实验一

数据集描述

本次实验使用的数据集由助教提供，该数据集已经分割为**训练集**、**验证集**与**测试集**三部分，其中分别有700,000、200,000与100,000个样本。其中每个样本有1个编号(id)，285维特征(feature[0,284])与1个二分类标签(target→(0,1))，即三个数据集中的第0列数据为`id`，第1到285列分别为`feature0`到`feature284`，最后一列为`target`。三个数据集的路径名称如下：

- 训练集：`Lab1_train.csv`
- 验证集：`Lab1_validation.csv`
- 测试集：`Lab1_test.csv`

实验要求

基本实验流程： 实现一个模型，然后在 **训练集** 上训练（根据训练样本的误差优化参数），每轮迭代后在 **验证集** 上验证，根据验证样本的误差决定是否停止训练、是否需要修改模型结构等超参数，如果需要修改超参数则修改模型后重新开始训练。最终获得一个最优超参数且完成训练的模型，用于预测 **测试集** 的结果。

具体要求：

1. 遵照助教划分的训练集、验证集、测试集，只在训练集上训练。
2. 不限制编程语言。
3. 使用多层感知机（全连接网络），分别测试只包含2层、3层、4层神经元的模型（即3个模型），自行调整其他超参数/参数（例如每层神经元数量、使用什么激活函数、是否使用 dropout 或 bias、epoch 数等属于超参数）；
4. 使用 SVM 模型，自行调整其他超参数/参数；
5. 根据**验证集**的 **F1 score** 选择上述4个模型的超参数，**重点**记录选择过程和选择理由；
6. 分析最终4个模型在验证集上的拟合情况，例如不同模型的过拟合、欠拟合程度等；
7. 分析4个模型在**测试集**的 F1 score、 ROC 曲线及 AUC（比如是否符合验证集的预期，是否观察到其他现象）；

实验报告

需要包括：

1. 简要说明自己使用的实验环境（机器设备情况、所用语言和库）、对实验数据的预处理过程（如果有）和读取方式；

2. 记录上述4个模型的训练过程，**重点**说明超参数选择过程和选择理由，列表说明使用过的超参数及对应实验结果；
3. 对比最终得到的4个模型在验证集上的拟合情况（过拟合/欠拟合程度），总结各模型的特点；
4. 分别记录最终4个模型在测试集的 F1 score、 ROC 曲线及 AUC，简要分析。

提交

将 **实验报告、源码**（不包括模型）打包成一个压缩文件，命名为 **学号+姓名+lab1** 提交到 www.bb.ustc.edu.cn。
截止时间：4月5日23:59。延期分数*80%。