机器学习项目报告 3: 多层感知机实战

1. 模型实现

（1）数据预处理

浏览数据

- 由于数据包含多个表格和超多的特征，而且MLP前期工作（特征筛选）要靠人来做，先浏览一遍，大体确定相关数据。另外要看'Key to Variable Encodings'，了解数字代表的意义（缺失值？是否？）
  数据加载
- 将kaggle zip下载到本地，创建字典，使用哈希验证确保数据的完整性，解压，csv读取
  创建训练集和测试集
- 通过 `FOLIO` 将两个数据集合并，筛选了与退学可能的相关特征，删除了无关特征，删除了目标特征中含缺失值的行
- 将数据转换为正态分布，提高模型收敛速度
- 使用train_test_split()按8:2比例划分训练集和测试集
- Dataframe > np数组 > pythorch张量
- 小批量加载数据，批次大小设置为64

（2）模型架构

MLP模型

- 输入层：对应特征个数
- 隐藏层：2个全连接层 + ReLU激活函数 + dropout
- 损失函数：交叉熵损失
- 优化器：Adam优化器
- 训练周期：根据测试集表现调整
- lr=0.01
  训练模块
- 清空梯度 > 前向传播 > 反向传播更新梯度 > 优化
- 用累加求loss,accuracy
  测试模块
- 不再更新梯度

2. 训练过程

- 由前几次反馈了解到Epoch设置=50左右合适
- lr试了=0.001，accuracy在60%，太慢了；=0.1，约为85%，步长过大不精准，最后定在了0.01
- 用的是CrossEntropyLoss,只支持二分类问题，但我未注意到目标特征有012三个值，采用删除法将target=缺失值（2）的行删去
- 优化：之后使用了Adam自适应优化器，根据更新频率自动调整lr,更好的拟合函数，基本上Epoch=11时就接近100%
- 加与不加

```
Epoch [1,Train Loss: 0.1757,Train Accuracy: 91.67%,
Epoch [11,Train Loss: 0.0948,Train Accuracy: 94.29%,
Epoch [20,Train Loss: 0.1230,Train Accuracy: 94.52%,

[5 rows x 67 columns]
Epoch [1,Train Loss: 0.1840,Train Accuracy: 93.57%,
Epoch [11,Train Loss: 0.2367,Train Accuracy: 87.35%,
Epoch [20,Train Loss: 0.2083,Train Accuracy: 92.37%,
```

3. 改进策略(以{问题>方法>效果}展开)

过拟合

- 增加Dropout函数
- 减少隐藏层数，由两个隐藏层减为一个
- 减小数据，将训练数据占比由80%改为60%，更好的测试模型

```
   FOLIO  q_personas  q_hombres  q_mujeres  x<30   pc  laptop  tv_plana  \
0      1           1          1          0     2  0.0     0.0       0.0
1      2           4          2          2     1  1.0     1.0       1.0
2      3           2          0          2     2  0.0     0.0       0.0
3      4           1          0          1     2  0.0     0.0       0.0
4      5           2          1          1     2  0.0     0.0       0.0

   tablet  smartphone  InternetF  sIF_causa  mejora_vida  mejora_trabajo  \
0     0.0         0.0        0.0        0.0          0.0             0.0
1     1.0         1.0        1.0        0.0          1.0             1.0
2     0.0         0.0        0.0        0.0          0.0             0.0
3     0.0         0.0        0.0        0.0          0.0             0.0
4     0.0         0.0        0.0        0.0          0.0             0.0

   mejor_decisor  ENT  FACTOR
0            0.0    1     925
1            1.0   24     809
2            0.0    2    1248
3            0.0    9    8713
4            0.0    5    1292
   FOLIO  SEXO  EDAD  edu_inicial  inscrito  nivel_edu  terminado  nt_causa  \
0      2     1    13          0.0       1.0        3.0        1.0       0.0
1      2     2    19          0.0       1.0        9.0        1.0       0.0
2      7     2     8          0.0       1.0        3.0        1.0       0.0
3      7     1    26          0.0       2.0        0.0        0.0       0.0
4      9     1    27          0.0       1.0        9.0        1.0       0.0

   asesorias  extraord  ...  tv_plana  tablet  smartphone  InternetF  \
0        2.0       0.0  ...       1.0     1.0         1.0        1.0
1        1.0       2.0  ...       1.0     1.0         1.0        1.0
2        2.0       0.0  ...       1.0     2.0         1.0        1.0
3        0.0       0.0  ...       1.0     2.0         1.0        1.0
4        2.0       2.0  ...       1.0     1.0         1.0        1.0

   sIF_causa  mejora_vida  mejora_trabajo  mejor_decisor  ENT  FACTOR_y
0        0.0          1.0             1.0            1.0   24       809
1        0.0          1.0             1.0            1.0   24       809
2        0.0          2.0             2.0            3.0   22       722
3        0.0          2.0             2.0            3.0   22       722
4        0.0          1.0             1.0            1.0   26       425

[5 rows x 67 columns]
Epoch [1,Train Loss: 0.0363,Train Accuracy: 98.77%, Test Accuracy: 100.00%
Epoch [11,Train Loss: 0.0050,Train Accuracy: 99.88%, Test Accuracy: 100.00%
Epoch [20,Train Loss: 0.0024,Train Accuracy: 99.92%, Test Accuracy: 99.98%
```

```
正在从file://national-survey-on-school-dropout.zip下载data\national-survey-on-school-dropout.zip...
正在从本地复制 national-survey-on-school-dropout.zip 到 data\national-survey-on-school-dropout.zip...
   FOLIO  q_personas  q_hombres  q_mujeres  x<30   pc  laptop  tv_plana  \
0      1           1          1          0     2  0.0     0.0       0.0
1      2           4          2          2     1  1.0     1.0       1.0
2      3           2          0          2     2  0.0     0.0       0.0
3      4           1          0          1     2  0.0     0.0       0.0
4      5           2          1          1     2  0.0     0.0       0.0

   tablet  smartphone  InternetF  sIF_causa  mejora_vida  mejora_trabajo  \
0     0.0         0.0        0.0        0.0          0.0             0.0
1     1.0         1.0        1.0        0.0          1.0             1.0
2     0.0         0.0        0.0        0.0          0.0             0.0
3     0.0         0.0        0.0        0.0          0.0             0.0
4     0.0         0.0        0.0        0.0          0.0             0.0

   mejor_decisor  ENT  FACTOR
0            0.0    1     925
1            1.0   24     809
2            0.0    2    1248
3            0.0    9    8713
4            0.0    5    1292
   FOLIO  SEXO  EDAD  edu_inicial  inscrito  nivel_edu  terminado  nt_causa  \
0      2     1    13          0.0       1.0        3.0        1.0       0.0
1      2     2    19          0.0       1.0        9.0        1.0       0.0
2      7     2     8          0.0       1.0        3.0        1.0       0.0
3      7     1    26          0.0       2.0        0.0        0.0       0.0
4      9     1    27          0.0       1.0        9.0        1.0       0.0

   asesorias  extraord  ...  tv_plana  tablet  smartphone  InternetF  \
0        2.0       0.0  ...       1.0     1.0         1.0        1.0
1        1.0       2.0  ...       1.0     1.0         1.0        1.0
2        2.0       0.0  ...       1.0     2.0         1.0        1.0
3        0.0       0.0  ...       1.0     2.0         1.0        1.0
4        2.0       2.0  ...       1.0     1.0         1.0        1.0

   sIF_causa  mejora_vida  mejora_trabajo  mejor_decisor  ENT  FACTOR_y
0        0.0          1.0             1.0            1.0   24       809
1        0.0          1.0             1.0            1.0   24       809
2        0.0          2.0             2.0            3.0   22       722
3        0.0          2.0             2.0            3.0   22       722
4        0.0          1.0             1.0            1.0   26       425

[5 rows x 67 columns]
Epoch [1,Train Loss: 0.1305,Train Accuracy: 97.37%, Test Accuracy: 97.92%
Epoch [11,Train Loss: 0.0777,Train Accuracy: 98.09%, Test Accuracy: 97.92%
Epoch [20,Train Loss: 0.0755,Train Accuracy: 98.09%, Test Accuracy: 97.92%
```

数据处理

- 问题1：回看了李沐的《动手-数据预处理》，发现没有处理置信度（0-9）和是否（01）之间可能对参数学习产生不同程度的问题
- 使用StandardScaler进行特征标准化，转换为正态分布，使不同尺度的特征（置信度和是否）具有可比性
- 效果：其实在accuracy上没有太体现出来()，但我至少安心一点了

- 问题2：冗余的无关数据过多
- 看了Kaggle的讨论区，学到了merge()函数，自己筛选了相关特征，在学长的提醒下关注到了inscrito
- 重新筛选后准确率不在虚高了
- 问题3：开始时做数据处理特征列没管缺失值
- 用mean()求均值代替了缺失值（9）

哈希验证逻辑

- 开始看代码不太理解哈希验证的逻辑
- 问了AI发现这种逻辑在语法中很常见也很美，基本逻辑就像先微分在积分的感觉，单项求解再累加的思想不论是在求loss，accyracy还是在密码验证上都很有用

内外函数

- 总是遇到error：not defined “xx”,先是dase_dir又是优化器的
- 发现都是内外函数设置的有问题导致的，就算我返回了内函数的值，仍需要再设置一个全局变量接受返回值
- data_dir = download_extract('national-survey-on-school-dropout')

相对路径和绝对路径

- 绝对路径：从根目录开始的完整路径，在任何位置都指向同一个文件或目录。（有usersname）
- 相对路径：相对于当前工作目录的路径，根据当前位置不同可能指向不同的文件。

4. 运行结果

```
正在从file://national-survey-on-school-dropout.zip下载data\national-survey-on-school-dropout.zip...
正在从本地复制 national-survey-on-school-dropout.zip 到 data\national-survey-on-school-dropout.zip...
   FOLIO  q_personas  q_hombres  q_mujeres  x<30   pc  laptop  tv_plana  \
0      1           1          1          0     2  0.0     0.0       0.0
1      2           4          2          2     1  1.0     1.0       1.0
2      3           2          0          2     2  0.0     0.0       0.0
3      4           1          0          1     2  0.0     0.0       0.0
4      5           2          1          1     2  0.0     0.0       0.0

   tablet  smartphone  InternetF  sIF_causa  mejora_vida  mejora_trabajo  \
0     0.0         0.0        0.0        0.0          0.0             0.0
1     1.0         1.0        1.0        0.0          1.0             1.0
2     0.0         0.0        0.0        0.0          0.0             0.0
3     0.0         0.0        0.0        0.0          0.0             0.0
4     0.0         0.0        0.0        0.0          0.0             0.0

   mejor_decisor  ENT  FACTOR
0            0.0    1     925
1            1.0   24     809
2            0.0    2    1248
3            0.0    9    8713
4            0.0    5    1292
   FOLIO  SEXO  EDAD  edu_inicial  inscrito  nivel_edu  terminado  nt_causa  \
0      2     1    13          0.0       1.0        3.0        1.0       0.0
1      2     2    19          0.0       1.0        9.0        1.0       0.0
2      7     2     8          0.0       1.0        3.0        1.0       0.0
3      7     1    26          0.0       2.0        0.0        0.0       0.0
4      9     1    27          0.0       1.0        9.0        1.0       0.0

   asesorias  extraord  ...  tv_plana  tablet  smartphone  InternetF  \
0        2.0       0.0  ...       1.0     1.0         1.0        1.0
1        1.0       2.0  ...       1.0     1.0         1.0        1.0
2        2.0       0.0  ...       1.0     2.0         1.0        1.0
3        0.0       0.0  ...       1.0     2.0         1.0        1.0
4        2.0       2.0  ...       1.0     1.0         1.0        1.0

   sIF_causa  mejora_vida  mejora_trabajo  mejor_decisor  ENT  FACTOR_y
0        0.0          1.0             1.0            1.0   24       809
1        0.0          1.0             1.0            1.0   24       809
2        0.0          2.0             2.0            3.0   22       722
3        0.0          2.0             2.0            3.0   22       722
4        0.0          1.0             1.0            1.0   26       425

[5 rows x 67 columns]
Epoch [1,Train Loss: 0.1440,Train Accuracy: 97.16%, Test Accuracy: 99.91%
Epoch [11,Train Loss: 0.0549,Train Accuracy: 98.93%, Test Accuracy: 100.00%
Epoch [20,Train Loss: 0.1347,Train Accuracy: 96.21%, Test Accuracy: 99.99%
```

5. 收获

- 具体的收获都在代码里啦，总的来说：
- 虽然前面有02做理论上的铺垫，但真正早就实现了MLP后，发现对于MLP数据预处理是非常非常关键的（这可能也是CNN比它好的地方，感觉机器学习的目标之一就是让机器的学习流程越来越自主化），MLP基本框架大差不差，决定拟合能力的是人的筛选能力；于是乎我掌握了很多处理方法，更重要的是在面对大量冗余的无关数据时心里有了基本的流程图：
- 组合数据集>筛选目标项、相关特征项>处理缺失值>数据标准化>划分训练、测试集>转换为张量
- 学习到语法背后美丽的逻辑和规则，学到了很多新的函数，让它们能为我所用；科学使用了AI工具
- 算是解决了一个实际问题很有成就感，有了一通百通的幻想(),也认识到了ml的魅力，强大的数据和算力很可能在解决问题上超越一个人多年的学识