人工智能导论第三次作业

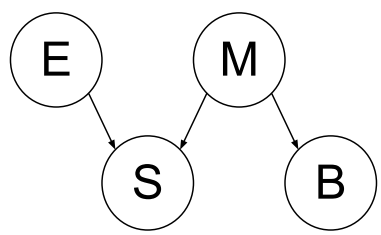
2023年5月

1 贝叶斯网络（10分）

考虑以下含有四个变量E, S, M, B的贝叶斯网络。相应的条件概率表如下。

$P(E)$	
$+e$	0.4
$-e$	0.6

$P(S\|E,M)$			
$+e$	$+m$	$+s$	1.0
$+e$	$+m$	$-s$	0.0
$+e$	$-m$	$+s$	0.8
$+e$	$-m$	$-s$	0.2
$-e$	$+m$	$+s$	0.3
$-e$	$+m$	$-s$	0.7
$-e$	$-m$	$+s$	0.1
$-e$	$-m$	$-s$	0.9



$P(M)$	
$+m$	0.1
$-m$	0.9

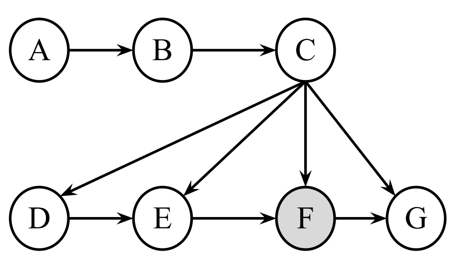
$P(B\|M)$		
$+m$	$+b$	1.0
$+m$	$-b$	0.0
$-m$	$+b$	0.1
$-m$	$-b$	0.9

计算以下概率:

(1) $P(-e, -s - m, -b)$
(2) $P(+b)$
(3) $P(+m \mid +b)$
(4) $P(+m \mid +s, +b, +e)$
(5) $P(+e \mid +m)$

2 变量消除法（10分）

考虑以下所有变量均为二元变量的贝叶斯网络。



我们想要使用变量消除法计算$P(B, D \mid F=f)$，变量消除的顺序为A, C, E, G。初始的所有因子如下：

$$P(a), P(b \mid a), P(c \mid b), P(d \mid c), P(e \mid c, d), P(f \mid c, e), P(g \mid c, f)$$

当消除A时，我们产生一个新的因子$\tau_1(b)$，如下所示

$$\tau_1(b)=\sum_a P(b \mid a) P(a)$$

剩余因子如下：

$$\tau_1(b), P(c \mid b), P(d \mid c), P(e \mid c, d), P(f \mid c, e), P(g \mid c, f)$$

请继续完成以下步骤：

(1) 当消除C时，产生新的因子τ_2，请写出其表达式并列出所有剩余因子。

(2) 当消除E时，产生新的因子τ_3，请写出其表达式并列出所有剩余因子。

(3) 当消除G时，产生新的因子τ_4，请写出其表达式并列出所有剩余因子。

(4) 如何利用剩余因子计算$P(B=b, D=d \mid F=f)$?

(5) 因子大小是变量消除法计算复杂度的关键因素。例如，假设所有的变量都是二元变量，则因子$P(b \mid a)$的大小是2，它有2^2种取值需要维护；因子$P(e \mid c, d)$的大小是3，它有2^3种取值需要维护。而由于f的值已观测，因此$P(g \mid c, f)$的大小只有2。你可能会发现，按照A, C, E, G的顺序并不是一个很好的顺序。请找出一个变量消除顺序，使得最大的因子最小。列出这个顺序，并给出在新的顺序下每次产生的因子大小。（提示：产生的最大新因子大小是2）

3 带方差的高斯线性回归（10分）

在高斯线性回归中，我们将σ也视作模型参数的一部分。假设$y_n \mid \mathbf{w}, \mathbf{x}_n, \sigma \sim \mathcal{N}(\mathbf{w}^T \mathbf{x}_n, \sigma^2)$，即

$$p(y_n \mid \mathbf{w}, \mathbf{x}_n, \sigma)=\frac{1}{\sqrt{2 \pi} \sigma} \exp \left(-\frac{1}{2 \sigma^2}\left(y-\mathbf{w}^T \mathbf{x}_n\right)^2\right),$$

其中$(\mathbf{x}_n, y_n)$为第n条数据，$\mathbf{w}$和σ为模型参数。假设数据集采样自独立同分布，并使用最大似然估计来求解，即

$$\log p(\mathcal{D}_n ; \mathbf{w}, \sigma)=\sum_{i=1}^N \log p(y_n \mid \mathbf{x}_n, \mathbf{w}, \sigma),$$

$$\hat{\mathbf{w}}, \hat{\sigma}=\underset{\mathbf{w}, \sigma}{\operatorname{argmax}} \log p(\mathcal{D}_n ; \mathbf{w}, \sigma).$$

求$\hat{\mathbf{w}}$和$\hat{\sigma}$.

4 采样（10分）

考虑一个在100×100网格上的采样问题。对于样本$\mathbf{x} = (x_1, x_2) \in \{1, 2, ..., 100\}^2$，其概率满足

$$p(x_1, x_2) \propto x_1 + \ln(x_1 x_2 + 2x_1 + 3x_2).$$

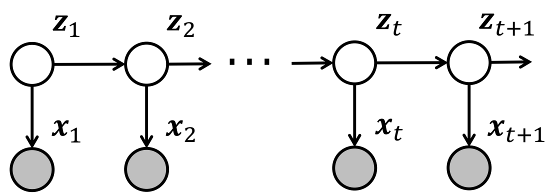
由于样本空间非常庞大，且归一化系数难以求解，我们使用马尔可夫链蒙特卡洛（Markov Chain Monte Carlo）进行采样。假设初始位于$(2, 2)$.

(1) 使用Metropolis-Hastings算法进行采样，所选的提议分布（proposal distribution）Q为均匀分布。若按照Q进行采样得到$(1, 3)$，则接受率$\alpha_{(2,2),(1,3)}$应该是多少？若采样得到的是$(3, 4)$，则接受率$\alpha_{(2,2),(3,4)}$又应该是多少？

(2) 使用Gibbs采样法进行采样。第一次采样过程需要采样x_1，请写出提议分布$q(x_1)$的形式（不需要计算）。在此基础上，若采样到$x_1 = 42$，请写出第二次采样过程需要采样的变量及其提议分布的形式（不需要计算）。

5 Baum-Welch算法（20分）

Baum-Welch算法是EM算法的一种，其解决了隐马尔科夫模型（Hidden Markov Model，HMM）三大主要问题中的学习问题。HMM的学习问题可以按如下方式定义：给定观测序列$X = \{x_1, ..., x_T\}$，在隐藏序列$Z = \{z_1, ..., z_T\}$ 未知的情况下，如何估计模型的最佳参数θ，使得$P(X \mid \theta)$最大。



参数$\theta = \{\pi, A, B\}$，包括初始概率分布$\pi = [\pi_i]_N$，转移（Transition）矩阵$A = [a_{ij}]_{N\times N}$，观测/发射（Emission）矩阵$B = [b_j(k)]_{N\times M}$，其中N表示隐状态总数，M表示可观测状态总数。在本题中我们将利用先前的知识完成该算法的推导。

(1) 首先进行**E**步的计算，在这一步中我们根据当前的网络参数θ^{old}得到隐变量Z的后验概率$q(Z) = P(Z, X \mid \theta^{(i)})$，然后固定住$q(Z)$，并列出ELBO得到优化目标。请根据ELBO写出$J(\theta)$，并证明：

$$\operatorname*{argmax}_\theta J(\theta) = \operatorname*{argmax}_\theta \sum_Z P(X, Z \mid \theta^{\text{old}}) \log P(X, Z \mid \theta)$$

(2) 使用模型的参数θ（包括π, A, B）来表示$P(X, Z \mid \theta)$，请给出相应的形式。

(3) 令$Q(\theta, \theta^{\text{old}}) = \sum_Z P(X, Z \mid \theta^{\text{old}}) \log P(X, Z \mid \theta)$，应用(2)中的结果，将$Q(\theta, \theta^{\text{old}})$拆分成三项之和，其中每一项仅与$\pi, A, B$中的一个参数有关。

提示：包含π的项为

$$\sum_Z P(X, Z \mid \theta^{\text{old}}) \log \pi_{z_1}.$$

(4) 上一小题实现了参数之间的解耦，可以进入**M**步的计算，此时暂时不需要考虑θ^{old}的处理。我们以π的求解为例。请首先证明

$$\sum_{Z} P(X, Z \mid \theta^{\text{old}}) \log \pi_{z_1} = \sum_{i=1}^{N} P(X, z_1 = i \mid \theta^{\text{old}}) \log \pi_i,$$

然后使用拉格朗日乘子法，求$\sum_Z P(X, Z \mid \theta^{\text{old}}) \log \pi_{z_1}$在

$$\sum_{i=1}^{N} \pi_i = 1$$

下取到极大值时π_i的取值。

(5) 参照上一小题的过程，求$Q(\theta, \theta^{\text{old}})$取极大值时$a_{ij}$和$b_j(k)$的取值。

(6) 注意到$\pi_i, a_{ij}, b_j(k)$的取值中都包含有关θ^{old}的项，现在我们对其进行求解。定义

$$\gamma_t^{\text{old}}(i) = P(z_t = i \mid X, \theta^{\text{old}})$$

$$\xi_t^{\text{old}}(i, j) = P(z_t = i, z_{t+1} = j \mid X, \theta^{\text{old}})$$

回顾课上讲过的前向-后向算法，对于给定参数θ^{old}的HMM，我们可以求出

$$\alpha_t^{\text{old}}(i) = P(x_{1:t}, z_t = i \mid \theta^{\text{old}}),\ \beta_t^{\text{old}}(i) = P(x_{t+1:T} \mid z_t = i, \theta^{\text{old}}),$$

请使用$\alpha_t^{\text{old}}(i), \beta_t^{\text{old}}(i), a_{ij}^{\text{old}}, b_j^{\text{old}}(k)$表示出$\gamma_t^{\text{old}}(i)$和$\xi_t^{\text{old}}(i, j)$.

(7) 回顾(4)(5)的结果，请使用$\gamma_t^{\text{old}}(i)$和$\xi_t^{\text{old}}(i, j)$表示出$\pi_i, a_{ij}, b_j(k)$，并给出Baum-Welch算法的伪代码。

6 LDA（40分）

请使用python实现Variational EM LDA。本次作业在*./dataset*中提供了三种不同的数据集，*dataset.txt*是英文的小规模数据集，*dataset_cn.txt*是中文的中等规模数据集，*dataset_cn_full.txt*是中文的大规模数据集。建议在较小数据集上验证实现正确性之后再使用较大的数据集。以下是作业要求：

(a) 根据提供的代码框架，写出Variational EM LDA的伪代码。

(b) 完成代码框架中缺失的变分推断部分。代码框架中已经实现了对于α, β的更新，只需要补充*main.py*的两个函数，计算ELBO并更新γ, ϕ。

(c) 设置主题个数K为5,10,20，使用*dataset_cn_full.txt*数据集，针对不同的K显示每个topic中出现频率最高的8个单词。

(d) 观察结果，找到主题分类效果最好的K，并分析原因。

补充说明：

1. 本次代码框架中使用了scipy，log(gamma(x))是gammaln函数，log(gamma(x))的导数是psi函数。

2. 本次代码框架没有引入λ，在变分推断更新γ和ψ时可能与课件有所出入，同学们可以参考原论文中的这一更新过程。

3. 考虑到时间问题，对于大规模数据集*dataset_cn_full.txt*，最大更新轮次（max_ epochs）设置为10轮即可。