人工智能导论第四次作业

2022年6月

1 采样方法（10分）

(a) 证明拒绝采样得到的分布就是原分布。

(b) 证明Gibbs采样法的过程满足细致平衡（detailed balance）。

2 Baum-Welch算法（40分）

Baum-Welch算法是EM算法的一种，其解决了隐马尔科夫模型（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$P(X|\theta)$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。

(a) 首先进行**E**步的计算，请根据ELBO写出$J(\theta)$，并证明

$$\underset{\theta}{\operatorname{argmax}} J(\theta) = \underset{\theta}{\operatorname{argmax}} \sum_{Z} P(X, Z|\theta^{(i)}) \log P(X, Z|\theta)$$

其中$\theta^{(i)}$表示第i个优化轮得到的参数。

(b) 令$Q(\theta, \theta^{(i)}) = \sum_Z P(X, Z|\theta^{(i)}) \log P(X, Z|\theta)$，请用$\pi_{z_1}, b_j(k), a_{ij}$表示$P(X, Z|\theta)$，并将$Q(\theta, \theta^{(i)})$展开为三项之和，每项只与一个参数有关。

(c) 上一小题实现了参数之间的解耦，可以进入**M**步的计算。请利用拉格朗日乘子法计算$Q(\theta, \theta^{(i)})$取极大值时，$\pi_i, a_{ij}, b_j(k)$的值。（提示：可以先证$\sum_Z \log \pi_{z_1} P(X, Z|\theta^{(i)}) = \sum_{i=1}^{N} \log \pi_i P(X, z_1 = i|\theta^{(i)})$，拉格朗日乘子法的等式是概率之和为1。）

(d) 定义

$$\gamma_t(i) = P(z_t = q_i|X, \theta)$$
$$\xi_t(i, j) = P(z_t = q_i, z_{t+1} = q_j|X, \theta)$$

其中q表示隐状态，请使用前向概率$\alpha_t(i)$和后向概率$\beta_t(i)$表示$\gamma_t(i)$和$\xi_t(i,j)$。（提示：前向概率定义为$\alpha_t(i) = P(x_{1:t}, z_t = i|\theta)$，后向概率定义为$\beta_t(i) = P(x_{t+1:T}|z_t = i, \theta)$。这个表示中可以使用$a_{ij}, b_j(k)$，但不能使用$P(X|\theta)$。）

(e) 请使用$\gamma_t(i)$和$\xi_t(i,j)$表示$\pi_i, a_{ij}, b_j(k)$，并给出Baum-Welch算法的伪代码。

3 LDA实现（50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ϕ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。