Variational Inference
变分推断

梁伟轩
weixuanliang@nudt.edu.cn

计算机学院, NUDT, 长沙

2020/11

Contents:

问题构建

首先我们定义要解决的问题。

推断问题

观测变量 $\mathbf{x} = \mathbf{x}_{1:n}$，而 $\mathbf{z} = \mathbf{z}_{1:m}$ 是相关的隐变量。推断问题是基于以上观测值，计算隐变量的条件概率 $p(\mathbf{z}|\mathbf{x})$ 密度。

而上述条件密度函数在各方面有重要的应用，例如生成新的样本点等等。$p(\mathbf{z}|\mathbf{x})$ 可以进一步化为：

$$p(\mathbf{z}|\mathbf{x}) = \frac{p(\mathbf{z},\mathbf{x})}{p(\mathbf{x})} = \frac{p(\mathbf{z},\mathbf{x})}{\int_{\mathbf{z}} p(\mathbf{z},\mathbf{x})\, d\mathbf{z}}$$

而在不少的模型中被称作 Evidence 的关于观测样本的先验 $p(\mathbf{x})$ 难以求解，或需要非常高的计算复杂度才能得到结果。下面我们举例说明。
以下是一个高斯混合模型 (Mixture of Gaussians)。假设我们有一个单位方差的单变量的混合高斯模型，它有 K 个混合分量，分别对应 K 个高斯分布。这些高斯分布的方差均为 1，均值分别为 $\{\mu_1, \mu_2, ..., \mu_K\}$。假设这些均值变量的概率密度函数为 $p(\mu_k)$，且有 $\mu_k \sim \mathcal{N}(0, \delta^2)$，其中 δ 为超参数。假设 $\mathbf{c}_i (i \in [n])$ 是一个 K 维的指示向量，其仅在一个位置上值为 1，其余均为 0，我们可以用 $\mathbf{c}_i$ 指示样本 $\mathbf{x}_i$ 属于哪一个具体的高斯分布。

整个模型具体可描述为：

$$\begin{aligned}
&\mu_k \sim \mathcal{N}(0, \delta^2) \qquad k = 1, ..., K \\
&\mathbf{c}_i \sim \text{categorical}(1/K, 1/K, ..., 1/K) \qquad i = 1, ..., n \\
&\mathbf{x}_i | \mathbf{c}_i, \boldsymbol{\mu} \sim \mathcal{N}(\mathbf{c}_i^\top \boldsymbol{\mu}, 1) \qquad i = 1, ..., n
\end{aligned}$$

对于上述模型，隐变量和观测变量联合分布的概率密度为

$$p(\boldsymbol{\mu}, \mathbf{c}, \mathbf{x}) = p(\boldsymbol{\mu}) \prod_{i=1}^{n} p(\mathbf{c}_i) p(\mathbf{x}_i | \mathbf{c}_i, \boldsymbol{\mu})$$

其 Evidence 可以写为：

$$\begin{aligned}
p(\mathbf{x}) &= \int p(\boldsymbol{\mu}) \prod_{i=1}^{n} p(\mathbf{c}_i) p(\mathbf{x}_i | \mathbf{c}_i, \boldsymbol{\mu}) \, d\boldsymbol{\mu} \\
&= \sum_{\mathbf{c}} p(\mathbf{c}) \prod_{i=1}^{n} p(\mathbf{x}_i | \mathbf{c}_i, \boldsymbol{\mu}) \, d\boldsymbol{\mu} \\
&= A \sum_{\mathbf{c}} p(\mathbf{c}) \int_{\mu_k} \cdots \int \exp\left(-\frac{1}{2\delta^2} \sum_{k=1}^{K} \mu_k^2 - \frac{1}{2} \sum_{i=1}^{n} (\mathbf{x}_i - \sum_{k=1}^{K} c_{ik} \mu_k)^2 \right) d\mu_1 \cdots d\mu_K
\end{aligned}$$

其中 A 为常数。根据 $n \times K$ 维矩阵 $\mathbf{c}$ 的不同值，我们要计算 $\mathcal{O}(K^n)$ 个 K 重积分。

变分推断

上面的计算复杂度关于样本 n 呈指数增长，这是令人难以接受的。所以，基于贝叶斯推断的精确求解是无法得到结果的。所以我们便想通过某个只关于隐变量的分布 $q(\mathbf{z})$ 去近似 $p(\mathbf{z}|\mathbf{x})$，即最小化如下 KL 散度

$$q^*(\mathbf{z}) = \underset{q(\mathbf{z})\in\mathcal{D}}{\operatorname{argmin}} \mathsf{KL}(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x}))$$

其中 $\mathcal{D}$ 为预先定义的 $q(\mathbf{z})$ 的搜索空间。
由于直接求解 KL 散度比较困难，上式可以进一步化为：

$$\begin{aligned} q^*(\mathbf{z}) &= \underset{q(\mathbf{z})\in\mathcal{D}}{\operatorname{argmin}} \mathsf{KL}(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x})) \\ &= \underset{q(\mathbf{z})\in\mathcal{D}}{\operatorname{argmin}} \mathbb{E}[\log q(\mathbf{z})] - \mathbb{E}[\log p(\mathbf{z}|\mathbf{x})] \\ &= \underset{q(\mathbf{z})\in\mathcal{D}}{\operatorname{argmin}} \mathbb{E}[\log q(\mathbf{z})] - \mathbb{E}[\log p(\mathbf{z},\mathbf{x})] + \log p(\mathbf{x}) \end{aligned} \tag{1}$$

我们定义分布 $p(\mathbf{x})$ 的证据下界 (Evidence Lower Bound, ELBO) 为

$$\text{ELBO}(q) = \mathbb{E}[\log p(\mathbf{z}, \mathbf{x})] - \mathbb{E}[\log q(\mathbf{z})] \tag{2}$$

由于优化 $q(\mathbf{z})$ 时，$\log p(\mathbf{x})$ 为常量，所以最小化 $\text{KL}(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x}))$ 等价于最大化 $\text{ELBO}(q)$。这是我们最终要优化的目标。下面阐述证据下界的得名由来。另一方面，我们有：

$$\begin{aligned}
\log p(\mathbf{x}) &= \log \left(\int p(\mathbf{x}, \mathbf{z}) \, d\mathbf{z} \right) \\
&= \log \left(\int q(\mathbf{z}) \frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} \, d\mathbf{z} \right) \\
&= \log \left(\mathbb{E} \left[\frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} \right] \right) \\
&\geq \mathbb{E} \left[\log \left(\frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} \right) \right] \quad \text{(Jensen Inequality)} \\
&= \text{ELBO}(q)
\end{aligned}$$

上述等号等且仅当 KL 散度为 0 时成立。可以看出 $\text{ELBO}(q)$ 即为 $p(\mathbf{x})$ 的下界，故而得名。

优化算法 (平均场变分族)

不同的密度搜索空间 $\mathcal{D}$，有不同的优化方法。此处介绍在平均场变分族 (Mean-field Variational Family) 的假设下，介绍式(2)的优化问题。

平均场变分族

平均场变分族意即隐变量之间相互独立，且其密度函数由各自的变分因子决定。此时有

$$q(\mathbf{z}) = \prod_{j=1}^{m} q_j(\mathbf{z}_j) \tag{3}$$

其中 $q_j(\mathbf{z}_j)$ 为 $\mathbf{z}_j$ 的概率密度函数。

基于如上假设，将(3)带入(2)中，可以得到

$$\text{ELBO}(q) = \underbrace{\int \prod_{j=1}^{m} q_j(\mathbf{z}_j) \log p(\mathbf{z}, \mathbf{x}) \, d\mathbf{z}}_{\text{Part1}} - \underbrace{\int \prod_{j=1}^{m} q_j(\mathbf{z}_j) \sum_{j=1}^{m} \log q_j(\mathbf{z}_j) \, d\mathbf{z}}_{\text{Part2}} \tag{4}$$

我们采取坐标上升 (轮替优化) 的方法对式(4)进行优化，即只优化一个变量，固定其余变量。我们针对 $q_j(\mathbf{z}_j)$ 进行优化时，式(4)第一部分可化为：

$$\text{Part1} = \int_{\mathbf{z}_j} q_j(\mathbf{z}_j) \left(\mathop{\mathbb{E}}_{i \neq j} [\log p(\mathbf{z}, \mathbf{x})] \right) d\mathbf{z}_j$$

而第二部分可化为：

$$\begin{aligned} \text{Part2} &= \int \prod_{k=1}^{m} q_k(\mathbf{z}_k) \log q_j(\mathbf{z}_j) \, d\mathbf{z} + \sum_{i \neq j}^{m} \int \prod_{k=1}^{m} q_k(\mathbf{z}_k) \log q_i(\mathbf{z}_i) \, d\mathbf{z} \\ &= \int_{\mathbf{z}_j} q_j(\mathbf{z}_j) \log q_j(\mathbf{z}_j) \, d\mathbf{z}_j + \text{const} \end{aligned}$$

通过如上演算，可将式(4)关于 $q_j(\mathbf{z}_j)$ 的优化问题变为：

$$\begin{aligned} \text{ELBO}(q) &= \int_{\mathbf{z}_j} q_j(\mathbf{z}_j) \left(\mathop{\mathbb{E}}_{i \neq j} [\log p(\mathbf{z}, \mathbf{x})] \right) d\mathbf{z}_j - \int_{\mathbf{z}_j} q_j(\mathbf{z}_j) \log q_j(\mathbf{z}_j) \, d\mathbf{z}_j \\ &= -\text{KL} \left(q_j(\mathbf{z}_j) || \exp \left(\mathop{\mathbb{E}}_{i \neq j} [\log p(\mathbf{z}, \mathbf{x})] \right) \right) \end{aligned} \tag{5}$$

通过式(5)的推导，我们将最大化证据下界的问题转换为一个最小化 KL 散度的问题。根据 KL 散度的性质，式(5)关于 $q_j(\mathbf{z}_j)$ 有闭式解：

$$q_j^*(\mathbf{z}_j) = \exp\left(\mathop{\mathbb{E}}_{i\neq j}\left[\log p(\mathbf{z}, \mathbf{x})\right]\right) \tag{6}$$

于是，我们在对 $q_j(\mathbf{z}_j)$ 初始化以后，利用(6)采用坐标上升的方法使式(2)收敛，最终得到局部最优解。上述演算过程比较抽象，下面我们举个例子，就能使得整个过程变得一目了然。

举例说明

接下来我们用变分推断求解开头所述的单变量高斯混合模型。假设控制 μ_k 的参数为均值 m_k 和方差 s_k，控制 $\mathbf{c}_i$ 的参数为 $\boldsymbol{\varphi}_i$[1]。
据此我们的证据下界可以化为：

$$\begin{aligned}\mathrm{ELBO}(\mathbf{m},\mathbf{s},\boldsymbol{\varphi}) &= \mathbb{E}[\log p(\mathbf{x},\mathbf{c},\boldsymbol{\mu});\mathbf{m},\mathbf{s},\boldsymbol{\varphi}] - \mathbb{E}[\log q(\boldsymbol{\mu});\mathbf{m},\mathbf{s}^2] - \mathbb{E}[\log q(\mathbf{c});\boldsymbol{\varphi}] \\ &= \sum_{i=1}^{n}\left(\mathbb{E}[\log p(\mathbf{x}_i|\mathbf{c}_i,\boldsymbol{\mu});\mathbf{m},\mathbf{s}^2,\boldsymbol{\varphi}_i] + \mathbb{E}[\log p(\mathbf{c}_i);\boldsymbol{\varphi}_i]\right) + \sum_{k=1}^{K}\mathbb{E}[\log p(\mu_k);m_k,s_k^2] \\ &\quad - \sum_{i=1}^{n}\mathbb{E}[\log q(\mathbf{c}_i);\boldsymbol{\varphi}_i] - \sum_{k=1}^{K}\mathbb{E}[\log q(\mu_k);m_k,s_k^2]\end{aligned}$$

接下来，我们对变量进行逐个优化。

[1] $\boldsymbol{\varphi}_i$ 为一个 K 维的实值向量，它代表第 i 个样本关于所有高斯分量的隶属度，详情可参见求解高斯混合模型的 EM 算法。

首先是关于控制 $\mathbf{c}_i$ 的量 φ_i 的优化，利用式(6)，我们有

$$q^*(\mathbf{c}_i;\varphi_i) \propto \exp\{\mathbb{E}[\log p(\mathbf{c}_i)] + \mathbb{E}[\log p(\mathbf{x}_i|\mathbf{c}_i,\boldsymbol{\mu})]\} \tag{7}$$

根据 $\mathbf{c}_i$ 的定义，可知 $p(\mathbf{c}_i) = 1/K$。因此，$\mathbb{E}[\log p(\mathbf{c}_i)] = -\log K$。而上式第二项可以写作：

$$\begin{aligned}
\mathbb{E}[\log p(\mathbf{x}_i|\mathbf{c}_i,\boldsymbol{\mu})] &= \mathbb{E}\left[\log\left(\prod_{k=1}^{K} p(\mathbf{x}_i|\mu_k)^{c_{ik}}\right)\right] \\
&= \sum_{k=1}^{K} c_{ik}\,\mathbb{E}[\log p(\mathbf{x}_i|\mu_k)] \\
&= \sum_{k=1}^{K} c_{ik}\,\mathbb{E}[-(\mathbf{x}_i-\mu_k)^2/2] + \text{const} \\
&= \sum_{k=1}^{K} c_{ik}(\mathbf{x}_i m_k - (m_k^2+s_k^2)/2) + \text{const}
\end{aligned}$$

综上，去掉无关的项，我们有 $\varphi_{ik} \propto \exp\{\mathbf{x}_i m_k - (m_k^2+s_k^2)/2\}$。

接下来是对控制 μ_k 的量 m_k 和 s_k 的优化。将无关的项看作常数，利用式(6)，我们有：

$$q^*(\mu_k; m_k, s_k) \propto \exp\{\mathbb{E}[\log p(\mu_k)] + \sum_{i=1}^{n} \mathbb{E}[\log p(\mathbf{x}_i|\mathbf{c}_i, \boldsymbol{\mu})]\} \tag{8}$$

根据 $\mathbf{c}_i$ 的定义，我们有 $\mathbb{E}[c_{ik}] = \varphi_{ik}$。所以我们有：

$$\begin{aligned}
\log q^*(\mu_k; m_k, s_k) &= \log p(\mu_k) + \sum_i \mathbb{E}[\log p(\mathbf{x}_i|\mathbf{c}_i, \boldsymbol{\mu})] + const \\
&= -\mu_k^2/2\delta^2 + \sum_i \mathbb{E}[c_{ik} \log p(\mathbf{x}_i|\mu_k)] \\
&= -\mu_k^2/2\delta^2 + \sum_i \mathbb{E}[c_{ik}] \log p(\mathbf{x}_i|\mu_k) \\
&= -\mu_k^2/2\delta^2 + \sum_i \varphi_{ik}(-(\mathbf{x}_i - \mu_k)^2/2) \\
&= (\sum_i \varphi_{ik}\mathbf{x}_i)\mu_k - \frac{1}{2}(\sum_i \varphi_{ik} + 1/\delta^2)\mu_k^2
\end{aligned}$$

由于 $q^*(\mu_k; m_k, s_k)$ 是一个 Gaussian 分布的密度函数，我们对比上式的一次项和二次项，可以得到 m_k 和 s_k 的更新策略如下：

$$m_k = \frac{\sum_i \varphi_{ik} \mathbf{x}_i}{\sum_i \varphi_{ik} + 1/\delta^2}, \qquad s_k^2 = \frac{1}{\sum_i \varphi_{ik} + 1/\delta^2} \tag{9}$$

至此，我们已经得到了所有隐变量的更新策略。虽然推导过程比较复杂，但是更新策略确实是非常简单的，每一步更新过程关于样本数量 n 是呈线性的，相比于传统的贝叶斯推断的指数复杂度，要低得多。下一步，我将把上面这个例子用程序实现，并且提供优化过程中分布的可视化表示。

参考文献：

David M. Blei, Alp Kucukelbir & Jon D. McAuliffe (2017) Variational Inference: A Review for Statisticians, Journal of the American Statistical Association, 112:518, 859-877.