实践作业

- 从四类方法中选三类方法，从选定的每类方法中，各选一种具体的方法，从给定的数据集中选一个数据集对这三种方法进行测试比较。

  第一类方法：：线性方法：线性SVM、 Logistic Regression
  第二类方法: 非线性方法：Kernel SVM， 决策树
  第三类方法: 集成学习：Bagging， Boosting
  第四类方法: 神经元网络：自选结构

- 提交内容：附件形式打包提交，命名“学号-姓名-作业三.zip”，包括：1. 实验报告 2.可运行代码。

应用作业-数据集

- 手写数字识别
- Mnist:训练集 (training set)

  由来自 250 个不同人手写的数字构成，其中 50% 是高中学生，50% 来自人口普查的工作人员. 测试集(test set) 也是同样比例的手写数字数据60,000 个训练样本，10000个测试数据
- 每个数字为 28*28的图像，即7[illegible]维的向量

$$p(y = j|x_1, x_2, \ldots, x_{784})$$

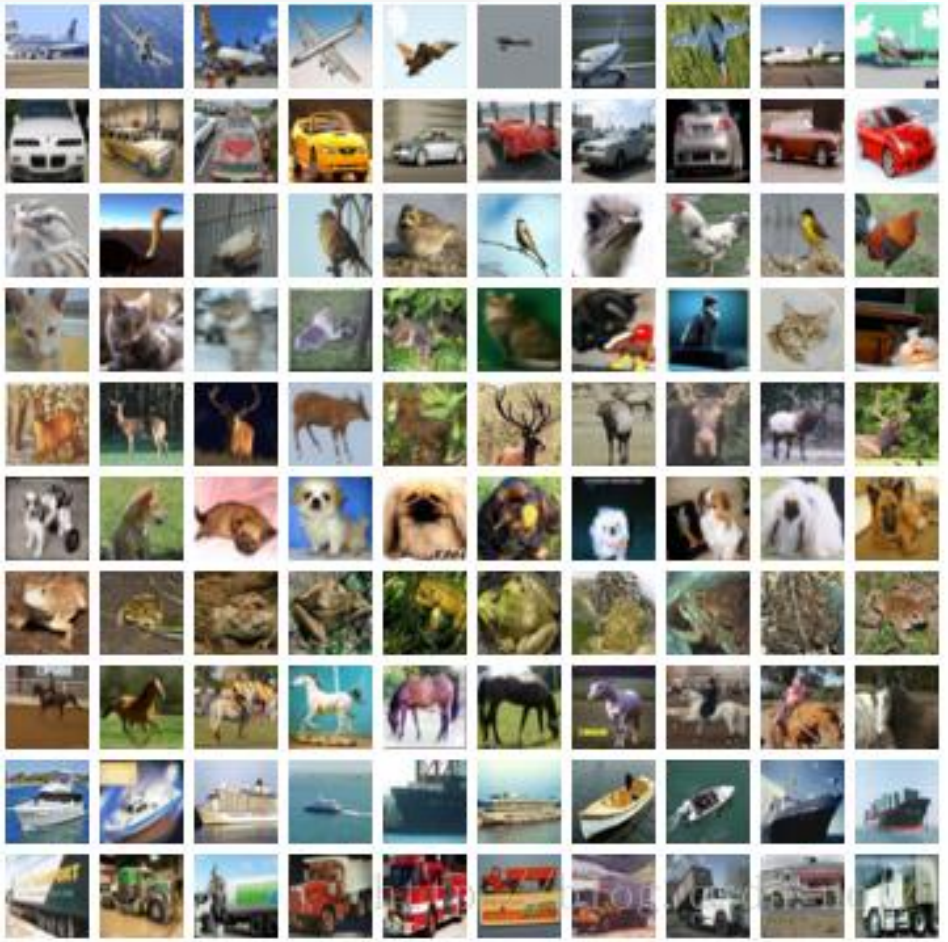
电信用户流失预测

- 当产品无法留住用户时，产品就像是一个筛子，这也使得放进的砂砾越多流失的也就越多。对于客户流失率而言，每增加5%，利润就可能随之降低25%-85%。随着市场饱和度的上升，电信运营商亟待解决增加用户黏性，延长用户生命周期的问题。因此，电信用户流失分析与预测至关重要。
- 案例数据集选自Kaggle——电信客户流失分析(Telco Customer Churn)
- 数据集共有7043条数据，20个字段，具体描述了电信用户是否流失及相关信息。

序号	字段名	字段描述
1	customerID	客户ID
2	gender	性别
3	SeniorCitizen	是否是老年人
4	Partner	是否有伴侣
5	Dependents	是否有家属
6	tenure	使用产品时长
7	PhoneService	是否开通电话服务业务
8	MultipleLines	是否开通多线业务
9	InternetService	是否开通互联网服务
10	OnlineSecurity	是否开通网络安全服务
11	OnlineBackup	是否开通在线备份
12	DeviceProtection	是否开通设备保护
13	TechSupport	是否订购技术支持服务
14	StreamingTV	是否订购网络电视
15	StreamingMovies	是否订购网络电影
16	Contract	签订合同方式
17	PaperlessBilling	是否开通电子账单
18	PaymentMethod	客户端支付方式
19	MonthlyCharges	月费用
20	TotalCharges	总费用
21	Churn	是否流失

CIFAR-10图像分类数据集

- 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32x32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13。
- 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。



CIFAR-10下载地址：https://www.cs.toronto.edu/~kriz/cifar.html