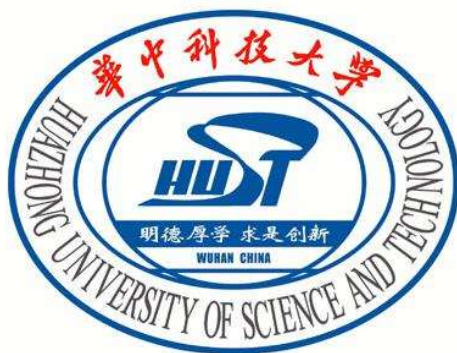


华中科技大学

计算机科学与技术学院

《机器学习》结课报告



专 业： 数据科学与大数据技术

班 级： 大数据 2201

学 号： U202215566

姓 名： 刘师言

成 绩：

指导教师： 张腾

完成日期： 2024 年 5 月 18 日

目录

1 实验要求.....	1
1.1 选题介绍.....	1
1.2 实验要求.....	1
2 算法设计与实现.....	1
2.1 数据处理.....	1
2.1.1 数据预览.....	1
2.1.2 类型转化.....	2
2.1.3 相关性热力图.....	2
2.1.4 特征构造.....	3
2.1.5 特征编码.....	3
2.1.6 归一化与标准化.....	3
2.1.7 训练集与测试集划分.....	3
2.2 模型构建.....	4
2.2.1 基于 Softmax 的逻辑回归.....	4
2.2.2 基于随机梯度下降的神经网络.....	4
2.2.3 基于信息增益划分的决策树.....	5
2.3 模型评估.....	5
2.3.1 混淆矩阵.....	6
2.3.2 准确率.....	6
2.3.3 精确率.....	6
2.3.4 召回率.....	6
2.3.5 F1 score.....	6
3 实验环境与平台.....	7
3.1 实验环境.....	7
3.2 实验平台.....	7
4 结果与分析.....	7
4.1 实验结果.....	7
4.1.1 基于 Softmax 的逻辑回归.....	7
4.1.2 基于随机梯度下降的神经网络.....	8
4.1.3 基于信息增益划分的决策树.....	9
4.1.4 Kaggle 线上性能评估.....	9
4.2 结果分析.....	9
4.2.1 训练效果分析.....	10
4.2.2 训练用时分析.....	11
4.2.3 总结.....	11
5 个人体会.....	12

1 实验要求

1.1 选题介绍

题目：肥胖风险的多分类预测

性能要求：准确率达到 80%及以上

1.2 实验要求

(1) 模型训练需自己动手实现，严禁直接调用已经封装好的各类机器学习库（包括但不限于 sklearn，功能性的可以使用，比如 sklearn.model_selection.train_test_split），但可以使用 numpy 等数学运算库（实现后，可与已有库进行对比验证）；

(2) 使用机器学习及相关知识对数据进行建模和训练，并进行相应参数调优和模型评估；

(3) 鼓励使用多种模型或不同数据集进行实验，并给出相应的分析思考；

(4) 鼓励自主拓展探索；

2 算法设计与实现

2.1 数据处理

机器学习任务的步骤主要分为数据处理、模型构建、模型训练、模型评估和模型预测等，而在构建与训练具体模型前首先要进行数据处理，从而使数据集更符合机器学习模型的格式及要求。

2.1.1 数据预览

首先，需要对数据的基本情况进行初步预览，了解数据形状、数据类型，检查是否有缺失值等等，为后续数据处理做铺垫。对数据集中各特征属性用表格进行呈现，如表 2.1 所示：

表 2.1 数据集中各特征属性表

特征	含义	类型	是否离散
id	序号	int64	是
Gender	性别	object	是
Age	年龄	float64	否
Height	身高	float64	否
Weight	体重	float64	否
family_history_with _overweight	家族史是否有超重问题	object	是

FAVC	是否食用高热量食物	object	是
FCVC	食用蔬菜频率	float64	否
NCP	一天中进餐次数	float64	否
CAEC	两餐之间进食频率	object	是
SMOKE	是否吸烟	object	是
CH2O	每日摄水量	float64	否
SCC	是否监控热量消耗	object	是
FAF	体育活动频率	float64	否
TUE	使用电子设备时长	float64	否
CALC	饮酒频率	object	是
MTRANS	乘坐的交通工具	object	是
Nobeyesdad	肥胖类别（目标）	object	是

2.1.2 类型转化

在了解完数据的基本情况后，需要对其中的非数值型数据进行类型转化。由于标签在后续预测时需要从数值重新转化成文本，因此对标签进行单独手动映射，其他文本特征则使用 `sklearn` 库的 `OrdinalEncoder` 编码器进行转化。

在转化完成后，再次对数据前若干行进行预览，确保转化过程成功无误。

2.1.3 相关性热力图

根据数据集绘制相关性热力图（如图 2.2 所示），能直观清晰地观察各特征与标签之间的相关程度，从而选择相关程度较高的特征作为主特征进行后续处理，对相关程度较低的特征考虑舍弃，进一步提高模型的准确性。

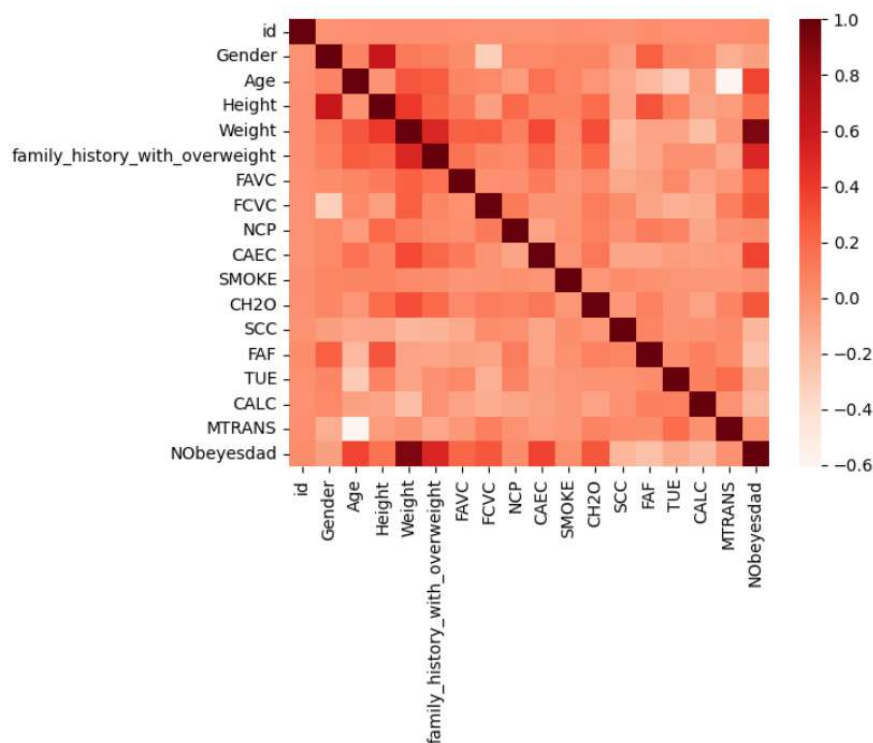


图 2.2 绘制相关性热力图

2.1.4 特征构造

通过查看数据集中各个特征的含义，结合生活常识，发现其中一些数据单独存在时和结果没有直接关联（如身高、体重等），因此，需要将这些特征进行一定地结合或者转化构造新的特征，如下所示：

（1）根据身高、体重对 BMI 进行计算，并将 BMI 作为一项新的特征加入数据集，使特征更具科学性。

（2）将 25 岁作为年龄划分界限，年龄超过 25 记为 1，反之记为 0，并将该年龄划分作为一项新的特征加入数据集，增强年龄一特征的利用价值。

在对上述特征构造完成后，再将无利用价值的特征从数据集中去除，如 id、身高、体重、年龄等，至此改造后数据集中的特征便都具有一定的参考价值。

2.1.5 特征编码

由表 2.1 可以看到，数据集中的部分特征为离散特征，当离散特征各取值仅有区分作用而无数值含义和偏序含义时，需要对其进行独热编码，从而消除距离对不同取值产生的影响。如 MTRANS 特征表示乘坐的交通工具，各取值没有数值含义和偏序含义，需要进行独热编码处理。

Gender、FAVC 等特征仅有 yes 与 no 两种不同取值，因此无需独热编码。而 CAEC 和 CALC 特征虽有多种取值，但其表示的是从事某活动的不同频率，存在一定偏序含义，因此也无需独热编码。

2.1.6 归一化与标准化

为消除各特征间量纲不同所带来的差异，对离散特征进行 Min-Max 归一化处理，而由于 Min-Max 归一化受离群值影响较大，对连续特征则进行 Z-score 标准化处理。

至此，数据预处理的所有工作就已基本完成，再次预览处理后的数据集，如图 2.3 所示：

Gender	family_history_with_overweight	FAVC	FCVC	NCP	CAEC	SMOKE	CH2O	SCC	FAF	TUE	CALC	NObeyesdad	BMI
0	1	1	-0.836259	0.314676	0.666667	0	1.206565	0	-1.171113	0.597424	0.5	3	-0.237856
1	0	1	-0.836259	0.338356	0.333333	0	-0.048348	0	0.021774	0.636498	1.0	1	-0.818311
2	0	1	-1.060306	-1.913377	0.666667	0	-0.195640	0	-0.138019	1.755197	1.0	0	-1.573703
3	0	1	1.039146	0.338356	0.666667	0	-0.584021	0	0.579882	0.271448	0.5	6	1.753549
4	1	1	0.438386	-1.119774	0.666667	0	-0.081467	0	1.176457	0.523099	0.5	3	-0.557083

CAEC	SMOKE	CH2O	SCC	FAF	TUE	CALC	NObeyesdad	BMI	Age_up_25	MTRANS_0	MTRANS_1	MTRANS_2	MTRANS_3	MTRANS_4
566667	0	1.206565	0	-1.171113	0.597424	0.5	3	-0.237856	0	0	0	0	1	0
333333	0	-0.048348	0	0.021774	0.636498	1.0	1	-0.818311	0	1	0	0	0	0
566667	0	-0.195640	0	-0.138019	1.755197	1.0	0	-1.573703	0	0	0	0	1	0
566667	0	-0.584021	0	0.579882	0.271448	0.5	6	1.753549	0	0	0	0	1	0
566667	0	-0.081467	0	1.176457	0.523099	0.5	3	-0.557083	1	0	0	0	1	0

图 2.3 预处理后的数据集

2.1.7 训练集与测试集划分

在正式进入机器学习模型构建前，需要对数据集按 7:3 的比例划分训练集与测试集，以便更好地对模型进行评估。划分后的样本不含标签，而标签则是各样本对应的肥胖类别。接下来便可开始机器学习的模型构建。

2.2 模型构建

2.2.1 基于 Softmax 的逻辑回归

逻辑回归是众多模型中实现简单且性能较好的一种机器学习模型。其中 **Softmax 回归** 是逻辑回归在多分类问题上的推广，此时逻辑回归的损失函数由原来的 Sigmoid 函数修改为 Softmax 函数。Softmax 回归的模型表达式如下：

$$P = \text{Softmax}(Wx + B)$$

其中 W 和 B 分别为模型对各特征的权重和偏置，输出 P 的元素个数与类别数相同， P 中各个值为样本属于对应类别的概率，概率值最高的类别即为对该样本的预测类别。

模型参数 W 和 B 主要采用梯度下降法进行迭代求解，迭代过程中每遍历一条样本便对参数进行计算更新，当迭代超过一定次数时，模型基本拟合，标志着训练的结束。Softmax 回归的基本实现原理如图 2.4 所示：

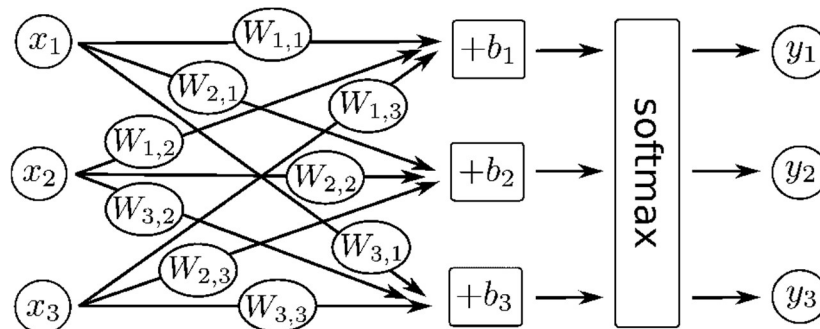


图 2.4 Softmax 回归基本原理图

2.2.2 基于随机梯度下降的神经网络

神经网络模型在逻辑回归模型的基础上加入了隐藏层，通过将线性分类器进行组合叠加，能够较好地进行非线性分类，因此相比逻辑回归更适合作为此课题的机器学习模型。

神经网络的主要步骤包括**前向传播**与**反向传播**。前向传播即从输入层到输出层，计算每一层各神经元的激活值。反向传播即根据前向传播计算出来的激活值，计算每一层参数的梯度，并从后往前进行参数的更新，主要采取的仍是梯度下降法。神经网络的基本实现原理如图 2.5 所示：

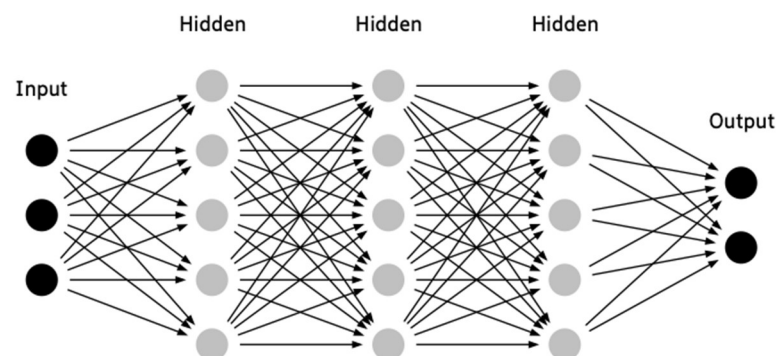


图 2.5 神经网络基本原理图

相比先前的逻辑回归模型，神经网络模型除新加入了隐藏层外，还分别在梯度下降和数据遍历上作了一定的优化：

a) **随机梯度下降法(SGD)**：每次从样本中随机抽出一组，训练后按梯度更新一次参数，然后再抽取一组，再更新一次。在样本量较大的情况下，可以不用训练完所有样本便获得一个损失值在可接受范围之内的模型；

b) **小批量训练(Mini-Batch)**：遍历数据集时，从样本中采样一批样本进行小批量训练，将这些批次的梯度视为真实梯度的近似值。小批量训练不需要完全遍历训练数据来更新参数，因此倾向于更快地收敛。

2.2.3 基于信息增益划分的决策树

决策树是根据数据特征构建一个树状结构，其中每个分支节点代表多个备选方案之间的选择，每个叶节点代表一个决策。程序设计中的条件分支结构(if-then)便是决策树思想的体现，决策树正是利用这类结构分隔数据的一种学习分类模型。决策树的基本实现原理如图 2.6 所示：

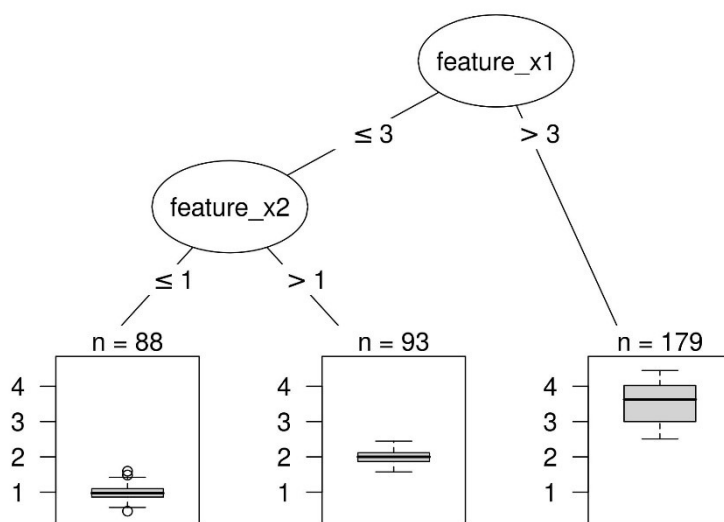


图 2.6 决策树基本原理图

各特征的**信息熵**是构建决策树的关键影响因素。某个特征的信息熵越大，该特征包含的信息越多，纯度越低；信息熵越小，该特征包含的信息越少，纯度越高。**信息增益**则表示根据某个特征划分前后的纯度差别，即信息熵的减小量，其公式为：

$$Gain(D) = Ent(D) - \sum_{v=1}^V \frac{nD^v}{nD} Ent(D = A_v)$$

在构建决策树时，优先选择信息增益值大的特征进行划分。循环信息增益计算与特征选择划分步骤，直至达到终止条件，形成最终的决策树模型。

2.3 模型评估

在机器学习中，对模型性能进行评估非常重要，选择适当的评估指标是保证模型成功的关键之一。当完成模型构建和模型训练工作后，需通过准确率

(Accuracy)、精确率 (Precision)、召回率 (Recall)、F1 等多项指标对模型进行全面评估，以全方位地了解模型性能好坏，为后续改进优化工作提供数据支撑。

2.3.1 混淆矩阵

混淆矩阵 (Confusion Matrix) 以矩阵的形式展示了真实类别与模型预测类别之间的关系，该不同关系 (如图 4.2 所示) 是各指标对模型进行评估的基础。

		真实标签	
		1	0
预测标签	1	TP	FP
	0	FN	TN

图 4.2 混淆矩阵中的不同关系

2.3.2 准确率

准确率是模型预测结果中正确结果的占比，是最简单直接的一个衡量指标。表达式为：

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

2.3.3 精确率

精确率是在所有预测为 1 的样本中，真实标签也为 1 的样本的占比。精确率又叫查准率，衡量模型对预测正样本的准确程度。精确率越高，说明在被预测为正的样本中，真实标签也为正的概率越大。表达式为：

$$Precision = \frac{TP}{TP + FP}$$

2.3.4 召回率

召回率是在所有真实标签为 1 的样本中，模型预测标签也为 1 的占比。召回率又叫查全率，衡量模型捞出正样本的能力，召回率越高，说明真实标签为正的样本，被预测为正的概率越大。表达式为：

$$Recall = \frac{TP}{TP + FN}$$

2.3.5 F1 score

F1 score 为精确率和召回率的综合，是两者的调和平均，能反映模型查的又准又全的能力。精确率越大，F1 越大；召回率越大，F1 越大。表达式为：

$$F1 = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}}$$

除了从上述几个指标对模型进行评估外，还可以记录模型在训练时平均花费的时间，并从训练时间、模型性能等多维度对不同模型进行比较和总结，从而更好地认识不同模型的特点与优劣，进而选用效果最佳的模型进一步改进优化。

3 实验环境与平台

3.1 实验环境

实验环境配置如表 3.1 所示：

表 3.1 实验环境配置

名称	配置信息
操作系统	Windows 11
开发语言	Python 3.9.12
CPU	Intel(R) Core(TM) i7-12700H 2.30 GHz
GPU	GeForce RTX 3060(6G)
内存	16G

3.2 实验平台

实验平台配置为：Jupyter Notebook 7.1.1

4 结果与分析

4.1 实验结果

4.1.1 基于 Softmax 的逻辑回归

初始化逻辑回归模型时，设置最大迭代次数为 300，学习率为 0.005。训练过程使用 time 模块对用时进行记录，结果如图 4.1 所示：



```
round 180, accuracy 84.89%
round 200, accuracy 84.88%
round 220, accuracy 84.90%
round 240, accuracy 84.91%
round 260, accuracy 84.92%
round 280, accuracy 84.94%
训练用时 45.26s
训练集准确率：84.93%
测试集准确率：84.51%
```

图 4.1 逻辑回归模型训练过程

图上 accuracy 表示训练过程中模型在训练集上的准确率。可以看到，在 200 轮左右时模型已接近收敛，说明模型构造基本完成。

训练完毕后对测试集进行预测。首先，查看由模型预测结果形成的混淆矩阵

(如图 4.3 所示)，以评估模型对于各个分类的预测能力：

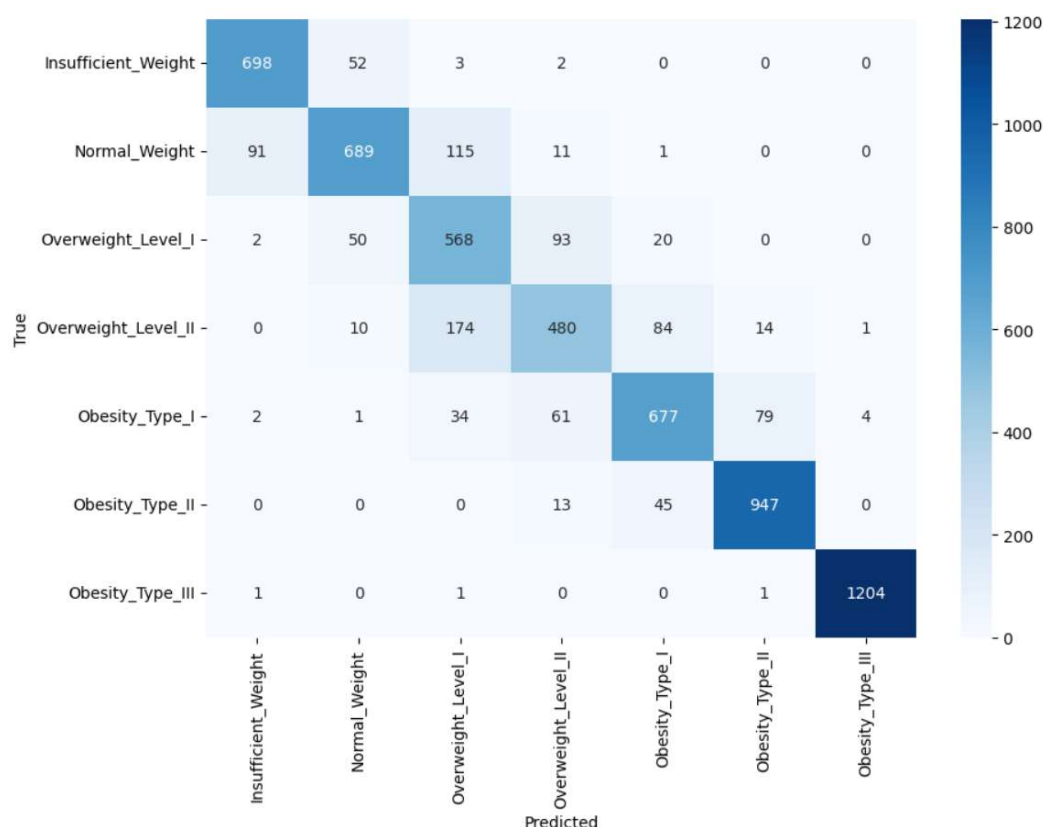


图 4.3 逻辑回归模型预测结果的混淆矩阵

由混淆矩阵可以看出，逻辑回归模型对于 **Obesity_Type_III** 类别的预测十分精确，出错样本占极少数，准确率高达 99.8%；而对于 **Overweight_Level_II** 类别的预测较为一般，准确率仅有 62.9%。

可能的原因是逻辑回归模型是一种线性分类模型，模型对各个类别的预测能力与类别本身特征鲜明与否密切相关。在本任务中，**Obesity_Type_III** 属于最高级别类别，特征较为突出，容易与其他类别区分开来；而 **Overweight_Level_II** 属于中间级别类别，特征与相邻类别区别不明显，预测时容易出现混淆。

根据预测结果，得到逻辑回归模型在测试集上各指标的性能，并与 sklearn 的逻辑回归模型进行比较，如表 4.1 所示：

表 4.1 逻辑回归模型性能

指标	MyLogisticRegression	sklean.LogisticRegression
准确率	84.51%	84.71%
F1 score	82.94%	83.12%
训练用时(s)	45.26	0.84

4.1.2 基于随机梯度下降的神经网络

初始化神经网络模型时，设置最大迭代次数为 300，学习率为 0.001，隐藏层层数为 30，小批量样本数为 5。训练过程使用 time 模块对用时进行记录。

训练完毕后对测试集进行预测。根据预测结果，得到神经网络模型在测试集上各指标的性能，并与 sklearn 的神经网络模型进行比较，如表 4.2 所示：

表 4.2 神经网络模型性能

指标	MyNeuralNetwork	sklean.NeuralNetwork
准确率	85.71%	85.79%
F1 score	84.20%	84.25%
训练用时(s)	18.29	20.50

4.1.3 基于信息增益划分的决策树

初始化决策树模型时，分别设置生成决策树的最大深度为 3、4 和 5，发现在最大深度为 3、5 时，模型在训练集上的预测准确率均较低。

原因可能是最大深度为 3 时模型欠拟合，即模型过于简单，无法捕捉到数据中的复杂模式；而最大深度为 5 时模型过拟合，即模型过于复杂，对训练数据进行了“记忆”而非“学习”。因此最终选取最大深度为 4 作为模型参数。训练过程使用 time 模块对用时进行记录。

训练完毕后对测试集进行预测。根据预测结果，得到决策树模型在测试集上各指标的性能，并与 sklearn 的决策树模型进行比较，如表 4.3 所示：

表 4.3 决策树模型性能

指标	MyDecisionTree	sklean.DecisionTree
准确率	83.85%	84.84%
F1 score	82.21%	83.44%
训练用时(s)	20.12	0.03

4.1.4 Kaggle 线上性能评估

分别使用三个模型对测试用数据进行预测，并将预测结果上传至 Kaggle 测评，得到各模型在线上的测评结果如图 4.4 所示：




	DecisionTree.csv Complete (after deadline) · 15m ago	0.83372	0.82839
	NeuralNetwork.csv Complete (after deadline) · 24m ago	0.86298	0.86235
	LogisticRegression.csv Complete (after deadline) · 25m ago	0.84799	0.84609

图 4.4 Kaggle 线上测评结果

从测评结果可以看到，三个模型的准确率均满足题目要求的 80%，其中神经网络模型的预测效果最好，准确率达 86.3%，远超预期目标。表明本次实验模型构建与训练十分成功。

4.2 结果分析

根据实验结果,对三种模型在各指标上的性能进行汇总对比,如表 4.4 所示:

表 4.4 各模型性能对比

指标	逻辑回归	神经网络	决策树
本地准确率	84.51%	85.71%	83.85%
线上准确率	84.80%	86.30%	83.37%
F1 score	82.94%	84.20%	82.21%
训练用时(s)	45.26	18.29	20.12

将上述三种模型性能以折线图的形式呈现,如图 4.5 所示:

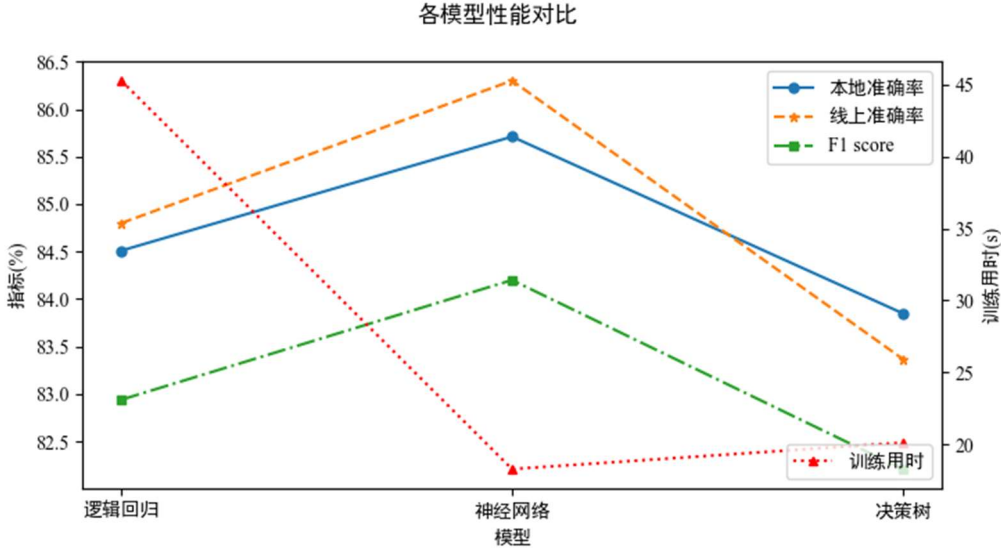


图 4.5 各模型性能对比

从折线图可以直观清晰地看出各模型之间的性能差异和训练用时差别。其中,神经网络模型的准确率最高,其次分别是逻辑回归模型和决策树模型;神经网络模型的训练用时最短,其次分别是决策树模型和逻辑回归模型。

4.2.1 训练效果分析

由于神经网络模型和逻辑回归模型都是通过梯度下降法对模型参数进行求解,具有一定的相似性。

而神经网络模型隐藏层的加入,使得线性分类器能够组合叠加,可较好地进行非线性分类(如图 4.6 所示),因此训练效果相较一般的线性逻辑回归模型更优。此外,相比于一般梯度下降法,神经网络模型采用的随机梯度下降法可以避免陷入局部最优解,训练结果更准确。

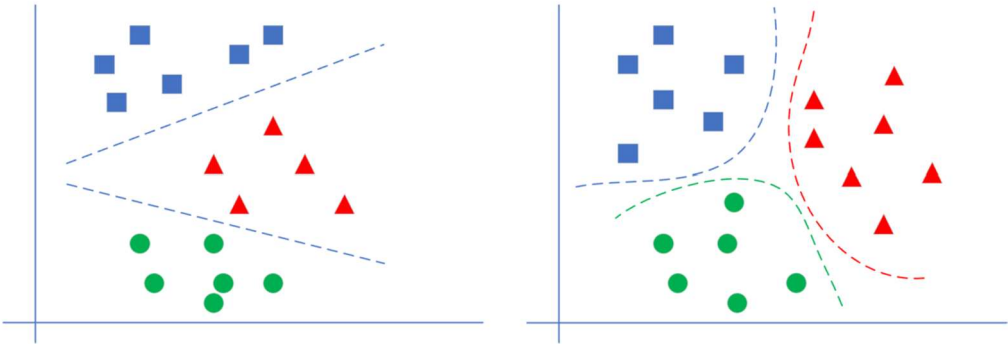


图 4.6 线性分类与非线性分类

不同于上述两种模型，决策树采用树状结构来构建模型，从而是一种非线性分类模型。决策树容易构造出易于理解和解释的规则，因此模型训练结果较为准确，但仍与前面两种模型有一定的差距。这主要是由决策树忽略数据集中属性之间的相关性导致的。

同时，决策树倾向于生成复杂的模型，容易过拟合训练数据，导致模型性能下降。需采取限制决策树最大深度、剪枝等方法来避免过拟合现象的产生，但这一定程度上又使模型构建不能达到最佳效果。因此对于诸如本任务的复杂关系，决策树的表现不如逻辑回归模型和神经网络模型优秀。

由图还可以看出，决策树自身在线上准确率相比本地有所下降，这也受到了模型欠拟合或过拟合的影响，表明决策树的稳定性较差，数据中微小的变化可能导致生成完全不同的树，不能很好地推广数据。

三种模型的准确率均达到了 80% 以上，且各模型 F1 分数与自身准确率的差距也较小，表明在类别不平衡的情况下，三种分类器在精确率和召回率之间依然取得了良好的性能。通过对 F1 指标的测评可有效避免准确率失效所带来的局限性与盲目性，从而更全面地对模型性能进行评估。

4.2.2 训练用时分析

相比逻辑回归模型，自己实现的神经网络模型训练用时缩短了超过 50%，这是由于神经网络模型运用的随机梯度下降法和小批量训练可以不用训练完所有样本就获得一个损失值在可接受范围内的模型，从而显著提高训练效率，大幅缩短训练用时。

而决策树的训练用时也相对较短，这是由于决策树在生成过程中限制了树的最大深度，同时对节点进行了剪枝，能指数级减少总节点个数，从而缩短训练用时。

4.2.3 总结

综合训练效果和训练用时两个层面，神经网络模型是本次任务的最佳分类模型，其较高的预测准确率、良好的 F1 分数和较短的训练用时确保能圆满完成预测任务，并且稳定性绝佳，在不同测试集上的表现效果均较好。

下面对三种模型各自的特点及优缺点进行对比总结，如表 4.5 所示：

表 4.5 各模型特点及优缺点

	逻辑回归	神经网络	决策树
分类器类型	线性	非线性	非线性
训练效果	较好	很好	较好
训练用时	较短	短	短
实现难度	容易	较难	中等
优点	模型简单，易于实现	具有高速寻找最优化解的能力	规则易于理解和解释
缺点	对于复杂任务准确率不高	难以解释推理过程和推理依据	容易欠拟合或过拟合，不稳定
应用场景	构造的特征线性可分，构造的特征基本线性相关	数据量庞大，参数之间存在内在联系的复杂问题	针对非数值型数据，一般用作其他模型的基础

5 个人体会

在学习《机器学习》这门课程之前，我一直认为机器学习是一个高深莫测的概念，不能通过科学有效的方法对其进行理解和解释。然而，当亲自动手从零开始一步步实现不同的机器学习模型时，原本看似神秘无比的算法在缜密的设计实现下有条不紊地运行起来，让我深谙机器学习的精妙与伟大。

通过对本次肥胖风险的多分类预测一任务的实践，从最初的数据分析、数据处理，到深层次的算法设计、模型构建，再到一锤定音的模型训练与模型预测，一路走来耗费了不少的时间与心血。但当看到模型训练时一轮轮迭代，准确率一点点提升，看到模型预测时准确率从起初的 70%慢慢优化改进到 80%甚至接近 90%时，心情无比激动而喜悦，感到一切的付出都是值得的。

可以说本次任务实践第一次打开了我计算机学习之旅中机器学习的大门。通过对不同模型的探索与构建、相应算法的学习与延伸，机器学习的原理和实现流程渐渐变得熟悉起来。经过一次次的算法设计、框架搭建、Debug 和改进优化，尽管自行实现的算法模型在本次任务的训练用时上与 `sklearn` 的模型相比仍望尘莫及，但预测效果已经与 `sklearn` 的模型相差无几，令我感到无比欣慰与满足，成就感十足。

然而，完成任务的过程并非一帆风顺。期间，除本次报告所选择的三种模型以外，我还尝试了诸如支持向量机、KNN 等其他模型，但许多以难度较大，编写不成功告终，有的则是由于训练效果较差而被舍弃。即便确立了本次采用的三种模型，在实现与优化的过程中也遭遇了大大小小的问题，比如训练用时过长、训练参数难以确定等等，好在通过查阅资料和询问老师同学都迎刃而解。最终经过不懈努力，圆满完成了三种模型的设计与实现，也进行了全面的模型评估与不同模型的横纵对比，可以说收获颇丰。

尽管机器学习的课程任务已经收官，但机器学习和计算机学习的道路仍在不断延伸，一路上还会遇到新的坎坷，也能收获惊喜与成长。希望在今后的学习生活中永葆持之以恒的探索精神，遇到陌生的问题与挑战时能临危不乱，耐心且细心地学习新的知识、新的方法，积少成多，久久为功。