暑期学校实验项目：高考志愿填报助手

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 小组名称 | AI算法实现A组 | | | | | | |
| 姓 名 | 王倩 | 专业 | 人工智能 | 班级 | 091181 | 学号 | 09118103 |
| 实验时间 | 2020.8.31-2020.9.23 | | 指导教师 | 孔祥龙 | | 成绩 |  |
| **一、实验背景和目的**  本实验致力于实现帮助高考生根据其所在地域以及分数排名，填报大学志愿的功能，为高考生们提供引导和帮助，选择最适合自己的大学。本学期的任务基于上学期收集的各个省份的一分一段表，并根据我们的需求，收集了各个省份以及大学的经纬度信息，从而使大学的预测更加合理。 | | | | | | | |
| **二、小组任务和个人任务**  ·小组任务：帮助考生推荐合适的大学  ·个人任务：设计推荐大学功能的实现流程，并编写模型代码和流程代码 | | | | | | | |
| **三、个人任务需求分析**  1. 个人任务需求：设计“实现用户输入自身的分数以及所在省份，模型输出为该用户推荐的大学，并根据本人情况，将推荐的大学分为冲一冲、稳一稳、保一保这三类大学”的实现流程，并编写流程实现代码  2. 分析过程：  1) 实现方法：对于实现推荐大学的功能，本次实验想要尝试通过神经网络的方法，首先根据需求设计合理的网络结构，然后用真实的学生数据对网络进行训练，让神经网络能够在输入学生分数和省份的情况下，输出对该学生所推荐的大学的预测。  为了实现“冲一冲、稳一稳、保一保”的推荐功能，首先根据预测选取多个符合条件的大学，经过中间的系列处理后，对最终的几个最佳大学在全国内的排名数据进行排序，令排名高的为冲一冲，低的为保一保，中的为稳一稳。  2) 细节处理：  a. 分数的处理：由于不同省份的分数所处排名以及排名百分比必定不同，所以在用户输入自身的分数后，除了一分一段表外，还需要有该省份的高考人数数据，来计算用户在该省份的排名百分比情况，使得相同排名、不同省份的考生数据更加公平。  b. 省份的处理：用户输入的省份只是字符串数据，为了满足更相近的地理数据更相似，需要搜集所有省份、所有大学的经纬度信息，并将数据归一化，消除与分数数据的量纲不一致问题。  c. 训练模型所需数据处理：由于模型实现的功能为输入（学生经度，学生纬度，学生排名），输出为某个特定的大学，所以训练数据也要构建成（学生经度，学生纬度，学生排名，所选大学）的格式。  但是由于只有一分一段表数据，所以我们需要根据各大学录取情况（包括录取人数和录取最高分和最低分），抽样生成一条一条的学生数据，作为真实的学生数据进行训练。 | | | | | | | |
| **四、实验过程（需附上关键代码及相关说明）**  实验内容主要分为三部分，一是预测流程的设计及实现，二是模型的设计及实现，三是在完成代码后，修改搜集到的大学数据为输入格式，输入并验证模型和流程的可行性。   1. 预测流程 2. 首先将学生数据输入神经网络1，令模型1输出预测的大学数据。本大学数据由（经度，纬度，排名）组成，是为通过训练神经网络，最终生成的数据，可能并非真实的大学数据。        1. 在所有的大学中，选出与第一步中预测的大学数组距离最近的前十所大学。实验中选取的是欧式距离。      1. 对于这十所大学，衡量它们的经纬度和排名因素，计算出大学得分。因为学生选学校要综合考虑学校的地理位置以及学校的排名，但是二者的重要程度占比不能确定，所以在这个步骤中我们通过训练神经网络2来计算两种因素的重要程度，从而输出综合考虑下大学的得分。得分越低，说明学校和学生的位置差异越小、学校的排名越靠前，从而越是理想的大学。实验中选取的位置差异为向量间的欧氏距离。      1. 选取前五个得分最低的大学，使得排名最靠前的为冲一冲，第二和第三靠前的为稳一稳，排名最末的两个为保一保。      1. 预测模型 2. 模型1：输入（学生排名，学生经度，学生纬度）->输出（大学经度，大学纬度，大学排名）   由于模型的输入为学生排名和学生经纬度所组成的数组，考虑到卷积神经网络更适合处理图片型数据，循环神经网络更适合处理文本类数据，本实验选取了全连接神经网络进行大学的预测。  为了深入学习到学生信息和大学信息之间的关系，模型1由五层全连接层组成。第一层为输入层，由三个神经元组成，分别代表学生的排名百分比、经度和纬度；第二、三、四层为中间层，分别由四个神经元组成，且附有sigmoid激活函数；第五层为输出层，由三个神经元组成，分别代表大学的排名、经度和纬度。最终通过一个softmax层将结果映射到0~1，完成输出。代码如下：     1. 模型2：输入（大学和学生的经纬度之差，大学排名）->输出（大学得分）   本模型由四层全连接层构成，第一层为输入层，由两个神经元组成，分别代表当前大学和学生的经纬度之差，以及大学的排名百分比；第二、三层为中间层，分别由四个神经元组成，且附有sigmoid激活函数；第四层为输出层，由一个神经元组成，代表大学的最终得分。代码如下：     1. 修改大学数据（去掉多余的行以及将字符串改成数字）：   原部分数据展示：    修改代码：    修改后的部分数据展示：     1. 验证流程及模型的可行性：   首先模拟用户输入的学生数据（经度，纬度，排名百分比）    经过上面的流程后，查看输出：   1. 流程1（模型1）的输出：      1. 流程2（与流程1相近的前十所大学）的输出：      1. 流程3（计算大学得分）的输出：      1. 流程4（得出最终的冲一冲稳一稳保一保结果）的输出： | | | | | | | |
| **五、实验结果与分析**  **实验结果：**如上面的验证过程结果所示，*流程1*的模型输出为三维向量，每维的数据范围0~1，为模型1为输入的学生数据预测的大学数据；*流程2*的输出结果为与模型1的结果最相近的十所大学数据；*流程3*的模型输出流程2的十所大学的对应得分；*流程4*的输出结果为通过对得分进行排序，选出的五所最佳大学及其冲稳保结果。  两个模型正常输出，流程代码正确实现。  **分析：***模型1*选取的输入特征为学生的地理信息和排名，为选取学校的主要因素，所以通过训练，可以使得模型较好地学习到学生信息和预测大学之间的内在关系，从而使得预测结果更加符合真实情况。*模型2*则在模型1的基础上，对与预测出的大学最相近的十所大学进行了评估，为了进行评估所选择输入的特征就是大学和学生的距离差以及大学的排名百分比，符合希望学校与家离得近、学校排名高的大部分心理。  *整体的预测流程*首先通过经过真实数据训练的模型，预测出符合真实情况的大学数据，然后计算出与预测的大学数据最相似的十所真实大学，作为学生根据自身情况筛选出的大学结果。接着再对这十所真实大学的自身数据进行评估，选出五所大学作为第二次筛选出的最佳结果。  **其中，第一次筛选强调的是学生的自身素质所匹配到的大学，第二次筛选则着重强调了大学的素质得分。**将学生数据与大学数据在最大程度上进行了利用。 | | | | | | | |
| **六、实验总结与心得体会**  通过这段时间的作业，我对于项目开发有了更加细致的了解，同时也体会到了团队协作的重要性。有很多看上去简简单单的任务实现起来却总是有诸多坎坷，很多细节问题都需要慢慢克服。  例如刚开始为模型准备数据的时候，并没有考虑到量纲问题，直到看到搜集到的经纬度数据，发现大部分数据都是在100以上，且彼此相差不多，才想起要数据归一化，免得使地理位置这维特征的重要性高出学生排名百分比（范围0~1）这维特征太多。  类似的问题还有很多，让我深刻体会到了开发一个项目的不容易。而对于小组协作方面，我们组在交流上还有所欠缺，不过也给了我们很好的经验，让以后的开发工作能够少走弯路，越来越顺利。 | | | | | | | |

2020年9月制