安徽工业大学
ANHUI UNIVERSITY OF TECHNOLOGY

2022 - 2023 学年第 1 学期

机器学习课程设计报告

题　目：	基于感知机的糖尿病数据分析
姓　名：	张建才
学　院：	计算机科学与技术学院
专　业：	人工智能
学　号：	209074458
班　级：	智能 202
日　期：	2022 年 12 月 25 日

评　分：

声　明

本人郑重声明，所提交的课程设计由本人独立完成，保证不存在剽窃、抄袭他人成果的现象，文中所引用他人观点、材料、数据，图表等文献资料均以注释说明来源。

课程设计作者（签名）：张建才

2022 年 12 月 25 日

目　录

摘　要 4
一、课设目的与思路 5
二、模型算法 6
1. 模型简介 6
2. 评价指标 6
三、实验结果与分析 7
1. 实验环境 7
2. 数据收集 7
3. 数据处理 7
4. 模型训练 7
5. 模型优化 7
6. 模型测试与评估 7
7. 模型 UI 设计（可选） 8
四、结语 9
参考文献 10
附录：相关代码 11

摘　要

感知机；糖尿病；感知机参数；热力图；数据处理想法；

一、课设目的与思路

目的：对糖尿病数据集(diabetes.csv)分析多维度数据并通过热力图辅助理解数据间关系与感知机参数的影响；

思路：使用 sklearn 库的 Perception 函数对数据集全部特征（维度）的二分；输出较高准确率的各个特征值的权重，并与热力图和现实情况进行定性对比。

二、模型算法

1.模型简介

感知机是一个二分类线性模型，其输入的是特征向量，输出的是类别。
算法：
Perceptron：
1 初始化参数
2 对所有数据进行判断，超平面是否可以把正实例点和负实例点完成正确分开。
3 如果不行，更新 w，b。
4 重复执行 2，3 步，直到数据被分开，或者迭代次数到达上限。

2.评价指标

AUC 的值就是处于 ROC curve 下方的那部分面积的大小；
对于 ROC：平面的横坐标是 false positive rate(FPR)，纵坐标是 true positive rate(TPR)：TPR 代表能将正例分对的概率，FPR 代表将负例错分为正例的概率。

TPR=TP/(TP+FN)
FPR=FP/(FP+TN)

AUC 值为 ROC 曲线所覆盖的区域面积，显然，AUC 越大，分类器分类效果越好。
AUC = 1，是完美分类器，采用这个预测模型时，不管设定什么阈值都能得出完美预测。绝大多数预测的场合，不存在完美分类器。
0.5 < AUC < 1，优于随机猜测。这个分类器（模型）妥善设定阈值的话，能有预测价值。
AUC = 0.5，跟随机猜测一样（例：丢铜板），模型没有预测价值。
AUC < 0.5，比随机猜测还差；但只要总是反预测而行，就优于随机猜测。

三、实验结果与分析

1. 实验环境

硬件环境：CPU
软件环境：python3.11;sklearn1.2;×（包括硬件环境和软件环境）

2. 数据收集

https://www.kaggle.com/datasets/saurabh00007/diabetescsv

3. 数据处理

把数据切分为特征 X 和标签 y

切分数据集

切分后训练集和测试集中的数据类型的比例跟切分前 y 中的比例一致

4. 模型训练

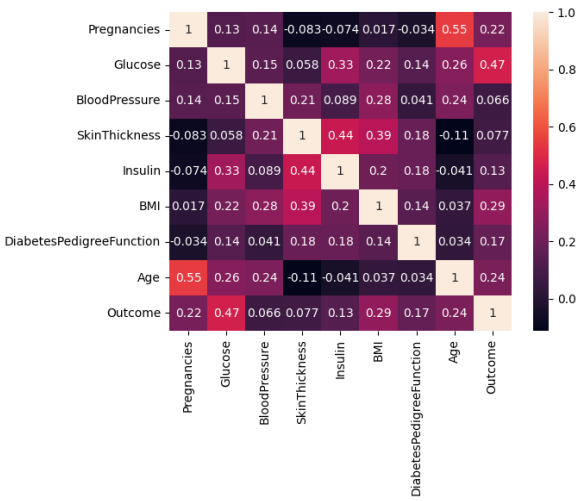
确定训练次数，添加惩罚项，每次数据输入时随机。

5. 模型优化

因为特征比较多，维度较高，增加了混合惩罚项，同时，数据数量较少，对损失函数的要求比较高。所以一般以训练次数为定值。

6. 模型测试与评估

```
[[159.74780099 238.25842101   0.         95.67187516   0.
  328.22316746  19.56645187  98.78537853]]
[-49847.]
```



可以观察到：Glucose 与 Outcome 的正相关系数比较大，同时它的权重也是比较大为 238；同时注意到 BMI 的权重比较大，说明这是更为准确的判断。

7. 模型 UI 设计（可选）

无

四、结语

整合课程设计使用的比较简单，同时也是最基础的感知机，对数据并没有过多处理，如:K 折交叉验证；如降维：只取一部分的特征，或是合并几个特征，或是减低数据的类别；如升维：将几个特征合并成一个新的特征；寻找他们之间的现实意义的相关性。

训练也是极其简单：并未自造车轮，调用 sklearn 库，方便不少，从复杂的程序语言中解脱出来。

模型评价却是比较复杂，在这个感知机中应该是不可能实现的，感知机是使用超平面划分数据的。也就是说，他输出的结果只有是或是否，所以 AUX 并不知道怎么使用。

参考文献

数据集：https://www.kaggle.com/datasets/saurabh00007/diabetescsv

思路：https://blog.csdn.net/hqllqh/article/details/108932368

AUX，ROC：

https://blog.csdn.net/weixin_44830815/article/details/105539747

附录：相关代码

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.linear_model import Perceptron
#loal data
diabetes_data = pd.read_csv(r'.\diabetes.csv')
plt.figure()
sns.heatmap(diabetes_data.corr(), annot=True)
#导入测试数据
from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test=train_test_split(diabetes_data.iloc[:,d
iabetes_data.columns !='Outcome'],diabetes_data['Outcome'],stratify=d
iabetes_data['Outcome'],random_state=66)
clf =
Perceptron(penalty='elasticnet',alpha=0.001 ,tol=None,fit_intercept=T
rue,max_iter=100000,shuffle=True,n_jobs=-1)
clf.fit(X_train, y_train)
print(clf.coef_)
print(clf.intercept_)
acc = clf.score(X_test,y_test)      # 使用测试集进行验证
print(acc)
```