模式识别与机器学习大作业

提交日期 6 月 22 日（考试后一周） 23:55 提交

在下列 4 类机器学习算法中，每类选一种算法：

- 线性方法: 线性 SVM、对数几率回归
- 非线性方法: Kernel SVM，决策树
- 集成学习：Bagging、Boosting
- 神经网络：自定义合适的网络结构

（1）数据集：下面两个数据集任选一个，或者采用自己科研中的数据集

（2）人员：最多三人一组

（3）编程：编程语言不限，可以调用工具包

（4）提交：程序提交原代码；实验报告以 docx、pdf 类型给出，说明每人的分工、数据特点、不同模型的性能比较及其原因；无需包含数据文件，所有文件打包成一个压缩包。

1. 结构型数据：电信用户流失数据集或者自己科研相关的数据集

电信用户流失数据集 WA_Fn-UseC_-Telco-Customer-Churn.csv，7043 个样本、每个样本 21 维特征。

https://www.kaggle.com/datasets/blastchar/telco-customer-churn?resource=download

特征说明如下：

变量	类型	说明
customerID	字符串	客户 ID
gender	字符串	性别
SeniorCitizen	数值型	是否老年人（是：1；否：0）
Partner	字符串	是否有伴侣（是：Yes；否：No）
Dependents	字符串	是否有需要抚养的孩子（是：Yes；否：No）
tenure	字符串	使用公司服务的月份数（0-72 之间）
PhoneService	字符串	是否办理电话服务（是：Yes；否：No）
MultipleLines	字符串	是否开通多条线路（是：Yes；否：No）
InternetService	字符串	是否开通网络服务和开通的服务类

		型 数字用户线路：DSL 光纤线：Fiber optic 未办理网络服务：No
OnlineSecurity	字符串	是否使用网络安全服务 是：Yes 否：No 未开通网络：No internet service
OnlineBackup	字符串	是否使用网络备份功能 是：Yes 否：No 未开通网络：No internet service
DeviceProtection	字符串	是否开启设备保护 是：Yes 否：No 未开通网络：No internet service
TechSupport	字符串	是否使用技术支持功能 是：Yes 否：No 未开通网络：No internet service
SteamingTV	字符串	客户是否办理数字电视功能 是：Yes； 否：No 未开通网络：No internet service
SteamingMovies	字符串	客户是否办理数字电影功能 是：Yes； 否：No 未开通网络：No internet service
Contract	字符串	客户的合约方式 每月签约：Month-to-month 一年：One year 两年：Two year

PaperlessBilling	字符串	是否使用无纸化账单 是：Yes；否：No
PaymentMethod	字符串	付款方式： 电子支票：Electronic check； 邮寄支票：Mailed check； 银行自动转账：Bank transfer（automatic）； 信用卡自动扣款：Credit card（automatic）
MonthlyuCharges	数值型	每月支出
TotalCharges	数值型	总支出
Churn	字符串	客户是否流失（已流失：Yes；未流失：No）

a) 性能评价指标：分类正确率

b) 编程：推荐使用 sklearn 工具包

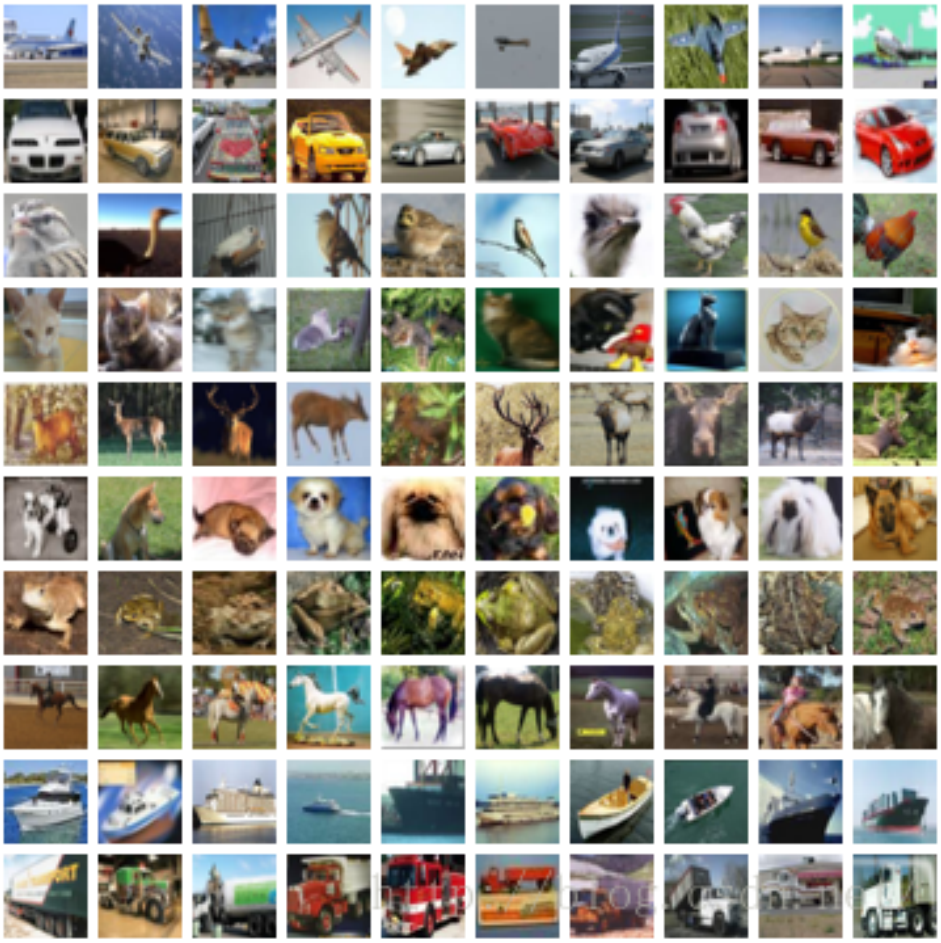
- 数据探索性分析：分析数据特点
- 数据预处理：数据缺失值处理、特征编码、特征选择、特征归一化/标准化等，注意不同的模型需要的数据预处理方式不完全相同
- 模型训练：寻找最佳模型超参数和参数：80%数据作为训练集，5 折交叉验证
- 模型评估:所选模型在剩余 20%的测试数据上的性能(分类正确率）

2. 非结构型数据：图片分类（文件名 cifar-10-python.tar.gz）

CIFAR-10 数据集是由 Alex Krizhevsky, Vinod Nair 和 Geoffrey Hinton 收集的一个用于识别普适物体的小型数据集。共包含 10 个类别的 RGB 彩色图片，如飞机、汽车、鸟类、猫、鹿、狗、蛙类、马、船和卡车。

CIFAR-10 数据集包含 60000 张32×32的彩色图像，每类包含 6000 张图片，其中 50000 张作为训练集，10000 张作为测试集。

CIFAR-10 数据集的每张图片是以被展开的形式存储，每一类的数据表示为 uint8，前 1024 个数据表示红色通道，接下来的 1024 个数据表示绿色通道，最后的 1024 个通道表示蓝色通道 3。



注意：

对传统机器学习方法，特征可以采用 PCA 提取，或者自行设计更好的特征提取方式；该数据集较大，如果要采用 kernel SVM 或 SVM 需要注意。

对神经网络方法，请自行设计合适的网络结构，采用端到端的方式实现图片分类。

性能评级指标为样本的分类正确率。