需求分析

本组计划用AI算法实现学校推荐

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。具体说明如下：

1. 整体概览

模型共分为三类，一类供文科生使用，一类供理科生使用，一类供不分文理的特殊省份学生使用。

1. 数据准备
2. 大学：每所大学根据其地理位置和大学档次进行编码，初步想法是将大学编码为二维向量，第一维是大学所在省份编号，第二维是大学排名
3. 省份：大学省份和学生省份的编号一致，要求地理位置上相近的省份之间的编号差距尽可能小，地理位置上相距很远的省份编号相差很大

三. 模型的输入输出

1. 输入：（学生排名，学生省份）

2. 输出：推荐大学

四. 模型的训练数据

学生数据（学生排名，学生省份，学生所选大学）

说明：因为目前对某个特定排名、省份的学生所选大学数据是无法获取的，所以首先我们需要根据实际情况自动生成。生成过程为：对于每所大学，获取其当年对每个省份的录取人数N以及最高最低分（或平均分也可），然后在这个分数区间内随机生成N个分数，然后对照一分一段表获取该分数所在排名，这样我们就有了类似（学生排名，学生省份，学生所选大学）这样的数据。且数据量很大。

五. 过程原理

对模型输入（学生排名，学生省份）后，先根据排名找出109所采集学校中可以上的大学，然后从这些大学中根据（大学排名，大学省份，学生省份）这三个特征条件进行筛选。

例如：小红可选的大学有 ([1,3], [2,4], [1,6], [7,10]) （其中每个二维数组都是一所大学），她自己的省份是（3）。在这几所学校中，小红所要选的学校，要满足：

H = a\*F(大学第一维特征-小红省份) + b\*G(大学第二维特征)

要尽可能小，即满足：小红省份和大学省份要近+大学排名要高。

（参数说明：其中a、b为权重，是我们训练神经网络需要优化的参数；F和G函数初步想法可以取二次幂函数，在后续开发过程中也可以根据效果进行修改。）

在可选的大学中，对每所大学计算它的H函数，最终选择H函数值最小的作为最佳推荐大学。

训练过程中，模型的Loss值可定义为：（设小红真正所选大学为Y，推荐小红所选大学为Y’）

Loss = (Y[0]-Y’[0])^2+(Y[1]-Y’[1])^2

跟据Loss值进行反向传播，优化参数a、b。

六. 功能扩展

1 最终模型的输出可以是：

1. 上述所说的最佳大学（H函数最小）
2. H函数较小的前若干个大学

2 可以为学生提供以下几个功能“

1. 学校省份选择：可以选择自己想去的若干个省份（筛选完学生排名可以上的大学后，再进行省份的筛选，最终在二次筛选的结果上进行最佳大学的评定）
2. 双一流选择：选择自己是否只想去双一流大学
3. 一二本：选择自己是否只想去一本

（b、c条件可能采集的109所大学都满足了？我不知道，如果不是的话还需要给大学增加是否是双一流、是否是一本的标签，便于筛选）

3 我们当前做的模型是为了那些想要离家近的学生准备的，且参数的权重也是基于此的，这符合大多数人的想法。但是为了功能扩展，后续有时间的话可以考虑为用户提供这样的功能 ：选择是否想要离家近。然后另外训练一个离家越远越好的大学推荐模型