人工智能导论第三次作业

2021年5月

1 第一题(10分)

从独立同分布的样本$\mathbf{x_1}, \ldots, \mathbf{x_n} \sim \mathcal{N}(\mu, \mathbf{\Sigma})$中估计$\mu$与$\mathbf{\Sigma}$，我们可以最大化对数联合密度函数:

$$\sum_{i=1}^{n} \log p(\mathbf{x}|\mu, \mathbf{\Sigma}) = -\frac{n}{2} \log |2\pi\mathbf{\Sigma}| - \frac{1}{2} \sum_{i=1}^{n} (\mathbf{x_i} - \mu)^T \mathbf{\Sigma}^{-1} (\mathbf{x_i} - \mu)$$

(1) 请计算验证μ与$\mathbf{\Sigma}$的最大似然估计具有如下形式:

$$\hat{\mu}_{\mathrm{MLE}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x_i}$$

$$\hat{\mathbf{\Sigma}}_{\mathrm{MLE}} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x_i} - \hat{\mu}_{\mathrm{MLE}})(\mathbf{x_i} - \hat{\mu}_{\mathrm{MLE}})^T$$

(2) 请证明：$\mathrm{E}[\hat{\mathbf{\Sigma}}_{\mathrm{MLE}}] = \frac{n-1}{n} \mathbf{\Sigma}$。

2 第二题(30分)

贝叶斯网络被广泛应用于医学诊断中，医生会通常根据患者的历史行为与症状判断病因。在新冠疫情流行期间，准确有效的医学检测是疫情防控的重中之重。对于新的病人，医生需要判断该病人患有普通肺炎R_1或新冠肺炎R_2。医生可以了解到该病人是否为密切接触者A_1，以及是否吸烟A_2，同时，可以获取肺部影像结果S_1，核酸检测结果S_2，是否干咳S_3以及是否呼吸困难S_4。已知密切接触者患有新冠肺炎的概率会更高，吸烟会提升患有普通肺炎的概率。新冠肺炎患者的四种症状均可能为阳性，而普通肺炎只可能导致干咳或者呼吸困难。

(1) 请画出最符合上述描述的贝叶斯网络图，并写出对应的条件概率乘积的分解形式。

(2) 请写出呼吸困难S_4的马尔可夫边界（Markov Boundary）。

(3) 描述（1）中的贝叶斯网络需要多少个独立参数？如果取消所有独立性假设，又需要多少个独立参数？（注：所有变量均为二值变量）

(4) 若病人有吸烟经历$A_1 = 1$，患哪些疾病的概率会发生变化；如果该患者已经观察到干咳症状$S_3 = 1$，吸烟经历又会影响患有哪些疾病的概率？

Table 1: 第三大题所用符号表。

公式符号	代码符号	描述
α	alpha	模型参数，document-topic分布的超参数
φ	varphi	模型参数，topic-word分布，代码中取了对数
γ	gamma	变分参数，当前推断文档的topic分布
ϕ	phi	变分参数，当前推断文档的每个词语的topic分布

(5) 在实际情况中，医生要根据所有的行为历史A_1，A_2与症状S_1，S_2，S_3，S_4判断病人是否患有新冠肺炎R_2。请写出对应该场景的条件概率，并利用贝叶斯公式与消元法写出计算过程。

3 第三题(60分)

请用python实现的Variational EM LDA。用于训练和预测的数据集分别放在两个文本文件中./dataset.txt与./infer.txt，一行表示一篇文档，可处理中文和英文。变分推断以及EM的收敛条件可以通过计算likelihood来实现，我们这里设置了固定的循环次数; 超参数α的更新被省略了，使用了初始值。几种重要参数的对应关系如Table 1。

本大题先用训练数据训练得到一个模型，输出得到的每个topic的top10的词语，然后对新来的测试文档进行了推断，通过 γ 参数得到预测的该文档属于的主题。下面是作业要求:

（1）根据提供的实例代码框架, 完成代码框架中缺失的变分推断部分。

（2）设置主题个数K分别为10, 20, 30, 50，并针对每种情况显示每个topic中出现频率最高的单词。

（3）观察结果，尝试找到该数据集的最佳K值，并分析原因。

本大题需要完成上述三项要求, 并且将结果与分析汇总成实验报告与代码一并提交, 实验报告不得超过3页。