

# 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  $O(n \cdot n)O$
4. 其余专业课的回忆版试卷也许可以在 [github.com/VayneDuan](https://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

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

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

$$p(\langle \text{EOS} \rangle | \text{book}) = \frac{c(\text{book } \langle \text{EOS} \rangle)}{\sum_w c(\text{book } w)} = \frac{1}{2}$$

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

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  $\lambda = \{A, B, \Pi\}$  的三个矩阵，用前向算法求  $P(O | \lambda)$ ，类似下图

## 6.4 前向算法-实例分析

观察集合是:  $V=\{\text{红}, \text{白}\}$ ,  $M=2$   
 状态集合是:  $Q=\{\text{盒子1}, \text{盒子2}, \text{盒子3}\}$ ,  $N=3$   
 球的颜色的观测序列:  $O=\{\text{红}, \text{白}, \text{红}\}$   
 初始状态分布为:  $\Pi=(0.2, 0.4, 0.4)$   
 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:

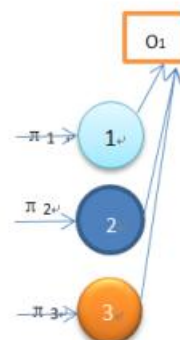
$$\alpha_1(1) = \pi_1 b_1(o_1) = 0.2 \times 0.5 = 0.1$$

隐藏状态是盒子2的概率为:

$$\alpha_1(2) = \pi_2 b_2(o_1) = 0.4 \times 0.4 = 0.16$$

隐藏状态是盒子3的概率为:

$$\alpha_1(3) = \pi_3 b_3(o_1) = 0.4 \times 0.7 = 0.28$$



## 6.4 前向算法-实例分析

球的颜色的观测序列:  $O=\{\text{红}, \text{白}, \text{红}\}$

(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:

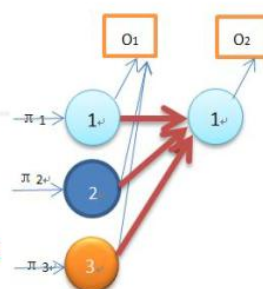
$$\alpha_2(1) = \left[ \sum_{i=1}^3 \alpha_1(i) a_{i1} \right] b_1(o_2) = [0.1 * 0.5 + 0.16 * 0.3 + 0.28 * 0.2] \times 0.5 = 0.077$$

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监督的 以及 基于词典的 实现方案

1. 文本分类系统实现方案: 输入文本, 首先进行分词, 然后进行特征提取, 之后根据上述特征对文档进行向量化表示, 权重用tf-idf表示。最后训练分类器, 可以用KNN/朴素贝叶斯算法。

2. 语义消歧的任务是确定一个多义词在给定的上下文语境中的具体含义。

Github VayneDuan