配分函数

在学习和推断中，对于一个概率的归一化因子很难处理，这个归一化因子和配分函数相关。假设一个概率分布：

$$p(x|\theta)=\frac{1}{Z(\theta)}\hat{p}(x|\theta),Z(\theta)=\int\hat{p}(x|\theta)dx \tag{1}$$

包含配分函数的 MLE

在学习任务中，采用最大似然：

$$\begin{aligned}\hat{\theta}&=\underset{\theta}{argmax}\,p(x|\theta)=\underset{\theta}{argmax}\sum_{i=1}^N\log p(x_i|\theta)\\&=\underset{\theta}{argmax}\sum_{i=1}^N\log\hat{p}(x|\theta)-N\log Z(\theta)\\&=\underset{\theta}{argmax}\frac{1}{N}\sum_{i=1}^N\log\hat{p}(x|\theta)-\log Z(\theta)=\underset{\theta}{argmax}\,l(\theta)\end{aligned} \tag{2}$$

求导：

$$\begin{aligned}\nabla_\theta\log Z(\theta)&=\frac{1}{Z(\theta)}\nabla_\theta Z(\theta)\\&=\frac{p(x|\theta)}{\hat{p}(x|\theta)}\int\nabla_\theta\hat{p}(x|\theta)dx\\&=\int\frac{p(x|\theta)}{\hat{p}(x|\theta)}\nabla_\theta\hat{p}(x|\theta)dx\\&=\mathbb{E}_{p(x|\theta)}[\nabla_\theta\log\hat{p}(x|\theta)]\end{aligned} \tag{3}$$

由于这个表达式和未知的概率相关，于是无法直接精确求解，需要近似采样，如果没有这一项，那么可以采用梯度下降，但是存在配分函数就无法直接采用梯度下降了。

上面这个期望值，是对模型假设的概率分布，定义真实概率分布为 p_{data}，于是，$l(\theta)$ 中的第一项的梯度可以看成是从这个概率分布中采样出来的 N 个点求和平均，可以近似期望值。

$$\nabla_\theta l(\theta)=\mathbb{E}_{p_{data}}[\nabla_\theta\log\hat{p}(x|\theta)]-\mathbb{E}_{p(x|\theta)}[\nabla_\theta\log\hat{p}(x|\theta)] \tag{4}$$

于是，相当于真实分布和模型假设越接近越好。上面这个式子第一项叫做正相，第二项叫做负相。为了得到负相的值，需要采用各种采样方法，如 MCMC。

采样得到 $\hat{x}_{1-m}\sim p_{model}(x|\theta^t)$，那么：

$$\theta^{t+1}=\theta^t+\eta(\sum_{i=1}^m\nabla_\theta\log\hat{p}(x_i|\theta^t)-\sum_{i=1}^m\nabla_\theta\log\hat{p}(\hat{x_i}|\theta^t)) \tag{5}$$

这个算法也叫做基于 MCMC 采样的梯度上升。每次通过采样得到的样本叫做幻想粒子，如果这些幻想粒子区域的概率高于实际分布，那么最大化参数的结果就是降低这些部分的概率。

对比散度-CD Learning

上面对于负相的采样，最大的问题是，采样到达平稳分布的步骤数量是未知的。对比散度的方法，是对上述的采样是的初始值作出限制，直接采样 $\hat{x}_i = x_i$，这样可以缩短采样的混合时间。这个算法叫做 CD-k 算法，k 就是初始化后进行的演化时间，很多时候，即使 $k = 1$ 也是可以的。

我们看 MLE 的表达式：

$$\begin{aligned}\hat{\theta} &= \underset{\theta}{argmax}\, p(x|\theta) = \underset{\theta}{argmax}\, \frac{1}{N}\sum_{i=1}^{N}\log p(x_i|\theta) = \mathbb{E}_{p_{data}}[\log p_{model}(x|\theta)]\\ &= \underset{\theta}{argmax}\int p_{data}\log p_{model}dx\\ &= \underset{\theta}{argmax}\int p_{data}\log\frac{p_{model}}{p_{data}}dx\\ &= \underset{\theta}{argmin}\, KL(p_{data}||p_{model})\end{aligned} \tag{6}$$

对于 CD-k 的采样过程，可以将初始值这些点表示为：

$$p^0 = p_{data} \tag{7}$$

而我们的模型需要采样过程达到平稳分布：

$$p^{\infty} = p_{model} \tag{8}$$

因此，我们需要的是 $KL(p^0||p^{\infty})$。定义 CD：

$$KL(p^0||p^{\infty}) - KL(p^k||p^{\infty}) \tag{9}$$

这就是 CD-k 算法第 k 次采样的目标函数。

RBM 的学习问题

RBM 的参数为：

$$h = (h_1, \cdots, h_m)^T \tag{10}$$
$$v = (v_1, \cdots, v_n)^T \tag{11}$$
$$w = (w_{ij})_{mn} \tag{12}$$
$$\alpha = (\alpha_1, \cdots, \alpha_n)^T \tag{13}$$
$$\beta = (\beta_1, \cdots, \beta_m)^T \tag{14}$$

学习问题关注的概率分布为：

$$\begin{aligned}\log p(v) &= \log\sum_h p(h,v)\\ &= \log\sum_h\frac{1}{Z}\exp(-E(v,h))\\ &= \log\sum_h\exp(-E(v,h)) - \log\sum_{v,h}\exp(-E(h,v))\end{aligned} \tag{15}$$

对上面这个式子求导第一项：

$$\frac{\partial \log \sum\limits_h \exp(-E(v,h))}{\partial \theta} = -\frac{\sum\limits_h \exp(-E(v,h))\frac{\partial E(v,h)}{\partial \theta}}{\sum\limits_h \exp(-E(v,h))} \tag{16}$$

$$= -\sum_h \frac{\exp(-E(v,h))\frac{\partial E(v,h)}{\partial \theta}}{\sum\limits_h \exp(-E(v,h))} = -\sum_h p(h|v)\frac{\partial E(v,h)}{\partial \theta}$$

第二项：

$$\frac{\partial \log \sum\limits_{v,h} \exp(-E(h,v))}{\partial \theta} = -\sum_{h,v} \frac{\exp(-E(v,h))\frac{\partial E(v,h)}{\partial \theta}}{\sum\limits_{h,v} \exp(-E(v,h))} = -\sum_{v,h} p(v,h)\frac{\partial E(v,h)}{\partial \theta} \tag{17}$$

所以有：

$$\frac{\partial}{\partial \theta}\log p(v) = -\sum_h p(h|v)\frac{\partial E(v,h)}{\partial \theta} + \sum_{v,h} p(v,h)\frac{\partial E(v,h)}{\partial \theta} \tag{18}$$

将 RBM 的模型假设代入：

$$E(v,h) = -(h^T wv + \alpha^T v + \beta^T h) \tag{19}$$

1. w_{ij}：

$$\frac{\partial}{\partial w_{ij}}E(v,h) = -h_i v_j \tag{20}$$

于是：

$$\frac{\partial}{\partial \theta}\log p(v) = \sum_h p(h|v)h_i v_j - \sum_{h,v} p(h,v)h_i v_j \tag{21}$$

第一项：

$$\sum_{h_1,h_2,\cdots,h_m} p(h_1,h_2,\cdots,h_m|v)h_i v_j = \sum_{h_i} p(h_i|v)h_i v_j = p(h_i=1|v)v_j \tag{22}$$

这里假设了 h_i 是二元变量。

第二项：

$$\sum_{h,v} p(h,v)h_iv_j=\sum_{h,v}p(v)p(h|v)h_iv_j=\sum_v p(v)p(h_i=1|v)v_j \tag{23}$$

这个求和是指数阶的，于是需要采样解决，我么使用 CD-k 方法。

对于第一项，可以直接使用训练样本得到，第二项采用 CD-k 采样方法，首先使用样本 $v^0=v$，然后采样得到 h^0，然后采样得到 v^1，这样顺次进行，最终得到 v^k，对于每个样本都得到一个 v^k，最终采样得到 N 个 v^k，于是第二项就是：

$$p(h_i=1|v^k)v_j^k \tag{24}$$

具体的算法为：

1. 对每一个样本中的 v，进行采样：
    1. 使用这个样本初始化采样
    2. 进行 k 次采样（0-k-1）：
        1. $h_i^l\sim p(h_i|v^l)$
        2. $v_i^{l+1}\sim p(v_i|h^l)$
    3. 将这些采样出来的结果累加进梯度中
2. 重复进行上述过程，最终的梯度除以 N