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 | 徐浩卿 | 专业 | 计算机科学与技术（人工智能） | 班级 | 091181 | 学号 | 09118120 |
| 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。 | | | | | | | |
| **二、小组任务和个人任务**   * 小组任务：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：根据生成的样本数据和设计的模型结构，进行模型的训练并保存参数。 | | | | | | | |
| **三、个人任务需求分析**   1. 实现设计好的预测模型：组内经过讨论后，得到一个或数个需要训练的模型设计，通过这些模型实现对合适大学的预测。需要将这些模型使用Pytorch实现，用于进一步的训练和实际应用。 2. 预处理和生成数据集：组内根据数据收集组发来的数据进行模拟数据生成，需要根据这些模拟数据做预处理。当训练好模型1后，还需要根据这些模拟数据生成用于模型2训练的数据集。也即，需要处理1、2两个数据集，分别的样本容量约为300万和3000万。 3. 训练模型：实现两个模型的训练代码，并使用数据集进行训练。其中模型具有一定量的超参数，需要在训练后进行验证并选取最佳的超参数。 4. 保存模型参数：训练后，在测试集上测试查看大致效果，之后将已训练好的模型参数保存下来，供后续处理以及投入使用。 | | | | | | | |
| **四、实验过程（需附上关键代码及相关说明）**  此处以训练理科预测模型为例，说明模型训练的代码与流程。   1. 数据预处理：将小组内准备好的训练原数据进行预处理工作，主要内容为将文理科及不分科的数据分开保存以便之后训练。 2. 训练模型（模型1与模型2的训练过程大致相同）：    1. 划分训练集：将数据集划分为训练集和测试集，此处选择75%的数据用于训练。   设定损失函数与优化器：此处选择MSE作为损失函数，并且选择带动量的mini-Batch SGD作为优化器。      训练：使用CUDA在GPU上进行模型的训练。      测试并保存：给出测试MSE并保存模型参数。 | | | | | | | |
| **五、实验结果与分析**  实验结果：  模型1进行训练后，训练集上的MSE大约为0.000143，平均在每个输出维度（大学排名、大学经纬度）上的误差约为6.1%。测试集上的MSE大约为0.000187，平均在每个输出维度上的误差为7.8%。  模型2进行训练后，训练集上的MSE大约为0.136，测试集上的MSE大约为0.169。  实验结果分析：  从实验结果来看，模型1训练后预测值贴合较好，预测准确度较高，主要可能是由于选取的特征与输出确实有较大的相关性。模型2训练后效果一般，准确率不高，可能是由于数据集中样本点非常稠密，2种输出有较大的混叠部分，导致模型分类效果不佳。后期优化可考虑加深网络或增加特征维度以获得更好的分类效果。 | | | | | | | |
| **六、实验总结与心得体会**  经过4周的小组项目开发工作，我深刻体会到了沟通讨论的重要性。小组合作项目中，一个人闷头干活是不可行的，必须要将想法与组员互相交流，确认彼此的想法和计划。否则，一个简单的小项目也可能成为十分困难的事情。而如果及时讨论出一个切实可行的具体方案，并将每个步骤分配给各人，一个宏大的工程也能迎刃而解了。  对于我们组所负责的工作：推荐大学算法，一开始大家也的确感到不知所措，不知道如何设计。好在经过我们数次的讨论后，确定了大致的算法流程，大家便感到这一复杂问题的解决容易了起来。在实现我们方案的过程中的确也遇到了不小的困难与失败，大家也都互相协助，将问题一个个解决，一步步地实现了整个模型。  此外，不仅是组内成员需要沟通，整个项目的数个小组也十分需要交流协作。本次软件实践中，我们在组间交流上有所欠缺，导致数据上出现了一些问题。在今后的学习生活中，在需要协同完成的任务上，交流的艺术我们应当更加重视。 | | | | | | | |

2020年9月制