Variational Inference 01 Background

Chen Gong

30 November 2019

这一小节的主要目的是清楚我们为什么要使用 Variational Inference，表达一下 Inference 到底有什么用。机器学习，我们可以从频率角度和贝叶斯角度两个角度来看，其中频率角度可以被解释为优化问题，贝叶斯角度可以被解释为积分问题。

1 优化问题

为什么说频率派角度的分析是一个优化问题呢？我们从回归和 SVM 两个例子上进行分析。我们将数据集描述为：$D = \{(x_i, y_i)\}_{i=1}^N, x_i \in \mathbf{R}^p, y_i \in \mathbf{R}$。

1.1 回归

回归模型可以被我们定义为：$f(w) = w^T x$，其中 loss function 被定义为：$L(w) = \sum_{i=1}^N ||w^T x_i - y_i||^2$，优化可以表达为 $\hat{w} = argmin\ L(w)$。这是个无约束优化问题。

求解的方法可以分成两种，数值解和解析解。解析解的解法为：

$$\frac{\partial L(w)}{\partial w} = 0 \Rightarrow w^* = (X^T X)^{-1} X^T Y \tag{1}$$

其中，X 是一个 $n \times p$ 的矩阵。而数值解中，我们常用的是 GD 算法，也就是 Gradient Descent，或者 Stochastic Gradient descent (SGD)。

1.2 SVM (Classification)

SVM 的模型可以被我们表述为：$f(w) = sign(w^T + b)$。loss function 被我们定义为：

$$\begin{cases} \min\ \frac{1}{2} w^T w \\ s.t. \quad y_i(w^T x_i + b) \geq 1 \end{cases} \tag{2}$$

很显然这是一个有约束的 Convex 优化问题。常用的解决条件为，QP 方法和 Lagrange 对偶。

1.3 EM 算法

我们的优化目标为：

$$\hat{\theta} = argmax\ \ \log p(x|\theta) \tag{3}$$

优化的迭代算法为：

$$\theta^{(t+1)} = argmax_\theta \int_z \log p(X, Z|\theta) \cdot p(Z|X, \theta^{(t)}) dz \tag{4}$$

2 积分问题

从贝叶斯的角度来说，这就是一个积分问题，为什么呢？我们看看 Bayes 公式的表达：

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \tag{5}$$

其中，$p(\theta|x)$ 称为后验公式，$p(x|\theta)$ 称为似然函数，$p(\theta)$ 称为先验分布，并且 $p(x) = \int_\theta p(x|\theta)p(\theta)d\theta$。什么是推断呢？通俗的说就是求解后验分布 $p(\theta|x)$。而 $p(\theta|x)$ 的计算在高维空间的时候非常的复杂，我们通常不能直接精确的求得，这是就需要采用方法来求一个近似的解。而贝叶斯的方法往往需要我们解决一个贝叶斯决策的问题，也就是根据数据集 X(N 个样本)。我们用数学的语言来表述也就是，$\widetilde{X}$ 为新的样本，求 $p(\widetilde{X}|X)$：

$$\begin{aligned} p(\widetilde{X}|X) &= \int_\theta p(\widetilde{X}, \theta|X)d\theta \\ &= \int_\theta p(\widetilde{X}|\theta) \cdot p(\theta|X)d\theta \\ &= \mathbf{E}_{\theta|X}[p(\hat{x}|\theta)] \end{aligned} \tag{6}$$

其中 $p(\theta|X)$ 为一个后验分布，那么我们关注的重点问题就是求这个积分。

3 Inference

Inference 的方法可以被我们分为精确推断和近似推断，近似推断可以被我们分为确定性推断和随机近似。确定性推断包括 Variational Inference (VI)；随机近似包括 MCMC，MH，Gibbs Distribution 等。