暑期学校实验项目：高考志愿填报助手

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 小组名称 | AI算法A组 | | | | | | |
| 姓 名 | 谈笑 | 专业 | 人工智能专业 | 班级 | 091181 | 学号 | 09118104 |
| 实验时间 | 2020.8.31-2020.9.23 | | 指导教师 | 孔祥龙 | | 成绩 |  |
| 1. **实验背景和目的**   新冠之下，为弘扬社会主义新时代的文化精神，贯彻东大人止于至善之高尚情操，本小组于今年上半年毅然投身于算法实现小组，立志于为即将进入大学的莘莘学子提供引导和帮助。高考志愿填报助手意在帮助高考完的考生根据自己所在省份的预测自身适合的大学与专业，有助于考生更加了解自己、明确目标定位。上学期，我们收集了各个省份的一分一段表以及部分高校的专业录取分数线，短学期内，我们组的任务目标为帮助考生预测可选大学，确定其冲一冲，稳一稳，保一保的若干所大学。 | | | | | | | |
| 1. **小组任务和个人任务**   **小组任务**：设计算法，根据考生的省份排名、地域位置、文理分科等信息帮助考生预测可选大学，确定其冲一冲，稳一稳，保一保的若干所大学。  **个人任务**：这次任务我被分配到的任务是完成算法的接口部分。包括给学生提供输入考试省份、高考排名、选科信息的接口，并通过与已有数据比对判断其正确性，随后将省份排名信息等进行处理转化为模型可以处理的数据形式，输入到模型中去。 | | | | | | | |
| **三、个人任务需求分析**  **1. 判断输入正确性**：对于省份信息的输入来说，我国共有34个省级行政区，对于学生输入的省份需要判断其是否在这34个区域名之中；对于排名信息的输入来说，数据类型则需要从原来输入的字符型改为正整数类型，并且其排名是有下限的；对于选科信息的输入来说，我国有两个地方的是不分文理科的，即上海和浙江，故对于这两个地方的考生来说，可以免去选科信息这一栏。  **2. 处理输入数据**：对于省份信息而言，为了给考生提供地域的选择，所以需要有可以数字化的数据来计算省份之间距离的远近。故而需要将考生输入的省份信息通过字典查找的方式转为省份归一化之后的经纬度；而对于排名信息而言，由于每个省份的考生人数不同，每个学校在各个省份的招生人数也不同，所以我们更需要的是该考生的排名在该省份所有人中对应的位置，故而我们需要从给定的每个省份的排名信息中，通过归一化得到考生的排名比例情况。 | | | | | | | |
| **四、实验过程（需附上关键代码及相关说明）**  **A. 判断输入正确性**  **1. 导入数据**：首先我们导入所有省份归一化之后的经纬度，并从中通过字典的key()函数提取出所有的省份。    **2. 省份信息**：假如身份不在provinces的列表中，那么需要则需要重新输入。    **3. 选科信息**：如果是来自上海或者浙江的考生，则直接跳过这一接口。    **4. 省份排名信息预处理**：导入省份排名信息，其中包括省份名称、省份最小排名和最大排名，同时为了保持数据的严谨性，验证最小排名和最大排名数字是否重合，如若重合则报错退出程序。    **5. 考生排名信息**：通过上面的数据处理我们接下来需要保证考生的高考名次位于最小排名和最大排名之间，否则无法通过。    **B. 处理输入数据**  **省份信息、考生排名信息**：我们需要通过归一化的式子将考生的排名信息转化为其在省份内的比例信息，接下来将信息按照模型要求的信息格式填入进行转化。 | | | | | | | |
| **五、实验结果与分析**  **A. 判断输入正确性**  **1. 区别分理科省份**（数据：江苏，文科，排名1000）      **2. 不区分文理科的省份**（数据：浙江，排名100）       1. **根本没有理科的西藏**     **B. 最后输入模型的数据**（数据：浙江，排名100） | | | | | | | |
| **六、实验总结与心得体会**  **A. 实验总结**  本次软件实践的课程中，我们小组主要负责是根据学生的地域和高考排名情况为学生进行合理的学校推荐的算法。深度学习作为一个跨领域的研究方向，其可应用领域是方方面面的。而在这次实践中，我们将深度学习中的人工神经网络用来解决回归问题，从学生省份、排名、选科的输入通过神经网络对信息的层层提取和筛选，最后为学生提供五所院校，保证学生能够根据结果选择冲一冲、保一保、稳一稳的院校。  但是在实际对模型进行训练构建的过程中，存在的**缺陷问题**仍然很多：   1. 模型使用的数据存在很大的数据不合理性，比如在数据清洗组提供的省份排名的csv文件中，福建文科总共仅有40个排名，西藏的文科仅有6人，甚至不存在理科，过于离谱。 2. 另一方面，推荐学校基于的地域过于宽泛，限于省份，因为缺少数据而不能扩展到城市，仅仅根据省份的大概经纬度来进行计算。   **B. 心得体会**  当然在本次实践中，我也学到了不少团队合作的心得体会。  1. 被划分做同一任务的一个团队的人，要经常互通有无，沟通任务进度，否则在前后交接之时，将花费大量的时间成本进行对接，无法让对方跟上自己的思路。  2. 对于那些队列式的工作，只有上一个人做完了，下面的人工作才能开始，这些工作应该给以明确的时间期限，人是容易拖延的，如果不规定时间期限，会导致后面的同学完成工程的时间减少，无法高质量准时完成任务。  3. 扩展的任务到底是做还是不做，从有决定的那一刻开始，应该积极联系数据组获得数据同时同步推进算法。 | | | | | | | |