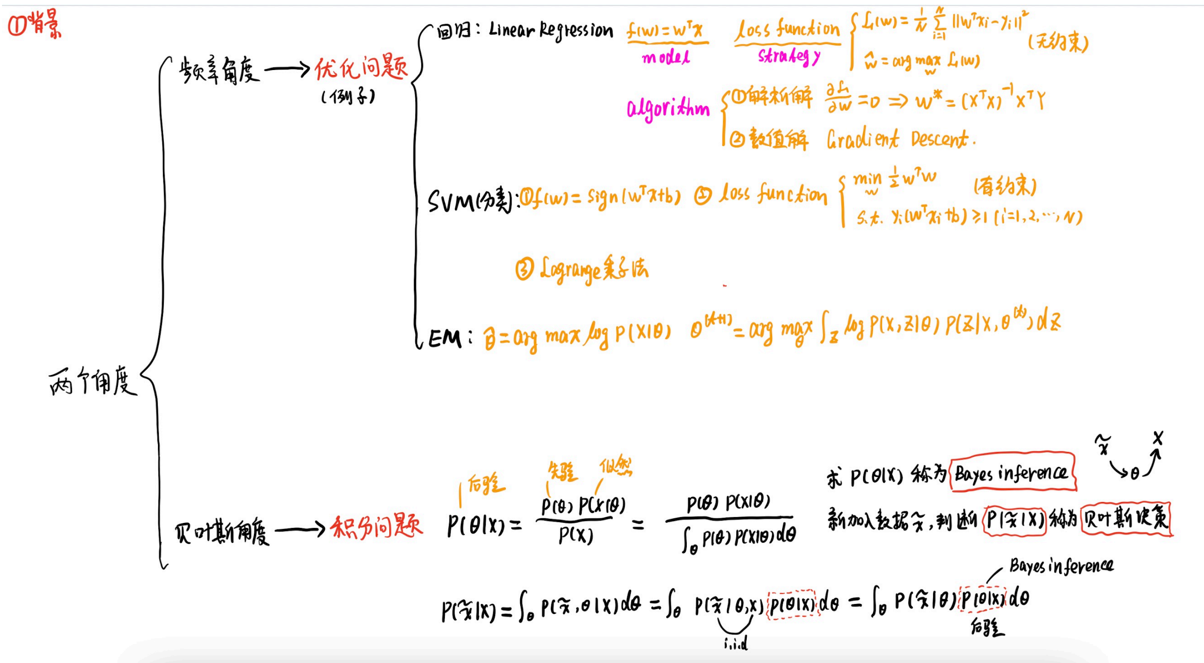
# 机器学习白板推导系列笔记 (12.1∼12.5)

Dexing Huang
dxhuang@bupt.edu.cn
Beijing University of Posts and Telecommunications

日期：2021 年 11 月 24 日

## 1 Inference 的进一步阐述




在引入变分推断之前, 对 Inference 进行进一步的说明, 也就是阐述为什么需要 Inference. 机器学习可以从两个角度分析和解决问题, 一个是频率角度, 另一个是贝叶斯角度. 从频率角度将问题转化为优化问题, 如线性回归, 支持向量机以及 EM 算法等. 而贝叶斯角度将其转化为积分问题, 何为积分问题呢.

在贝叶斯派的认识里, 参数 $\theta$ 也是一个概率分布, 因此终极目的是求出后验 $P(\theta|X)$, 这一过程称为Bayesian Inference. 得到后验后可以干什么呢, 若出现了一个新的数据 $\hat{x}$, 那么可以求出其概率分布 $P(\hat{x}|X)$.

$$P(\hat{x}|X) = \int_\theta P(\hat{x},\theta|X)d\theta = \int_\theta P(\hat{x}|\theta,X)P(\theta|X)d\theta = \int_\theta P(\hat{x}|\theta)P(\theta|X)d\theta \tag{1}$$

因此重点是后验 $P(\theta|X)$ 的求解

$$P(\theta|X) = \frac{P(\theta)P(X|\theta)}{P(X)} = \frac{P(\theta)P(X|\theta)}{\int_\theta P(\theta)P(X|\theta)d\theta} \tag{2}$$

从上式可以看到, 分母积分的求解是这个问题的关键, 通常这个积分是非常难解的. 因此需要一些 Inference 的方法来求出 $P(\theta|X)$, 通常 Inference 分为两种

- 精确推断

  在 9. 中讲过几种精确推断的方法, 如变量消除法以及信念传播算法
- 近似推断

  分为确定性近似和随机性近似. 确定性近似的代表就是本章要说的变分推断, 随机性近似包括之后要讲的MCMC, Gibbs 采样等.

下面开始进入正题, 变分推断.

## 2 基于平均场理论的变分推断

先做一下符号说明, $X$ 为观测变量, $Z$ 包括隐变量以及参数. 因为变分推断的目标是找到后验的近似分布, 因此把 $\theta$ 包含在 $Z$ 中. VI 的思想和 EM 基本类似, 即最大化对数似然函数 $\log P(X)$.

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

对 $q(Z)$ 求期望

$$\log P(X) = \underbrace{\int_Z q(Z) \log \frac{P(X,Z)}{q(Z)} dZ}_{\mathcal{L}(q)} - \underbrace{\int_Z q(Z) \log \frac{P(Z|X)}{q(Z)} dZ}_{KL(q(Z)||P(Z|X))} \tag{3}$$

$P(Z|X)$ 就是要求的后验, 它求不出来, 因此需要找到一个分布来逼近它, 根据分析知道

$$\log P(X) \geq \mathcal{L}(q) \tag{4}$$

那么如果让 $\mathcal{L}(q)$ 尽可能大, 大到取等号, 那么

$$q(Z) \approx P(Z|X)$$

**注.** 因为 KL 散度是衡量两个概率分布相似性程度的, 且非负

就可以认为找到了一个近似分布 $q(Z)$ 可以表示后验 $P(Z|X)$.

$$\hat{q}(Z) = \arg\max_q \mathcal{L}(q) \tag{5}$$

那么该如何求解呢, 这里需要注意, $q(Z)$ 是人为设置的, 因此也出现了许多种假设的方法. 在这里用到一种统计物理学中平均场理论的方法, 设分布 $q(Z)$ 可以拆分为 $M$ 块相互独立的部分, 那么

$$q(Z) = \prod_{i=1}^{M} q_i(Z_i) \tag{6}$$

那么 $\mathcal{L}(q)$ 化为

$$\begin{aligned}
\mathcal{L}(q) &= \int_Z q(Z) \log P(X,Z) dZ - \int_Z q(Z) \log q(Z) dZ \\
&= \underbrace{\int_Z \prod_{i=1}^{M} q_i(Z_i) \log P(X,Z) dZ}_{(1)} - \underbrace{\int_Z \prod_{i=1}^{M} q_i(Z_i) \log q(Z) dZ}_{(2)}
\end{aligned}$$

对于式 (1), 假设先考虑 $q_j$ 分量

$$
\begin{aligned}
(1) &= \int_{Z_j} q_j(Z_j)\left(\int_{Z/Z_j} \prod_{i=1/j}^{M} q_i(Z_i) \log P(X,Z) dZ/Z_j\right) dZ_j \\
&= \int_{Z_j} q_j(Z_j) \mathbb{E}_{\prod_{i=1/j}^{M} q_i(Z_i)}[\log P(X,Z)] dZ_j
\end{aligned} \tag{7}
$$

对于式 (2), 有

$$
\begin{aligned}
(2) &= \int_{Z} \prod_{i=1}^{M} q_i(Z_i) \log q(Z) dZ \\
&= \sum_{i=1}^{M} \int_{Z_i} q_i(Z_i) \log q_i(Z_i) dZ_i \\
&= \int_{Z_j} q_j(Z_j) \log q_j(Z_j) dZ_j + const \quad (\text{因为只关心 } q_j)
\end{aligned} \tag{8}
$$

**注.** 将式 (6) 代入, 将每项加法拆开易得

设 $\mathbb{E}_{\prod_{i=1/j}^{M} q_i(Z_i)}[\log P(X,Z)] + const = \log \tilde{P}(X,Z_j)$

$$
\begin{aligned}
(1)-(2) &= \int_{Z_j} q_j(Z_j) \log \tilde{P}(X,Z_j) dZ_j - \int_{Z_j} q_j(Z_j) \log q_j(Z_j) dZ_j + const \\
&= \int_{Z_j} q_j(Z_j) \log \frac{\tilde{P}(X,Z_j)}{q_j(Z_j)} dZ_j + const \\
&= -KL(q_j(Z_j)||\tilde{P}(X,Z_j)) \le 0
\end{aligned} \tag{9}
$$

因此, 当 $q_j(Z_j) = \tilde{P}(X,Z_j)$ 时 $\mathcal{L}(q)$ 取得最大值, 即

$$
\log q_j(Z_j) = \mathbb{E}_{\prod_{i=1/j}^{M} q_i(Z_i)}[\log P(X,Z)] + const \tag{10}
$$

$$
\begin{aligned}
q_j(Z_j) &= \exp\{\mathbb{E}_{\prod_{i=1/j}^{M} q_i(Z_i)}[\log P(X,Z)]\} \cdot exp\{const\} \\
\int_{Z_j} q_j(Z_j) dZ_j &= \int_{Z_j} \exp\{\mathbb{E}_{\prod_{i=1/j}^{M} q_i(Z_i)}[\log P(X,Z)]\} \cdot exp\{const\} dZ_j \\
\exp\{const\} &= \frac{1}{\int_{Z_j} \exp\{\mathbb{E}_{\prod_{i=1/j}^{M} q_i(Z_i)}[\log P(X,Z)]\} dZ_j}
\end{aligned}
$$

那么

$$
q_j(Z_j) = \frac{\exp\{\mathbb{E}_{\prod_{i=1/j}^{M} q_i(Z_i)}[\log P(X,Z)]\}}{\int_{Z_j} \exp\{\mathbb{E}_{\prod_{i=1/j}^{M} q_i(Z_i)}[\log P(X,Z)]\} dZ_j} \tag{11}
$$

由上式子可以看出, 该过程是迭代的, 并且被证明是收敛的.

下面来简单说说基于平均场理论的变分推断的不足, 首先是 $q(Z)$ 的假设本身就太强了, 大部分都不适用. 同时这种迭代的方法计算量也很大, 有时候也无法计算, 因此需要对变分推断进行优化.

# 3 SGVI

基于平均场理论的变分推断更新的方式相当于坐标上升法, $q(Z)$ 的假设也存在一些问题. 那么就想是否可以使用梯度的方法来得到最优的 $q(Z)$.

假设分布 $q_\phi(X)$ 含有参数 $\phi$, 当参数给定那么分布也就确定, 因此 (5) 可以变为

$$\hat{\phi} = \arg\max_{\phi} \mathcal{L}(\phi) \tag{12}$$

$$\begin{aligned}\mathcal{L}(\phi) &= \int_Z q_\phi(Z) \log P(X,Z) dZ - \int_Z q_\phi(Z) \log q_\phi(Z) dZ \\ &= \mathbb{E}_{q_\phi(Z)}[\log P(X,Z) - \log q_\phi(Z)]\end{aligned} \tag{13}$$

参数的更新公式为

$$\phi^{(t+1)} \leftarrow \phi^{(t)} + \lambda \nabla_\phi \mathcal{L}(\phi) \tag{14}$$

因此重点是求出梯度 $\nabla_\phi \mathcal{L}(\phi)$

$$\begin{aligned}\nabla_\phi \mathcal{L}(\phi) &= \nabla_\phi \mathbb{E}_{q_\phi(Z)}[\log P(X,Z) - \log q_\phi(Z)] \\ &= \nabla_\phi \int_Z q_\phi(Z)(\log P(X,Z) - \log q_\phi(Z)) dZ \\ &= \int_Z \nabla_\phi \{q_\phi(Z)(\log P(X,Z) - \log q_\phi(Z))\} dZ \\ &= \int_Z \nabla_\phi q_\phi(Z) \cdot (\log P(X,Z) - \log q_\phi(Z)) dZ + \int_Z \nabla_\phi (\log P(X,Z) - \log q_\phi(Z)) \cdot q_\phi(Z) dZ \\ &= \int_Z \nabla_\phi q_\phi(Z) \cdot (\log P(X,Z) - \log q_\phi(Z)) dZ - \int_Z \nabla_\phi \log q_\phi(Z) \cdot q_\phi(Z) dZ\end{aligned}$$

先来计算 $\int_Z \nabla_\phi \log q_\phi(Z) \cdot q_\phi(Z) dZ$

$$\begin{aligned}\int_Z \nabla_\phi \log q_\phi(Z) \cdot q_\phi(Z) dZ &= \int_Z \frac{1}{q_\phi(Z)} q_\phi(Z) \nabla_\phi q_\phi(Z) dZ \\ &= \int_Z \nabla_\phi q_\phi(Z) dZ \\ &= \nabla_\phi \int_Z q_\phi(Z) dZ \\ &= \nabla_\phi 1 = 0\end{aligned}$$

因此

$$\nabla_\phi \mathcal{L}(\phi) = \int_Z \nabla_\phi q_\phi(Z) \cdot (\log P(X,Z) - \log q_\phi(Z)) dZ \tag{15}$$

又因为

$$\nabla_\phi q_\phi(Z) = q_\phi(Z) \nabla_\phi \log q_\phi(Z) \tag{16}$$

那么

$$\begin{aligned}\nabla_\phi \mathcal{L}(\phi) &= \int_Z q_\phi(Z) \nabla_\phi \log q_\phi(Z) \cdot (\log P(X,Z) - \log q_\phi(Z)) dZ \\ &= \mathbb{E}_{q_\phi(Z)}[\nabla_\phi \log q_\phi(Z) \cdot (\log P(X,Z) - \log q_\phi(Z))]\end{aligned} \tag{17}$$

上式期望可以采用蒙特卡洛采样的方法求得, 选取 $Z^{(l)} \sim p_\phi(Z), l = 1, 2, \cdots, L$

$$\begin{aligned}\nabla_\phi \mathcal{L}(\phi) &= \mathbb{E}_{q_\phi(Z)}[\nabla_\phi \log q_\phi(Z) \cdot (\log P(X,Z) - \log q_\phi(Z))] \\ &\approx \frac{1}{L} \sum_{l=1}^{L} \nabla_\phi \log q_\phi(Z^{(l)}) \cdot [\log P(X, Z^{(l)}) - \log q_\phi(Z^{(l)})]\end{aligned} \tag{18}$$

但是上述方法也有缺点, 观察 $\nabla_\phi \log q_\phi(Z^{(l)})$ 项, 若采样得到的样本在 0 附近会造成值不稳定, 即方差很大的情况, 方差大, 而且本身通过参数估计分布也有误差 $\phi \to p_\phi(Z)$, 误差叠加该方

法就根本没什么用. 为了减小方差需要增加采样点数, 但是无限增加采样点数目显然也是不可行的.

## 4 Reparameterization Trick

减少方差有很多办法, 这里使用的是 Reparameterization Trick[1].

假设一随机变量 $\epsilon \sim p(\epsilon)$(这是人为设置的已知的), 且 $Z$ 跟 $\epsilon$ 存在以下关系

$$Z = g_\phi(\epsilon, X) \tag{19}$$

那么

$$\begin{aligned}
\nabla_\phi \mathcal{L}(\phi) &= \nabla_\phi \mathbb{E}_{q_\phi(Z)}[\log P(X,Z) - \log q_\phi(Z)] \\
&= \nabla_\phi \mathbb{E}_{p(\epsilon)}[\log P(X,Z) - \log q_\phi(Z)] \\
&= \mathbb{E}_{p(\epsilon)}[\nabla_Z(\log P(X,Z) - \log q_\phi(Z))\nabla_\phi Z] \\
&= \mathbb{E}_{p(\epsilon)}[\nabla_Z(\log P(X,Z) - \log q_\phi(Z))\nabla_\phi g_\phi(\epsilon, X)]
\end{aligned}$$

最后对 $\epsilon$ 采样, 得 $\{\epsilon^{(l)}\}_{j=1}^L$, 那么

$$\nabla_\phi \mathcal{L}(\phi) \approx \sum_{l=1}^{L} \nabla_Z(\log P(X,Z) - \log q_\phi(Z))\nabla_\phi g_\phi(\epsilon^{(l)}, X) \tag{20}$$

**注.** $Z = g(\epsilon^{(l)}, X)$

这里有些步骤并没有进行严格的数学证明, 先意会即可, 更多的内容在变分自编码器 (VAE) 中.

[1]https://gregorygundersen.com/blog/2018/04/29/reparameterization/