重庆某双一流大学自然语言处理期末真题

考试时间：2020 年 12 月 19 日 14:00-16:00
回忆整理：Vayne Duan

写在前面：

1. 试卷总体比较简单，开卷，可以带计算器(有计算量，不过写分数也行应该).
2. 本回忆版真题于 2020 年 12 月 19 日下午 16:22 写成，刚回到寝室就开始写了
3. 计院专业课的试卷似乎都不准老师发出来，希望有学弟学妹们能将我”回忆试卷”的习惯传承下去，为之后的学弟学妹们做一点微小的贡献 O(∩_∩)O
4. 其余专业课的回忆版试卷也许可以在 **github.com/VayneDuan** 找到，记得 **star & follow**!

一、填空题(10 空 * 2 分 = 20 分)

1.”他将来学校讲学”：属于**组合型**歧义
2. 支持向量机的目标是寻找 **最大类间界限** 的超平面
3. 除了互信息，**困惑度** 也可以用于评价语言模型
4. 信息熵是用来度量 **不确定性** 的指标
5. 文本表示中，**向量空间**模型将文本分解为空间中的向量
6. 基于语义词典的消歧方法，用**语义范畴**作为主要因素[可能题目记错了，答案是这个没错]
7. 朴素贝叶斯，上下文的词语依赖于**文本类别**，词之间是**独立**的[书上原话]
8. 答案是 **概念 属性**，题目忘记了，是文本分类或者消歧相关的内容，是书上原话

二、简答题(5 道* 4 分 = 20 分)

1. 什么是数据平滑？为什么要使用数据平滑？
2. 什么是生成式模型?什么是判别式模型?
3. 简要叙述 n 元模型分词原理
4. 什么是生预料？什么是标注预料？
5. [忘记了]

三、计算题(2 道* 15 分 = 30 分)

1. 类似下面图片里的题，只不过句子换成了 “<BOS> 他 是 研究 生物 的 <EOS>”

模型参数估计

例如，给定训练语料：

“*John read Moby Dick*”，

“*Mary read a different book*”，

“*She read a book by Cher*”

根据 **2** 元文法求句子*John read a book*的概率？

模型参数估计

$$p(John|<BOS>)=\frac{c(<BOS>John)}{\sum_w c(<BOS>w)}=\frac{1}{3}\quad p(a|read)=\frac{c(read\quad a)}{\sum_w c(read\quad w)}=\frac{2}{3}$$

$$p(read|John)=\frac{c(John\quad read)}{\sum_w c(John\quad w)}=\frac{1}{1}\qquad p(book|a)=\frac{c(a\quad book)}{\sum_w c(a\quad w)}=\frac{1}{2}$$

$$p(<EOS>|book)=\frac{c(book\quad <EOS>)}{\sum_w c(book\quad w)}=\frac{1}{2}$$

$$p(John\ read\ a\ book)=\frac{1}{3}\times 1\times\frac{2}{3}\times\frac{1}{2}\times\frac{1}{2}\approx 0.06$$

训练集：

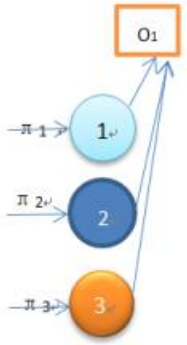
<BOS>John read Moby Dick<EOS>
<BOS>Mary read a different book<EOS>
<BOS>She read a book by Cher<EOS>

2. 给定 λ = {A,B, Π} 的三个矩阵，用前向算法求 P(O| λ)，类似下图

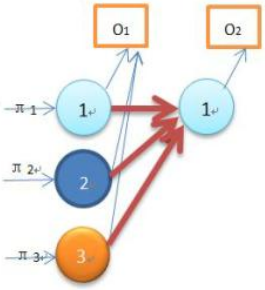
6.4 前向算法-实例分析

观察集合是: V={红，白}，M=2
状态集合是: Q={盒子1，盒子2，盒子3}，N=3
球的颜色的观测序列: O={红，白，红}
初始状态分布为: Π=(0.2,0.4,0.4)
其它转移概率、发射概率均已知。

（1）首先计算时刻1三个状态的前向变量：时刻1是红色球，
隐藏状态是盒子1的概率为:
α1(1)= π1b1(o1) = 0.2×0.5=0.1
隐藏状态是盒子2的概率为:
α1(2)= π2b2(o1) = 0.4×0.4=0.16
隐藏状态是盒子3的概率为:
α1(3)= π3b3(o1) = 0.4×0.7=0.28



6.4 前向算法-实例分析



球的颜色的观测序列: O={红，白，红}

（2）开始递推，时刻2三个状态的前向概率：时刻2是白色球
隐藏状态是盒子1的概率为:

$$\alpha_2(1) = \left[\sum_{i=1}^{3} \alpha_1(i)a_{i1}\right] b_1(o_2) = [0.1*0.5 + 0.16*0.3 + 0.28*0.2] \times 0.5 = 0.077$$

隐藏状态是盒子2的概率为:

$$\alpha_2(2) = \left[\sum_{i=1}^{3} \alpha_1(i)a_{i2}\right] b_2(o_2) = [0.1*0.2 + 0.16*0.5 + 0.28*0.3] \times 0.6 = 0.1104$$

隐藏状态是盒子3的概率为:

$$\alpha_2(3) = \left[\sum_{i=1}^{3} \alpha_1(i)a_{i3}\right] b_3(o_2) = [0.1*0.3 + 0.16*0.2 + 0.28*0.5] \times 0.3 = 0.0606$$

三、应用分析题(2 道* 15 分 = 30 分)

1. 简要叙述文本分类的主要任务和模型，请设计一个中文文本分类的系统实现方案
2. 简要叙述语义消歧的主要任务，请**分别**设计一个基于有有监督的 以及 基于词典的 实现方案