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 | 胡艺凡 | 专业 | 计算机科学与技术（人工智能） | 班级 | 091181 | 学号 | 09118127 |
| 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**  为贯彻东大人止于至善的精神，为莘莘学子提供更加便利和明确的目标指引，AI算法实现A组毅然投身于设计出能够为高三毕业生填报高考志愿的算法系统。为同学们的未来谋福祉，为同学们的梦想寻出路。高考志愿填报助手是一个根据学生的考试成绩来预测学生能够报考大学的AI助手。能够为学生提供“保一保、稳一稳、冲一冲”三个功能以供学生选择，使其对未来更加自信且充满希望。 | | | | | | | |
| **二、小组任务和个人任务**  小组任务：根据学生的高考分数排名，以及所采集大学的录取分数线和录取人数，可以训练一个能够预测学生报考学校的神经网络分类器  具体实施方案为根据各个省份的一分一段表以及88所高校的专业录取分数线，随机模拟若干名考生名次作为预训练数据，再利用神经网络分类器预测输出，还可以对输出进行筛选等副操作。  个人任务：对各组工作进行统筹协调，并且清洗数据模拟得到的数据，将各个大学的id对应成编码并且对接各个部分成员，将数筛选成三个dataset（文科、理科、综合）,使得数据集满足后续训练要求。 | | | | | | | |
| **三、个人任务需求分析**  作为组长，首先需要确定小组算法实现的目标，并且进行分工和管理。在召开短会之后，  我们首先确立了目标:用神经网络分类器预测输出学生报考的学校。随后我根据小组人数  经行了分工，为每个部分小组工作进行协调和统筹。  正由于我组长角色的存在，所以我负责整个小组的衔接工作，所以进行数据清洗使得数据  集满足后续训练要求。首先是第一小组给出了大学的排名、大学的经纬度和省份的经纬度。  算法训练组的要求为输入需要缩放至[0,1]，按照要求缩放后，考虑到各个大学和省份的地  理位置经度方面差异较小，纬度方面差异较大，所以要考虑不同的缩放标准来扩大其差异。  与此同时进行的是数据模拟小组得到的数据之后，我首先要对其进行核实检查，查看模拟  结果是否符合实际和预期。检查每个省份的最高排名和最低排名分别是多少，例如理想状  态下每个省份的最高排名应该都是1。但是在实际的操作过程中我们发现根据数据集模拟  出来的结果并不是每个省份都能满足这一情况。这首先肯定是数据集的纯度存在一定的问  题，在问题发现后我们马上进行复查，在对于清华北大这种特殊学校我们修改了模拟的阈  值，并且对有问题的数据进行了人为的修改和筛选。对于西藏这些数据量几乎不存在的地  区，我们最终决定将其纳入，因为对于神经网络的学习训练不存在太多的影响。 | | | | | | | |
| **四、实验过程（需附上关键代码及相关说明）**  # 以下是省份对应的经纬度并对其缩放  provinces = {  '吉林': [125.326800, 43.896160], '黑龙江': [126.662850, 45.742080],  '辽宁': [123.429250, 41.835710], '内蒙古': [111.765220, 40.817330],  '新疆': [87.627100, 43.793430], '青海': [101.780110, 36.620870],  '北京': [116.407170, 39.904690], '天津': [117.199370, 39.085100],  '上海': [121.473700, 31.230370], '重庆': [106.550730, 29.564710],  '河北': [114.469790, 38.035990], '河南': [113.753220, 34.765710],  '陕西': [108.954240, 34.264860], '江苏': [118.762950, 32.060710],  '山东': [117.020760, 36.668260], '山西': [112.562720, 37.873430],  '甘肃': [103.826340, 36.059420], '宁夏': [106.258670, 38.471170],  '四川': [104.075720, 30.650890], '西藏': [91.117480, 29.647250],  '安徽': [117.285650, 31.861570], '浙江': [120.153600, 30.265550],  '湖北': [114.342340, 30.545390], '湖南': [112.983400, 28.112660],  '福建': [119.296590, 26.099820], '江西': [115.910040, 28.674170],  '贵州': [106.707220, 26.598200], '云南': [102.709730, 25.045300],  '广东': [113.266270, 23.131710], '广西': [108.327540, 22.815210],  '香港': [114.165460, 22.275340], '澳门': [113.549130, 22.198750],  '海南': [110.348630, 20.019970], '台湾': [121.520076, 25.030724],  }  list\_1=[]  list\_2=[]  for i in provinces:  list\_1.append(int(provinces[i][0]))  list\_2.append(int(provinces[i][1]))  max\_1=max(list\_1)  min\_1=min(list\_1)  max\_2=max(list\_2)  min\_2=min(list\_2)  for i in provinces:  provinces[i][0]=(int(provinces[i][0])-min\_1)/(max\_1-min\_1)  provinces[i][1] = (int(provinces[i][1]) - min\_2) / (max\_2 - min\_2)  #将省份的经纬度存入list以此查出最大最小值进行放缩  # 找出大学对应的省份经纬度和排名进行同样的放缩  head=['大学','经度','纬度','排名'] d3=356 d3min=1 d2=22.934949800000002 d2min=22.8065434 d1=39.0372239 d1min=87.6061172 data = [] with open('college\_code(1).json', 'r', encoding='utf8') as f:  datas = json.load(f)  with open('colle\_shrink.csv','w',newline='') as p:  writer=csv.writer(p)  writer.writerow(head)  for i in datas:  D1=(float(i['college\_code'][0])-d1min)/d1  D2 = (float(i['college\_code'][1]) - d2min) / d2  D3 = (float(i['college\_code'][2]) - d3min) / d3  DataRow=[i['college\_name']]+[D1]+[D2]+[D3]  print(DataRow[0])  data.append(DataRow)  writer.writerows(data)  #缩放完毕，得到colle\_shrink.csv文件  #根据以上得到的文件建立起{‘大学’：[[经度，纬度],排名]}的字典以便于索引  with open ("colle\_shrink.csv",'r',encoding='gbk') as f:  collegue\_List=csv.reader(f)  print(collegue\_List)  sheet={}  for rows in collegue\_List:  if line!=0:  collegue\_name = ""  collegue\_number = ""  print(rows)  sheet[rows[0]]=[[float(rows[1]),float(rows[2])],float(rows[3])]  print(sheet[rows[0]])  line+=1  #将前面小组成员的文件全部合并  Province = ['北京','天津','河北','山西','内蒙古','辽宁','吉林','黑龙江','上海','江苏','浙江',  '安徽','福建','江西','山东','河南','湖北','湖南','广东','广西','海南','重庆','四川',  '贵州','云南','西藏','陕西','甘肃','青海','宁夏','新疆','香港','澳门','台湾'] category = ["文科","理科", "all"] print(sheet) NameList=['09118209祁丁然.csv','09118206陶特.csv','09118205王昕彤.csv','09118145邵彤.csv','09118122邵一展.csv', '09118101高捷.csv','清北.csv']  with open('out.csv', 'a', newline='', encoding="utf-8-sig") as q:  writer = csv.writer(q)  row=['rank','province','category','college']  writer.writerow(row) # for names in NameList:  print(names)  c=0  with open('./新建文件夹/'+names, 'r', encoding='UTF-8') as f:  reader=csv.reader(f)  print(reader)  with open('out.csv','a',newline='',encoding="utf-8-sig") as q:  writer=csv.writer(q)  for row in reader:  if c!=0:  writer.writerow(row)  c+=1  #讲合并后的数据全部按要求经行放缩  with open('result\_1.csv','a',encoding='utf8') as o:  tittle=['rank','province','cate','coll\_rank','locate']  writer=csv.writer(o)  for P in Province:   for c in category:  Rank = []  with open('out.csv',encoding='utf-8') as p :  reader=csv.reader(p)  for row in reader:  if row[2]==c and row[1]==P:  Rank+=[int(row[0])]  if Rank:  with open('out.csv', encoding='utf-8') as p:  reader = csv.reader(p)  for row in reader:  if row[2]==c and row[1]==P:  print(min(Rank),max(Rank),P,c)  print(row[3])  rank=(int(row[0])-min(Rank))/(max(Rank)-min(Rank))  location=provinces[row[1]]  print(location)  Out\_Put=[rank,location,c,sheet[row[3]][1],sheet[row[3]][0]]  writer.writerow(Out\_Put)  #最后是对省份排名经行检查功能的函数  for P in Province:   for c in category:  Rank = []  with open('out.csv',encoding='utf-8') as p :  reader=csv.reader(p)  for row in reader:  if row[2]==c and row[1]==P:  Rank+=[int(row[0])]  if Rank:  print(P,c,max(Rank),min(Rank)) | | | | | | | |
| **五、实验结果与分析**  E:\qqfile\1070324893\Image\C2C\Image1\VYG{78XLPG%[(GJ)7E0$1LR.png  最终得到合并交给训练组的数据集如上如所示。分别为学生排名、经度、纬度、文理科、所在省  份的经度、纬度、和所选大学排名。这样讲前面的省份组和模拟组的数据全部结合起来共给训  练组训练，能够达到较好的效果和预期。 | | | | | | | | |
| **六、实验总结与心得体会**  本次软件实践，我承蒙小组成员们的厚爱担任了小组长。深感对于整个小组的统筹和把握有许  多不到位的地方，在此深表惭愧。这次的实践对我们小组的每一位成员来说是一次创新的尝试，  也是一次对自我的挑战。我们需要考虑许多方面的因素，从学生的角度出发，我们需要回顾当  年填报志愿时我们最需要的是什么；从数据的角度出发，我们需要考虑如何利用现有的有限  的数据得到一个最为完善的结果；从算法实践的角度出发，我们需要考虑使用什么模型能够让  我们的结果最为理想。为此我们召开了许多次会议，作为组长，我得到了规划和总结能力的锻  炼。同时如何巧妙地催促组员在最短时间内完成任务也是一门学问。在整合清洗数据的过程中，  我的代码能力也得到了进一步的提升。在和算法实现小组同学的交流过程中，我学习倒了新的  知识和想法，视野思路也因此得到了更进一步的拓宽。由于我们是整个小组共同承担算法任务，  如何去分配每个人的任务也是一项挑战。俗话说‘不患寡而患不均’，能够分配出一个让同学们  没有异议的任务方案，是一个非常大的挑战。作为小组长，我采取了自愿认领和强制分配相结  合的手段，让同学们都能够很乐于接受自己的任务。于此同时，我感受到了合作交流的重要性。  每一次的会议，都能够推进小组的一大进步。所以再今后的学习生活之中，当有小组任务之时，  交流和沟通是非常重要的。如果有机会的话，我仍然愿意去承担小组长这个职位。 | | | | | | | | |

2020年9月制