关于艾滋病患者是否在特定时间窗口内死亡的因素探究

Github: https://github.com/fita23689/Code-Final

中国人民大学统计与大数据研究院
郑濠锋
2023103371

一、背景介绍

艾滋病（艾滋病毒感染症病）是由人类免疫缺陷病毒（HIV）引起的慢性传染病。HIV 通过破坏人体免疫系统的工作机制，逐渐削弱机体的免疫功能，使感染者容易受到各种疾病和感染的侵袭。如果未经治疗，艾滋病最终会导致艾滋病最终期，即获得性免疫缺陷综合征（AIDS），这是一种严重的、晚期的 HIV 感染并引发多种严重感染和肿瘤的病症。目前，该病尚无有效的治疗方法，人们一旦感染艾滋病毒，就会终生携带，但通过妥善治疗，艾滋病毒可得到控制。艾滋病患者的存活时间与许多因素有关，如抗逆转录病毒治疗（ART）、免疫功能状态、年龄、饮食和生活方式等。本文使用 Logistic 回归，探究影响艾滋患者是否在特定时间窗口内死亡的因素，并试图对其背后的原理作出一定解释，为治愈或缓解艾滋病的研究提供数据支持。

二、数据介绍

本文选取 AIDS Clinical Trials Group Study 175 数据集，数据集的链接为艾滋病临床试验小组研究 175 - UCI 机器学习存储库，该数据集包含有关被诊断患有 AIDS 的患者的医疗保健统计数据和分类信息。该数据集最初发表于 1996 年。预测任务是预测每个患者是否在特定时间窗口内死亡，本文使用其探究影响患者是否在特定时间窗口内死亡的重要因素,并完成预测任务。

表 1 代表了数据集中的特征以及一些基本信息，从表中可看出该数据集共有 24 个变量，其中一个变量 cid 为 Target，即为需要预测的目标，而其余 23 个皆为可能对患者是否在特定时间窗口内死亡有一定影响，这将在下文详细展开探讨。可以从表中看出，因变量（cid）的类型为 Binary 零一变量，1 代表目标死亡，0 代表度过特定窗口时间，因此在之后的建模中可以考虑使用广义线性模型对模型进行拟合，此外，观察除因变量外的其他变量，可以看到有整数 Integer、连续型 Continuous 以及零一变量 Binary 三种，对不同变量类型来说，在数据处理时需要进行不同的处理，

这将在数据处理阶段详细展开。此外，在数据集的描述中注明此数据集并无缺失值，因此无需对缺失值进行处理。

表 1 数据集中的特征描述

count	name	role	type
0	pidnum	ID	Integer
1	cid	Target	Binary
2	time	Feature	Integer
3	trt	Feature	Integer
4	age	Feature	Integer
5	wtkg	Feature	Continuous
6	hemo	Feature	Binary
7	homo	Feature	Binary
8	drugs	Feature	Binary
9	karnof	Feature	Integer
10	oprior	Feature	Binary
11	z30	Feature	Binary
12	zprior	Feature	Binary
13	preanti	Feature	Integer
14	race	Feature	Integer
15	gender	Feature	Binary
16	str2	Feature	Binary
17	strat	Feature	Integer
18	symptom	Feature	Binary
19	treat	Feature	Binary
20	offtrt	Feature	Binary
21	cd40	Feature	Integer
22	cd420	Feature	Integer
23	cd80	Feature	Integer
24	cd820	Feature	Integer

通过对数据集基本特征的观察，可以认为使用Logistic回归进行对患者是否在特定时间窗口内死亡影响因素的探究，接下来将对数据集进行数据处理以使得其能满足Logistic回归对数据集的基本要求以及基本假设。

表2 VIF（方差膨胀因子）的数值

Features	VIF Factor	VIF Factor2	VIF Factor3
time	14.524	1.464	1.450
trt	7.185	2.551	1.021
age	18.376	1.135	1.127
wtkg	35.208	1.113	1.109
hemo	1.946	1.786	1.783
homo	8.141	2.765	2.752
drugs	1.289	1.124	1.121
karnof	73.850	1.061	1.052
oprior	1.227	1.200	1.035
z30	14.312	6.437	-
preanti	6.330	3.930	-
race	1.673	1.204	1.202
gender	12.936	2.227	2.212
str2	28.066	11.679	-
strat	75.660	13.344	1.138
symptom	1.276	1.067	1.061
treat	10.654	2.652	-
offtrt	2.000	1.335	1.328
cd40	18.689	1.934	1.546
cd420	17.173	2.262	1.865
cd80	15.799	3.025	-
cd820	16.379	3.023	1.092
const	-	372.514	362.095

三、数据处理过程

进行 Logistic 回归需要对数据进行的处理一般包括缺失值处理、异常值处理、特征选择、数据标准化以及处理样本不平衡等方法，下面试图使用这些方法对本文数据集进行处理，由上文介绍已知数据集并无缺失值，所以无需对数据集进行缺失数据的检查，且发现数据集中‘zprior’一列的值全为 1，其含义是在 175 天前是否服用过 ZDV（齐多夫定）药物，这表明所有样本再 175 天前都服用过 ZDV，所以该变量不存在变异性，将其删去。

由于变量较多，考虑对变量进行筛选，首先考虑变量之间的多重共线性问题，计算变量之间的 VIF（方差膨胀因子），如表 2 第二列所示，可以看到有许多变量都存在 VIF 值较大的问题，说明数据存在比较严重的多重共线性，尝试加入常数项 const 之后，从表 2 第三列中可以看到 VIF 值都有显著下降的现象，但还是有一些变量的 VIF 值较大，考虑使用相关系数进行筛选。

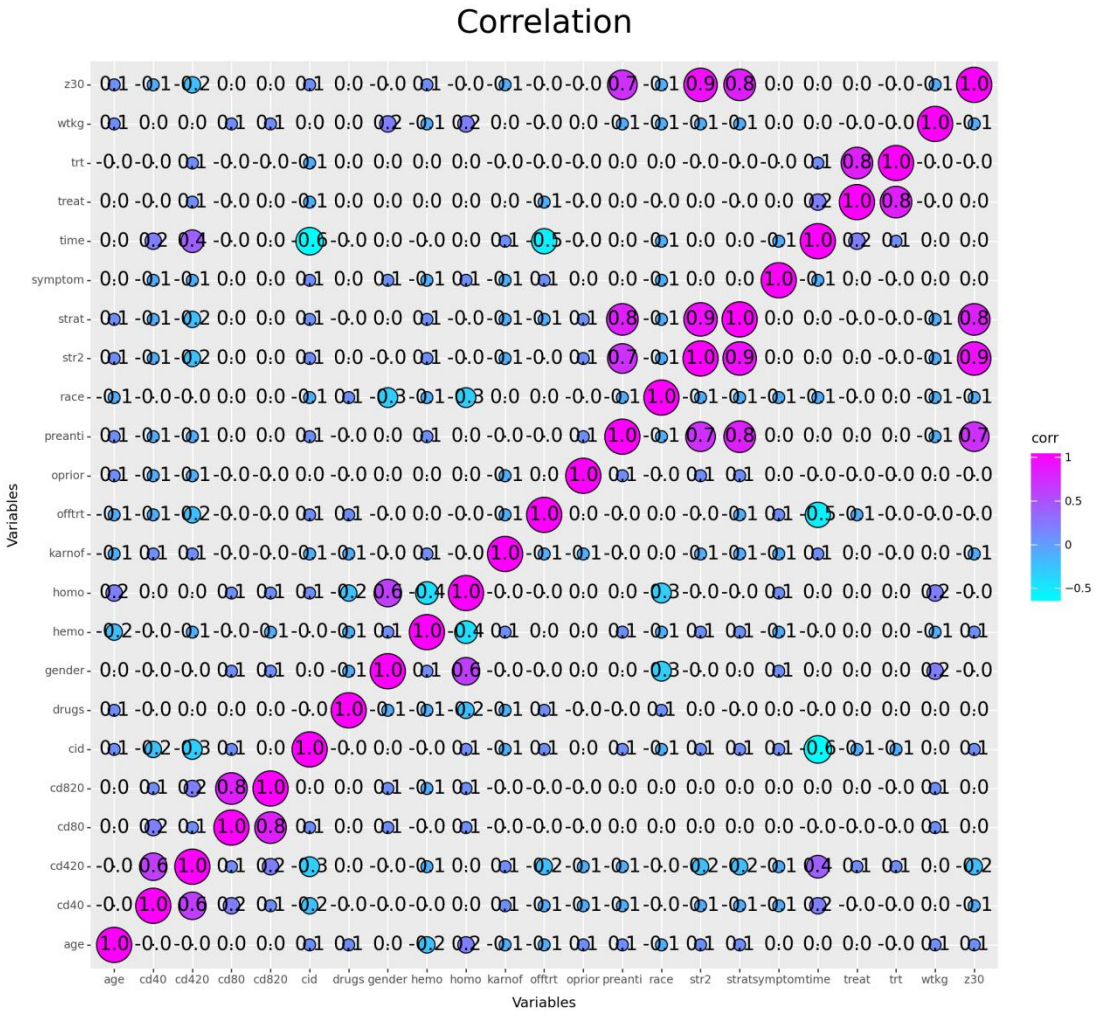

图 1 变量相关系数图

画出相关系数图如图 1 所示，可以看出因变量中存在一些变量之间确实有相关系数较强的现象，通过筛选相关系数在 0.7 以上的变量，得到以下几组变量如表 3 所示，可以看出几组变量不仅在数值层面上相关系数较高，且从含义上来看也存在较大关联，于是综合考虑选择删去 treat、z30、preanti、str2 以及 cd80 五个变量，之后进行 VIF 值的计算得到结果如表 2 中第四列所示，除了常数项每一个变量的 VIF 值都在 3 以下，从 VIF 的角度来说可以认为剩下的变量不存在多重共线性。

表 3 相关系数大于 0.7 的因子及其含义

factor_1	factor_2	corr	factor_1 meaning	factor_2 meaning
trt	treat	0.776	治疗指标：0 代表只有 ZDV；1 代表 ZDV+ddl；2 代表 ZDV+Zal；3 代表只有 ddl	治疗指标：1 代表只有 ZDV，2 代表其他
z30	str2	0.903	175 天前的 30 天前是否服用过 ZDV	是否服用过抗逆转录病毒药物
z30	strat	0.849	同上	抗逆转录病毒治疗史：1 代表未经过该治疗；2 代表在 1 周前且 52 周内接受过该治疗；3 代表在 52 周前接受过该治疗
preanti	strat	0.833	在 175 天前的前多少天进行过抗逆转录病毒治疗	同上
str2	strat	0.917	上述已解释	同上
cd80	cd820	0.756	实验开始时的 cd8 指标	在 20±5 周时的 cd8 指标

由 trt 与 strat 的含义可知这两个变量为分类变量，且类别种数分别为 3 与 4，于是将这两个变量转化为哑变量，分别得到变量 trt_0，trt_1，trt_2 与 strat_1，strat_2。

由于变量的数量较多，接下来使用 stepwise 即逐步回归法对变量进行筛选，逐步回归法所进行的加入变量与删除变量的具体过程如表 4 以及图 2 所示，使用的标准为AIC准则。stepwise得到的结果相比于之前的结果删去了 trt_0，trt_1，trt_2，wtkg，homo，oprior，gender，cd40，其中 trt 的含义在表 3 中已给出，

删去 trt 可能代表不同的治疗方式可能对艾滋患者是否在特定时间窗口内死亡并无影响，也可能是因为该变量是通过作用在其他变量上以造成影响；wtkg 代表体重，这可能说明体重的影响较小，因为体重可能通过很多因素共同影响，对艾滋患者是否在特定时间窗口内死亡并无影响；homo 代表是否同性恋，可能是同性恋只对是否患上艾滋病有影响，而对艾滋病的后续发展并无影响；oprior 代表在 175 天前是否接受过非 ZDV 抗逆转录病毒治疗，可能对因变量无影响；gender 代表性别，可能性别对因变量无影响。注意在逐步回归函数中，使用了 sm.add_constant 来对自变量加入常数项，因此在调用逐步回归函数前需要先将已有的常数项删去，为了防止自变量数据中不存在常数项引发 Keyerror，因此使用异常捕获的方法，如果不存在常数项，就直接将自变量数据作为逐步回归函数的参数。

表 4 逐步回归的过程

Step	Number of variances	Add	Drop	AIC
1	1	time	-	1705.571
2	2	offtrt	-	1582.750
3	3	cd420	-	1496.232
4	4	race	-	1474.114
5	5	cd820	-	1461.671
6	6	age	-	1453.408
7	7	strat_1	-	1446.578
8	8	symptom	-	1441.945
9	9	drugs	-	1439.265
10	10	karnof	-	1436.230
11	11	strat_2	-	1433.813
12	12	hemo	-	1431.485
13	13	cd40	-	1431.130
14	14	gender	-	1432.104
15	13	-	gender	1431.130

画出箱线图观察变量的异常值情况，结果如图 3 所示，由于变量过多，对变量进行筛查后，再次画出箱线图，如图 4 所示，异常值主要集中在 cd420 与 cd820 两个变量上，观察发现异常值都是大于均值的值，而 cd8 与 cd4 都是 CD4 和 CD8 都是 T 淋巴细胞，可能是在用药之后能够引起患者体内这两种细胞的极速增加导致，因此不作异常值的处理。

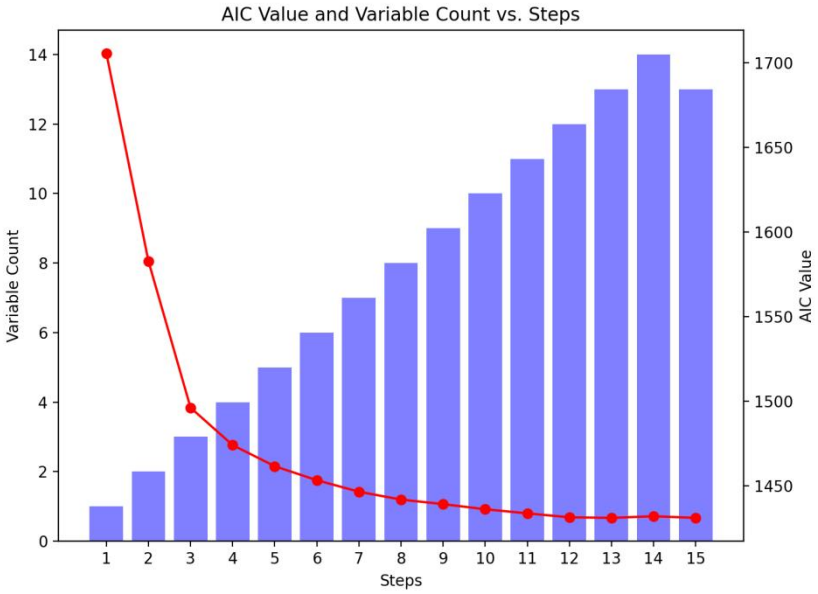

图 2 变量选择过程图

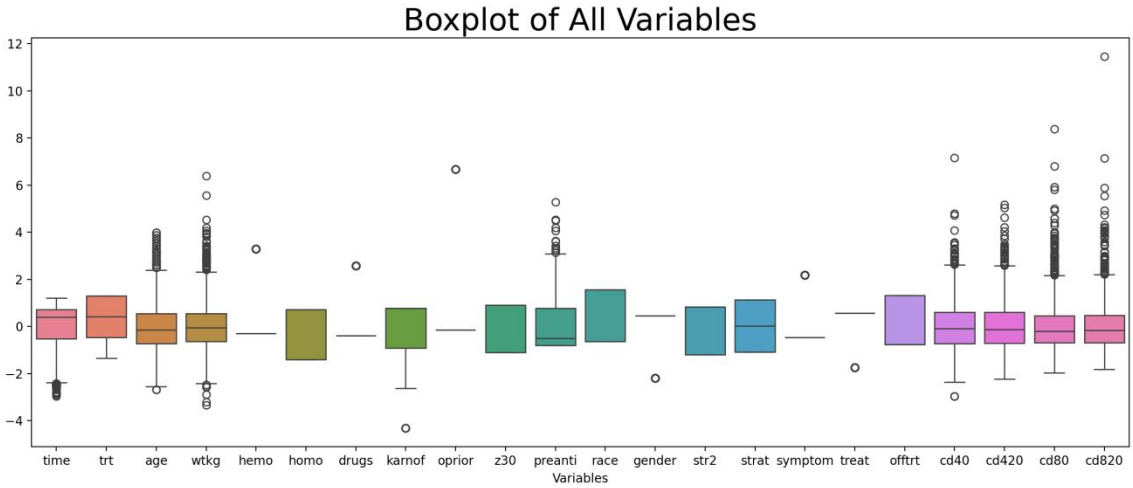

图 3 所有变量的箱线图

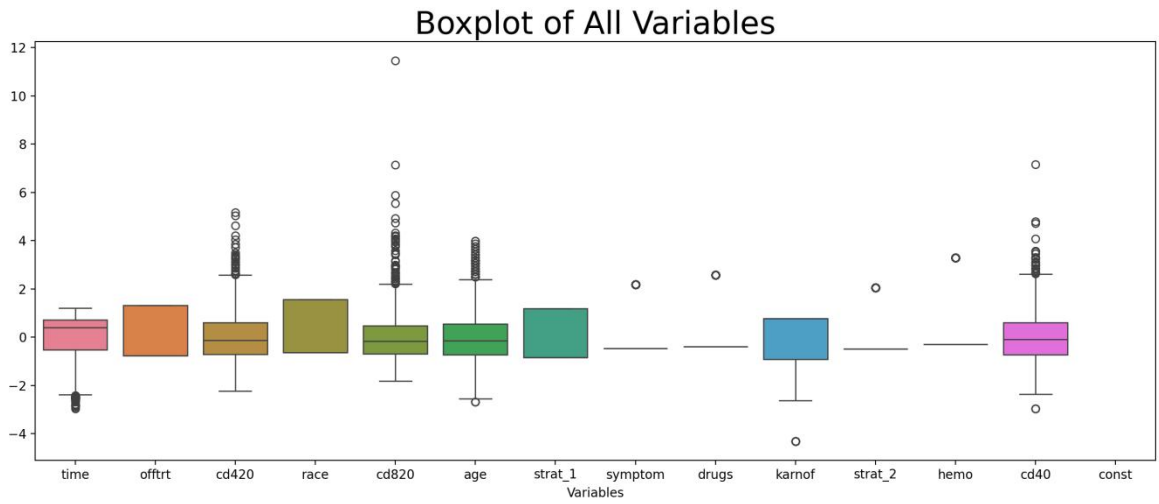

图 4 筛选出来变量的箱线图

为保持变量的量纲一致，对数据进行标准化处理，公式为

$$\widetilde{X}_{ij} = (X_{ij} - \overline{X}_j) / sd(X_j)$$ 。

其中$\widetilde{X}_{ij}$代表标准化后的数据，X_{ij}代表第 i 行第 j 列的数据，$\overline{X}_j$代表第 j 列的均值，sd 代表标准差。

四、建立解释模型

经过对数据集的处理，模型满足了样本之间相互独立、自变量之间基本不存在多重共线性等基本假设，下面使用 Python 中 statsmodels 库中的 Logit 方法对模型进行拟合，标准输入为 X 与 Y，X 为包含自变量的样本数据，Y 为对应的因变量的样本数据，使用极大使然估计法进行参数拟合，使用牛顿算法进行迭代求解。

表 5 回归系数及检验

Variance	Coef	P>|z|	Meaning
time	-1.8456	0.000	从实验开始到死亡或到达特定窗口时间的时间
offtrt	-1.8897	0.000	是否在 96±5 周前接受过 trt 治疗，是为 1，否为 0，下同
cd420	-0.7972	0.000	在实验开始后 20±5 周时的 cd8 指标
race	-0.6116	0.000	种族（是否白人）
cd820	0.2272	0.001	在前文已有解释
age	0.1529	0.029	年龄
strat_1	-0.6069	0.000	哑变量，在前文已有解释
symptom	0.4062	0.018	有无症状
drugs	-0.5166	0.020	有无使用过静脉注射药品
karnof	-0.1630	0.019	Karnofsky 功能状态评分标准，得分越高，健康状况越好，越能忍受治疗给身体带来的副作用
strat_2	-0.4315	0.025	哑变量，在前文已有解释
hemo	-0.5431	0.038	是否患有血友病
cd40	0.1366	0.123	实验开始时的 cd4 指标
const	-0.4769	0.001	常数项

求解得到模型系数以及对回归系数的 Z 检验如表 5 所示。通过各变量的 P 值可以看出在 0.05 的假设下所有变量都能够通过 Z 检验，即回归系数显著不为 0。此外，模型的对数使然函数值为-701.57，而只包含截距项的模型的对数似然函数值为-1187.5，似然比检验的 P 值接近 0，说明似然比检验通过，模型相比于只包含截距项的模型拟合能力显著提高。

从系数的正负以及大小比较上来看，时间过得越长，度过特定时间窗口的概率也就越大；在 96±5 周前接受过 trt 治疗，度过特定时间窗口的概率越大；cd4

细胞的数量上升，通常表示人体免疫系统的活性增强，度过特定时间窗口的概率越大；白人度过特定时间窗口的概率越大，可能受与贫富差距以及身体状况等其他因素影响；关于抗逆转录病毒治疗史，在 1 至 52 周内接受该治疗的人，从未接受过该治疗的人，52 周前接受过该治疗的人，他们度过特定时间窗口的概率依次减少；根据样本分布，年龄集中在 13 至 56 岁之间，可以认为在这个年龄段年龄越小，度过特定时间窗口的概率越大；无症状的样本度过特定时间窗口的概率越大；Karnofsky 评分越高，度过特定时间窗口的概率越大，以上变量的解释都是符合常理的。

而对于 cd820，其值越高死亡概率越大，可能是由于药物使用后 cd8 细胞的增加越多，代表需要消灭的 hiv 病毒越多，但是 cd8 细胞无法战胜 hiv 病毒所导致的；对于 drugs，在数据集中无清晰描述是静脉注射治疗药还是静脉注射毒品，因此在本文中不作解释；对于 hemo 变量，患有血友病的样本度过特定时间窗口的概率越大，这不符合常理，且其 p 值也是所有变量中除了 cd40 最大的一个，通不过 0.01 的检验，可以认为其回归系数可能是错误的；对于 cd40 变量，其回归系数为正，与 cd420 的相反，这可能是药物生效的结果，当然其检验 P 值为 0.123 较大，也可能是错误拟合的结果。

五、建立预测模型

尽管 Logit 回归可以从系数上对变量作出较好的解释，但如果我们只关心对一个新的样本如何使用模型进行预测其是否会在特定窗口时间内死亡，从准确率角度来说 Logit 回归可能不足以满足要求（通过模型比较也证实确实 Logit 回归的准确率不高），因此下面建立深度神经网络模型以达到更好的预测效果。

在本文中建立模型使用的是 PyTorch 框架，对于数据集，定义类 AidsDataset 继承 PyTorch 中的 Dataset 类，并重写__init__、__getitem__与__len__方法，在__init__方法中，对连续性变量进行筛选，然后进行标准化，并使用 sklearn 中的 train_test_split 将数据集划分为训练集与测试集，比例为 3:1，以及__init__方法包含参数 selected，默认值为 None，如果不作输入则将所有特征作为模型的输入，如果输入了一个列表 selected（如前文逐步回归的返回结果），则使用 selected 中包含的特征作为模型输入，并将数据以及标签转换为 PyTorch

的 FloatTensor 类型，并保存到 self.data 和 self.target 中；__getitem__方法每次返回一个样本；__len__方法返回数据集的大小，即数据集的样本数量。

然后定义函数 prep_dataloader，通过 PyTorch 的 DataLoader 类来加载网络训练所需要的数据流，其中可以定义 batch_size、使用的线程数以及是否需要 shuffle 等

接下来定义 NeuralNet 类继承 PyTorch 中的 nn.Module 类，在 NeuralNet 类的中 __init__方法中，我们可以自定义神经网络的结构，在本文中，神经网络的结构为输入维度×64 的线性层，连接一个 ReLU 层，引入非线性，帮助神经网络模型更好地学习和表示复杂的非线性关系，再连接一个 64×16 的连接层，再连接一个 16×1 得到最终结果的维度，最后连接一个 Sigmoid 层，将上一层输出压缩到[0, 1]区间用于分类，损失函数选择 nn.BCELoss 即二分类交叉熵损失函数；由于网络输出是一维的数，定义 forward 函数进行模型输出的前向传递，在函数中将张量由(batch_size, 1)的张量转化为(batch_size,)的张量；最后定义 cal_loss 函数，在该函数中可以将将 L1 或 L2 正则项作为模型参数个数的惩罚加到损失函数上，对加总的函数进行梯度下降求解系数。

接着定义 train 函数对网络进行训练，主要思想是利用优化器（Adam，SGD 等）对网络损失进行梯度下降求解，当测试集的损失下降时保存模型，最后一个被保存的模型即为训练过程中的在测试集上表现最优的模型。此外还需要定义 test 函数用于计算测试集上的损失。

至此模型的准备工作已经做完，接下来定义模型的存放路径、是否使用 gpu 加速运算以及网络训练的超参数，包括 epochs_size、batch_size、optimizer、optimizer 的参数（学习率，动量大小等）以及提早停止迭代的 early_stop 等。经过调参过程，最终选用筛选后的变量进行训练，epochs_size 为 3000，batch_size 为 300，optimizer 选用 SGD，optimizer 的学习率为 0.001，动量为 0.9，early_stop 为 500。

接下来就可以开始模型的训练，首先调用 prep_dataloader 函数得到数据流，再实例化 NeuralNet 类，然后调用 train 函数进行梯度下降更新模型系数。

模型的结果训练曲线如图 5 所示，可以看出模型基本收敛，在训练集与测试集上都有较小的损失。对于模型结果，结果大于等于 0.6 的认为属于 1 类(死亡)，

小于 0.6 的认为属于 0 类（度过窗口时间）。

除了深度神经网络外，本文还训练了 Logit 回归、决策树、支持向量机、高斯朴素贝叶斯以及 K 近邻模型，对这几个模型，训练使用的是 sklearn 库中对应的方法，特征同样为筛选后的特征，训练集与测试集与神经网络保持一致，超参数为默认参数，得到的结果如表 6 所示，其中精确率、召回值、F1 值为加权平均的结果，权重为正负样本的比例。可以看出 DNN 在各个指标上都优于传统的机器学习模型，这可能是由于深度神经网络通过多层的叠加，对样本的特征提取的更加完整，但神经网络的缺点是其训练时间花费更久，以及模型的超参数过多，需要比较好的调参才能真正得到一个比较好的模型，而传统的机器学习的超参数相对来说则少的多，可以通过交叉验证以及网格搜索等算法得到一个较好的超参数组。

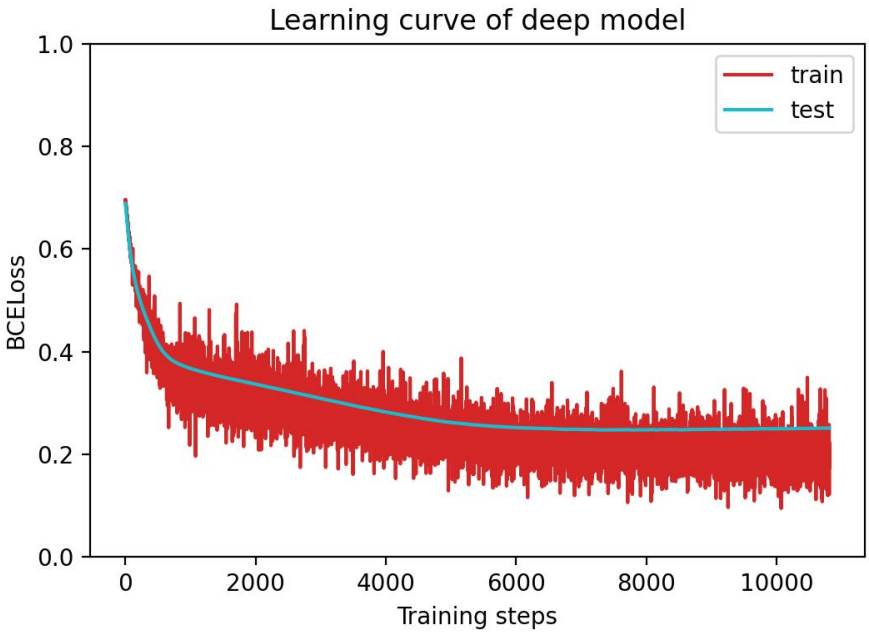

图 5 神经网络训练曲线

表 6 不同模型的结果对比

model	precision	recall	f1-score	accuracy
DNN	0.90	0.90	0.89	0.90
Logit	0.84	0.85	0.84	0.85
DecisionTree	0.78	0.79	0.79	0.84
SVM	0.88	0.89	0.88	0.83
GaussianNB	0.79	0.79	0.79	0.79
GNN	0.83	0.84	0.83	0.84

六、总结

本文以 AIDS Clinical Trials Group Study 175 数据集作为数据，通过背景介绍、数据介绍、数据处理、模型拟合、模型检验、模型解释以及模型比较等步骤探讨了关于艾滋患者是否在特定时间窗口内死亡的因素，最终得到大多数变量的解释都符合常理的结果，但也有部分变量的回归系数与常理不符，本文给出了一定的猜测。最后通过训练一个神经网络模型，以及 5 个传统的机器学习分类模型，比较它们的不同，使得作者对类的定义、函数的编写以及代码模块化等编程能力有了一定提高，且从实践角度比较深度学习与传统机器学习模型的优劣，加深了作者对这两类模型的认识。从数据的角度，通过对艾滋病数据集的探索，作者多了许多对艾滋病治疗的了解，以及希望能提出一定的数据支持说明艾滋病治疗影响因素的准确性。