Semi-Supervised Mean Fields

Fei Wang Shijun Wang Changshui Zhang Ole Winther
State Key Lab of Intelligent Technologies and Systems Intelligent Signal Processing
Department of Automation Informatics and Mathematical Modelling
Tsinghua University, Beijing, P.R.China Technical University of Denmark

【强化学习 104】Mean-field for Semi-supervised，也聊 Ising

张楚珩
清华大学 交叉信息院博士在读

本文提出使用 Ising 模型的平均场论解法来解决半监督学习问题。

原文传送门

Wang, Fei, et al. "Semi-supervised mean fields." Artificial Intelligence and Statistics. 2007.

特色

这篇文章先介绍了 Ising 模型（这是一个很有趣的模型，记得本科电磁学课的小作业还写过 Ising 模型的模拟程序）。接下来这篇文章把半监督学习问题用 Ising 模型来建模，并且使用平均场方法（naive mean field approach）来解决。文章还指出，贝叶斯方法和该方法在本质上的联系；特别地，半监督学习问题的本质就是要利用数据标签在数据空间上的平滑特性，可以把数据空间建模为一个图（graph），这又和我们前面在研究的 spectral graph theory 有一定的联系。

过程

1. Ising 模型

考虑 N 个粒子，每个原子有一个自旋（spin），自旋是二值量子化的，只能够取向上或者向下两种情况，即 $S_i \in \{-1,+1\}$ 。这样，原子的自旋方向的状态（configuration）就有 2^N 种情形，记这样的状态为加粗的 $\boldsymbol{s}$ 。

那么系统究竟处于哪一种状态下呢？考虑这是一个热力学系统，系统会以一定的概率处于某一个状态下，根据热力学的知识，一个粒子可分辨系统处于不同状态的概念遵循玻尔兹曼分辨，即系统处于状态 $\boldsymbol{s}$ 的概率为 $P(\boldsymbol{S}) = \frac{1}{Z}\exp(-E(\boldsymbol{S})/kT)$ ，其中 k 为玻尔兹曼常数，T 为系统当前的温度，E 为系统处于该状态时的能量，Z 为配分函数，也可以理解它为归一化系数；在这里可以认为 kT=1。可以看到，当温度较低时，系统处于最低能量状态的概念会大大超过其他状态，因此系统会基本上处于最低的能量状态；当温度较高时，系统处于各个能量状态的概率基本上一样，因此系统会以几乎差不多的概率分布在各个状态上。

某个状态下的能量函数定义如下

$$E(\mathbf{S}) = -\sum_{<i,j>} J_{ij} S_i S_j - \sum_i \theta_i S_i, \qquad (4)$$

其中求和符号表示对于系统中任意两个粒子对进行求和，J 表示这两个粒子对的相互作用能，θ 表示外场对于某个粒子的作用能。

有了这样的模型之后，我们的问题是：如果给定相互作用能 J 和外场作用能 θ，如何求到每一个粒子所处自旋状态的概率分布或者均值。我们可以把第 i 个粒子处于状态 S_i 的概率分布写作整个系统状态分布的边缘分布：

$$P(S_i) = \int \prod_{j \neq i} dS_j P(\mathbf{S}). \qquad (7)$$

即，对于所有在第 i 个粒子上处于状态 S_i 的状态的概率的和。

接着，我们还可以写出该粒子自旋状态的均值：

$$\langle S_i \rangle = \int dS_i P(S_i), \qquad (8)$$

我们可以看到，这里计算的难点就在于要计算几乎所有可能的状态的概率，并对其求和。然而，系统的状态数目随着粒子的数目呈指数级增长，因此这样直接计算的方法是不可行的。

2. 平均场方法求解 Ising 模型

其实，如果系统处于某个状态的概率（联合概率分布）如果具有某些好的性质，比如能拆分成多个独立的项，其实我们也能够比较方便地进行计算。观察系统处于某个状态的概率：

$$P(S) = \frac{1}{Z}\exp\left(\sum_i (\theta_i + \sum_j J_{ij} S_j) S_i\right) \propto \prod_i \exp\left((\theta_i + \sum_j J_{ij} S_j) S_i\right)$$

但是注意到，它并没有被拆分为多个独立的项，因为每一项都还是与全局其他粒子的自旋相关。接下来，就到了最为关键的一步了，就是把外界对于每一个粒子的作用都用一个**平均场** $\bar{m}_i$ 来代替。把这样近似后的概率分布函数写作 Q。令

$$Q(S) = \prod_i Q_i(S_i), \quad Q_i(S_i) = \frac{\exp(\bar{m}_i S_i)}{\sum_{S_i} \exp(\bar{m}_i S_i)} \approx \frac{1 + S_i m_i}{2}$$

考虑到每个粒子只有两种状态，而这两种状态对应的 Q_i 之和为 1，因此也可以被写作后面一种形式；同时，注意到 m_i 也可以看做是分布 Q 的一个参数；如果是后一种情形，m_i 也可以被解释为一个隐变量，代表着粒子 i 处的平均自旋，$m_i = \langle S_i \rangle_Q$ 。

下面我们要做的就是求解出近似的概率分布 Q，也就是要求解出相应的平均场。我们最小化分布 P 和分布 Q 之间的 KL divergence

$$KL(Q\|P) = \sum_{\mathbf{S}} Q(\mathbf{S}) \ln \frac{Q}{P}. \qquad (9)$$

$$KL(Q\|P) = \ln Z + V(Q) - S(Q), \qquad (10)$$

其中 S 为分布 Q 的熵

$$S(Q) = -\sum_{\mathbf{S}} Q(\mathbf{S}) \ln Q(\mathbf{S}) \qquad (11)$$

V 为 variational entropy

$$V(Q) = \sum_{\mathbf{S}} Q(\mathbf{S}) E(\mathbf{S}) \tag{12}$$

带入到 KL divergence 中，然后求导，可以求解出每个粒子的平均自旋 m_i

$$m_i = \tanh\left(\sum_j J_{ij} m_j + \theta_i\right). \tag{18}$$

注意到每个位置的平均自旋的解是通过其他位置的解来定义的，因此要准确求出还需要做数值迭代，即通过下面的算法来得到。

Table 1: The NMF Method For Ising Model.

Initialization:
- Start from a *tabular rasa* $\langle \mathbf{S} \rangle = (\langle S_1 \rangle, \langle S_2 \rangle, \cdots, \langle S_N \rangle) = \mathbf{0}$ (or small values if $\langle \mathbf{S} \rangle = \mathbf{0}$ is a fixed point).
- Learning rate $\eta = 0.05$.
- Fault tolerance $ft = 10^{-3}$.

Iterate:

do:

for all i**:**

Compute m_i by Eq.(18).

$\delta \langle S_i \rangle = m_i - \langle S_i \rangle$

end for

for all i**:**

$\langle S_i \rangle = \langle S_i \rangle + \eta \delta \langle S_i \rangle$

end for

while $\max_i |\delta \langle S_i \rangle|^2 > ft$

平均场近似究竟在什么地方做了近似？

首先，平均场近似的结果肯定不是准确的结果。在平均场中，原本的“粒子-粒子”相互作用，通过一个平均场来作为中介，变成了“粒子-平均场-粒子”的相互作用。观察到，原本 P 函数中存在交叉项 $S_i S_j$ 。假设在热平衡状态下，每个粒子的平均状态为 $\langle S_i \rangle$ ，定义粒子偏离平均状态的量为 $\delta S_i = S_i - \langle S_i \rangle$ 。在平均场中，其实做了如下的近似

$$S_i S_j = \langle S_i \rangle \langle S_j \rangle + \delta S_i \langle S_j \rangle + \langle S_i \rangle \delta S_j + \delta S_i \, \delta S_j \approx \langle S_i \rangle \langle S_j \rangle + \delta S_i \langle S_j \rangle + \langle S_i \rangle \delta S_j$$

$\Rightarrow 2S_iS_j \approx S_i\langle S_j\rangle + S_j\langle S_i\rangle$

这样，原本的“粒子 i-粒子 j”变成了“粒子 i-平均场-粒子 j”。最后加上外面的求和，就可以得到最后的平均场近似。文章里面直接用了一个 tractable 的 Q 分布，然后最小化它和真实概率分布的 KL 散度，这种方法叙述会更简单，但是容易使人忽略其物理实质，同时也让人对于“平均场”这个词不太理解。建议大家参看更为严格的物理推导：

Franz Utermohlen: Notes on "Mean Field Theory Solution of the Ising Model"

接下来，我们来看看究竟在哪里做了近似，其实最直接的就是这里假设了 [illegible] ；在热扰动（thermal fluctuation）较小的时候，这个式子成立。物理上来理解，在维度较低的时候，热扰动可能会比较明显，因此平均场可能不太准确；在维度较高的时候，会更准确。

3. Ising 模型的临界温度

既然我们得到了这个解

$$m_i = \tanh\left(\sum_j J_{ij} m_j + \theta_i\right).$$

那么下面可以顺水推舟，顺便讲一下系统的临界温度问题。现在考虑这 N 个粒子构成了一块固体，每个粒子的自旋就是其磁矩，那么这个物体就会因此产生一个宏观的磁矩 m。这个磁矩以及外场的作用在该物体的任何位置应该都是一样的，因此我们可以抹去上述公式的下标。同时，注意到，我们前面把温度等一些物理相关的常数去掉了，现在我们把它们捡回来，可以得到

$m = \tanh(\frac{\theta + Jm}{kT}),\ J = \sum_j J_{ij} \quad J_{ij} > 0$ ，

先考虑没有外场的情形（ $\theta = 0$ ）。我们发现当温度较高的时候，即 $J/kT < 1$ 的时候，上述方程只有一个解 m=0；我们把这种状态下的材料称为顺磁性材料（paramagnetic），这种材料在出现外场的时候，会顺着外磁场产生一定的磁矩，相应的比例就是磁化率。当温度较低的时候，即 $J/kT > 1$ 的时候，上述方程只有两个稳定解 $m = \pm m_0$ ，则意味着概材料在没有外磁场的情况下也具有一定的磁性，我们称这种材料为铁磁性材料（ferromagnetic）。而相应的临界温度 $T = J/k$ 就成为系统的临界温度。

4. MCMC 方法求解 Ising 模型

关于 Ising 模型，我们前面研究的问题是：对于任意一个粒子，其自旋状态的边缘分布怎样。为了解决这样一个问题，我们做了平均场近似，即把其他粒子对某一粒子的作用等效为一个平均场对于该粒子的作用，从而实现了联合概率分布的解耦合，使得我们可以顺利求解这个系统。该方法不仅可以求某个粒子状态的边缘分布、求解系统的宏观状态，还可以从分布中采样。

我们现在只考虑采样问题：当 J_{ij}, θ_i 都确定下来之后，其联合概率分布函数 P 就确定下来了，如何从这个分布中采集一个系统状态的样本？这时候我们还可以利用 Markov Chain Monte Carlo（MCMC）方法来进行求解。

考虑系统的每一个状态（configuration）也是马科夫链上的一个状态（state），它们一一对应。如果马科夫链满足遍历性（ergodicity）和细致平稳（detailed balance），那么该马科夫链就具有唯一的稳态分布；即从该马科夫链上的任意一个状态出发，经过无限步之后所处状态的分布就是该唯一状态分布。从联合概率分布函数 P 中采样的问题，可以转化为设计一个马科夫链的概率转移矩阵，使得它：1）满足遍历性和细致平稳；2）其稳态分布是 P。

考虑这样一个马科夫链：每一个状态描述 N 粒子系统的一个状态 $s \in \{-1, +1\}^N$ ，而每次只考虑翻转一个粒子，即当且仅当 s 和 s' 只有至多一个粒子上的自旋不一样时， s 可以一步转移到 s' 。

遍历性：我们注意到，只要我们能够保证从任意一个状态出发，它转移到任意和它只有一个粒子不同的状态的概率不为零，就能够保证整个系统的遍历性；即从任意状态出发都一定能以非零的概率以最多 N 步转移到任意一个其他状态。

$$P_{s\to s'} > 0,\ \forall d(s',s) \le 1$$

其中，P 表示一步概率转移矩阵，d 两个状态之间的距离，它表示两个状态之间有多少个粒子自旋不同。

细致平稳：细致平稳讲的是达到稳态分布之后，从任意状态流向另一任意状态的概率流等于反向的概率流，假设稳态分布在状态 s 上的概率为 d_s ，那么细致平稳条件可以写作

$$d_s P_{s\to s'} = d_{s'} P_{s'\to s}$$

稳态分布：我们希望设计的这个马科夫链的稳态分布是 P（不好意思，符号有点乱，稳态分布的 P 我们用函数形式表示，一步转移矩阵 P 我们用下标形式表示，大家区分一下），那么该条件可以写作

$$\frac{d_s}{d_{s'}} = \frac{P(s)}{P(s')} = \exp(\Delta E_{s',s}),\ \Delta E_{s',s} := E(s') - E(s)$$

我们同时注意到，由于 s 和 s' 只相差了一个粒子，因此，其能量差距可以较为容易得计算得到。

收敛速率：虽然只要满足以上条件，就可以保证从马科夫链上任意状态出发经过无穷多步之后能够收敛到稳态分布，但是我们还是希望这个收敛的过程需要快一些，因此我们希望各个状态之间的转移概率足够大。不妨假设 s' 的能量大于等于 s 的能量，联立上述方程，有

$$\frac{P_{s'\to s}}{P_{s\to s'}} = \exp(\Delta E_{s',s}) \ge 1$$

要使得收敛速率最大，可以令 $P_{s'\to s} = 1$ ， $P_{s\to s'} = \exp(-\Delta E_{s',s})$ 。

转移方案：把上述条件全部考虑进来，我们可以得到如下的概率转移方案：

在状态 s 上，以某一个概率分布从 N 个粒子中采样一个粒子，该概率分布需要以非零的概率采集任意一个粒子。考虑翻转该粒子形成状态 s'，并且计算 $\Delta E_{s',s}$ 。如果 $\Delta E_{s',s} \le 0$ 就以 100% 的概率转移到状态 s'；如果 $\Delta E_{s',s} > 0$ ，就以 $\exp(-\Delta E_{s',s})$ 的概率转移到状态 s'。

比较：下面我们来比较一下 MCMC 方法和平均场方法。

- 平均场方法没有考虑热扰动，做了近似，但是 MCMC 方法没有该近似。
- 平均场方法可以求解出系统和各个热力学统计量之间的关系，便于我们理解系统的宏观规律，而 MCMC 方法基于数值模拟，无法让我们得到更多对于物理系统的观察。
- 相比平均场方法而言，MCMC 方法在计算上比较慢，并且由于 MCMC 方法基于马科夫链，因此其本质上是一个 sequential 的方法，比较难并行化或者向量化加速。不过值得一提的是，我们刚刚讲的这种每次考虑翻转一个粒子的方法叫做 Matropolis-Hastings 算法，后续还发展了又每次翻转一坨粒子（cluster）的 Swendsen–Wang 算法和 Wolff 算法，它们可以一定程度上加速 MCMC。

注：以上比较来源于一篇比较老的文献 [1]，因此不确定还有没有更新的研究进展。关于更详细的介绍以及 Wolff 算法可以参考这一系列知乎 [2]。

5. 用 Ising 模型建模半监督学习问题

绕了一圈再回到文章中的这个问题，即如何用 Ising 模型来建模半监督学习问题。先考虑一个半监督学习问题，它有一个数据集 x ；其中一部分数据有 -1 或者 +1 的标签，记这一部分数据为 x_L ；另一部分数据没有标签，记这一部分数据为 x_U 。在这种情况下，我们可以把 N 个数据点看做是 N 个粒子，而这些数据点的标签看做是 N 个粒子的自旋状态；用粒子间相互作用 J_{ij} 来刻画不同数据点之间的相似程度；用外场作用 θ_i 来刻画有标签数据点的标签。通过求解热平衡状态下的系统，根据粒子的期望自旋状态 $\langle S_i \rangle$ 来推断无标签数据点的标签。

粒子-粒子相互作用 J_{ij}

文章先定义了数据点之间的距离度量，然后再根据距离度量来定义相似性，并且把相似性作为 J_{ij} ；可以发现，相似性越高，说明这两个数据点之间的标签越可能相似，因此赋予其一个较大的相互作用常数。注意到距离度量和相似性度量都需要根据一定的先验知识来选取适用于相应数据的度量，文章里面比较了很多种，这里就介绍一下文章里面提到的比较特别的（也是最后实验使用到

的）度量。

首先，相似性可以根据通过距离来度量，度量方式为 weighted exponential similarity

$$w_{ij} = \exp\left(-\frac{d_{ij}}{\sigma}\right), \qquad (19)$$

接下来，定义距离度量为 connectivity distance

$$d_{ij} = \min_{p \in \mathcal{P}_{ij}} \max_{1 \leqslant k \leqslant |p|-1} d^{E}_{p_k p_{k+1}},$$

其想法是考虑有一条路径连接第 i 个数据点和第 j 个数据点，那么这条路径上面两两数据点之间的最大距离可以刻画这两个数据点之间的是不是被其他的数据点较好地“联通”；同时，在所有的路径之间找一条路径，使得这两个数据点能够被最好的联通。注意到，在这里，这个距离度量的定义仍然是基于数据点之间的欧氏距离来定义的（上标 E）。

该距离度量有一个 soften 的版本，定义如下

$$d_{ij} = \frac{1}{\rho} \ln\left(1 + \min_{p \in \mathcal{P}_{ij}} \sum_{k=1}^{|p|-1} \left(e^{\rho d^{E}_{p_k p_{k+1}}} - 1\right)\right),$$

其中有一个参数 ρ；它趋向于零的时候，上述度量更接近于度量路径的总长度；它趋向于无穷大的时候上述度量更接近于度量路径中的最大一段的长度（即前面没有 softened 的版本）。

外场作用 θ_i

如果有标签，外场作用就设置为标签；如果没有标签就设置外场作用为零。

$$\theta_i = \begin{cases} l_i, & if\ \mathbf{x}_i\ is\ labeled\ as\ l_i \\ 0, & if\ \mathbf{x}_i\ is\ unlabeled \end{cases}. \qquad (22)$$

算法

最后的算法就是先计算相互作用和外场，然后利用前面的平均场方法来求解到每个粒子的平均自旋，根据平均自旋的符号来给出标签的推断。只需要注意一点就是，相似性矩阵计算出来之后，还又经过了一个归一化处理。

Inputs:
- Dataset $\mathcal{X}$, Scale σ, Temperature T
- A proper distance function $d(\cdot,\cdot)$

Outputs:
- The labels of the unlabeled data.

1. Calculate the distance matrix $\mathbf{D}$ with its (i,j)-th entry $D_{ij} = d(\mathbf{x}_i, \mathbf{x}_j)$;
2. Calculate the interaction matrix $\tilde{\mathbf{J}}$ with its (i,j)-th entry $\tilde{J}_{ij} = \exp(-D_{ij}/\sigma)$;
3. Normalize $\tilde{\mathbf{J}}$ by $\tilde{\mathbf{J}} = \mathbf{R}^{-1/2}\tilde{\mathbf{J}}\mathbf{R}^{-1/2}$, where $\mathbf{R}$ is a diagonal matrix with the i-th element on its diagonal line $R_{ii} = \sum_j \tilde{J}_{ij}$;
4. Compute the matrix $\mathbf{J}$ with its (i,j)-th entry $J_{ij} = \beta\tilde{J}_{ij}$, where β is defined as in Eq.(3),
5. Calculate $\langle S_u \rangle$ $(\mathbf{x}_u \in \mathcal{X}_U)$ by the method shown in Table 1, and fix $\langle S_l \rangle$ $(\mathbf{x}_l \in \mathcal{X}_l)$ to be t_l, which is the label of $\mathbf{x}_l$;
6. If $\langle S_i \rangle > 0$, then classify $\mathbf{x}_i$ as $+1$, else if $\langle S_i \rangle < 0$, classify $\mathbf{x}_i$ as -1

6. 平均场方法和贝叶斯方法的联系

考虑如下概率图模型，即从数据 X 生成一个数据的类别隐变量 y，然后从该隐变量生成数据的标签 t。

那么从半监督学习数据集中做推断，可以写作

$$P(\mathbf{y}|\mathcal{D}) = P(\mathbf{y}|\mathcal{X}, \mathbf{t}_L) = \frac{1}{Z}P(\mathbf{y}|\mathcal{X})P(\mathbf{t}_L|\mathbf{y})$$

其中，花 D 表示半监督学习数据集；后一个等式是贝叶斯公式，只不过把分母有归一化常数代替；最后的推断由两部分构成，前一项表示数据集上的先验信息，后一项表示产生已有标签的 likelihood。

Prior Information term

为了计算得到前一项先验信息，我们可以把数据看做一个全连接的图：图上的每一个节点就是一个数据点，图之间的连边强度就是数据点之间的相似性。做半监督学习的基本假设就是数据和标签之间需要平滑，在图上，我们可以定义一个数据的平滑性指标（smoothness）

$$\mathcal{S}_{\mathbf{y}} = \mathbf{y}^T\mathbf{M}\mathbf{y}, \qquad (24)$$

其中，向量 y 为数据的标签，每一个维度分别对应一个数据；M 为 smoothness matrix，根据我另外一个专栏你们讲的 spectral graph theory，可以自然想到，使用 normalized graph Laplacian 来作为 M 矩阵

$$\mathbf{M} = \mathbf{I} - \mathbf{R}^{-1/2}\tilde{\mathbf{J}}\mathbf{R}^{-1/2} \qquad (26)$$

因此，可以规定 prior information term 为

$$P(\mathbf{y}|\mathcal{X}) = \frac{1}{Z_p}\exp\left(-\mathbf{y}^T(\mathbf{I}-\mathbf{R}^{-1/2}\tilde{\mathbf{J}}\mathbf{R}^{-1/2})\mathbf{y}\right).$$

Likelihood term

考虑到数据标签只有两类，并且假设有标签的数据相互独立，因此这一项可以被定义为

$$P(\mathbf{t}_L|\mathbf{y}) = \prod_{i=1}^{l} P(t_i|y_i) = \frac{\exp\left(\sum_{i=1}^{l} t_i y_i\right)}{\prod_{i=1}^{l}(\exp(y_i)+\exp(-y_i))}$$

最后综合起来，可以得到

$$P(\mathbf{y}|\mathcal{D}) = \frac{\frac{1}{Z}e^{-\mathbf{y}^T\mathbf{y}+\mathbf{y}^T\mathbf{R}^{-1/2}\tilde{\mathbf{J}}\mathbf{R}^{-1/2}\mathbf{y}+\sum_{i=1}^{l}\tilde{\theta}_i y_i}}{\prod_{i=1}^{l}(e^{y_i}+e^{-y_i})}.$$

其中 θ 就是数据的标签。考虑到可以对 y 做标准化，因此半监督学习问题就变成了最大化如下函数

$$\mathcal{J} = \frac{\frac{1}{Z}e^{\mathbf{y}^T\mathbf{R}^{-1/2}\tilde{\mathbf{J}}\mathbf{R}^{-1/2}\mathbf{y}+\sum_{i=1}^{l}\tilde{\theta}_i y_i}}{\prod_{i=1}^{l}(e^{y_i}+e^{-y_i})},$$

和前面的 Ising 模型的对比可以看出，这里需要在最小化能量的同时也最大化隐变量 y 的分布；而 Ising 模型是在“软化地”最小化能量（服从 Boltzmann 分布）。

文章中还说到，用平均场解这个问题复杂度会比直接优化该目标更低。

One more thing

有一点需要注意的是文章中的 Ising 模型讲的是任意两个粒子/数据点间都有相互作用；但是在实际的系统中，距离越远的粒子，相互作用越弱，在很多文献中经常考虑的只是最近邻粒子之间的相互作用。这样的简化并不会改变问题的性质。

这篇文章发表在 2007 的 AISTATS 上，有若干公式打印有误。这里在截图之后已经做了更正。

参考资料

[1] Fiig, Thomas, et al. "Mean-field and Monte Carlo calculations of the three-dimensional structure factor for YBa 2 Cu 3 O 6+ x."*Physical ReviewB*54.1 (1996): 556.

[2] hlpyatne：当蒙特卡罗方法遇见伊辛模型（下）

编辑于 2020-02-18

量子场论 理论物理 强化学习 (Reinforcement Learning)

文章被以下专栏收录

强化学习前沿
读呀读paper