2020年 机器学习课程结课项目

(任选其一完成)

题目1　疫情期间网民情绪识别

（1）　题目背景

2019 新型冠状病毒（COVID-19）感染的肺炎疫情发生对人们生活生产的方方面面产生了重要影响，并引发国内舆论的广泛关注，众多网民参与疫情相关话题的讨论。为了帮助政府掌握真实社会舆论情况，科学高效地做好防控宣传和舆情引导工作，本题目针对疫情相关话题开展网民情绪识别的任务。

（2）　数据集

数据集依据与“新冠肺炎”相关的 230 个主题关键词进行数据采集，抓取了2020年1月1日—2020年2月20日期间微博数据，并对其进行人工标注，标注分为三类，分别为：1（积极），0（中性）和-1（消极）。

训练数据以csv格式存储在train.csv文件中，其中包含45000条微博数据，具体格式如下：

[微博中文内容，情感倾向]

1.微博中文内容，格式为字符串

2.情感倾向，标签取值为{1,0,-1}

（3）　任务描述

根据train.csv文件中的微博数据，设计算法对test.csv文件中的4500条微博内容进行情绪识别，判断微博内容是积极的（1）、消极的（-1）还是中性的（0）。

将结果存储在 csv 文件中，编码采用 UTF-8 编码，格式如下：

微博中文内容　　情感倾向

新冠肺炎……　　1

（4）　评测标准

基于以下混淆矩阵(confusion matrix)，采用 Precision，Recall，F1-score 三个指标评价算法结果，要对比 3 种以上算法的结果，可进一步自由发挥，做算法参数敏感性的实验及对比分析等。

Confusion matrix		真实值	
		positive	negative
预测值	positive	TP	FP
	negative	FN	TN

其中，TP 是真阳例，TN 是真阴例，FP 是假阳例，FN 是假阴例。

Precision：精确率（ 查准率 ），即为在预测为 1 的样本中，预测正确（实际为 1）的人占比，,用混淆矩阵中的字母可表示为：

$$Precision = TP/(TP+FP)$$

Recall：召回率（查全率），即为在实际为 1 的样本中，预测为 1 的样本占比,用混淆矩阵中的字母可表示为：

$$Recall = TP/(TP+FN)$$

F1-score：F1 分数（F1 Score），是统计学中用来衡量二分类模型精确度的一种指标。它同时兼顾了分类模型的准确率和召回率。F1 分数可以看作是模型准确率和召回率的一种加权平均，它的最大值是 1，最小值是 0。

$$F1=2*Precision * Recall /(Precision + Recall)$$

题目2　传感器异常数据检测和预警

（1）　题目背景

有一座数字化电厂，在生产过程中会产生大量的传感器数据，在不同工况下数据表现或分布是不一样的，现在有运行正常的数据（工况未知）用来训练，如何对生成过程中的异常数据进行检测和预警？请利用所学的机器学习方法给出解决方案，并分析其中的参数或阈值的设置和优化　(数据见附件)

（2）　数据集

数据见附件

（3）　任务描述

自行选择模型

（4）　评测标准

可以用预测 loss 表示模型性能。

题目3　人脸识别

（1）　题目背景

人脸识别功能包括人脸检测（从给定图片中用矩形框框住人脸位置，并裁剪出人脸照片）、人脸特征提取（输入人脸照片，输出是高维矢量特征）、人脸匹配（通常用 K-NN 算法，找出待查询人脸的最近邻居，在具体应用中，一般是 1-NN）。可利用现有开源框架，完成一个人脸识别系统，能够在自己台式机或笔记本上实现

实时人脸识别。测试时，人脸库中的照片可以是家庭成员，也可以是班级同学。该项目主要考核同学们对于 K-NN 算法和核函数的掌握，具体测试时实验设置如下：

① 可以调整人脸库中单个人注册的人脸照片数；

② 采用不同距离函数（L2-distance）或和核函数（余弦函数）的设置；测试不同参数设置情况下，参数变化对人脸识别系统影响。

对于熟悉 C++同学，可以采用中科视拓（https://github.com/seetafaceengine/SeetaFace2）的人脸检测和特征提取模型；对于熟悉 python 的同学，人脸检测可以使用 retinaface(https://github.com/peteryuX/retinaface-tf2)，人脸特征提取可以使用 LightCNN (https://github.com/AlfredXiangWu/LightCNN)。对于开源框架使用不限于推荐的两种，同学们可以自己搜索更好的资源。

（2） 数据集

自行获取

（3） 评测标准

分类正确率。

题目4 个人收入预测

（1） 数据集

给定训练集 income.csv，要求根据每个人的属性值来判断此人年收入是否大于 50K。

训练集介绍：

（1）CSV 文件，大小为 4000 行×59 列;

（2）4000 行数据对应着 4000 个人，ID 编号从 1 到 4000;

（3）59 列数据中，第一列为 ID，最后一列 label(1 或 0)表示年收入是否大于 50K，中间的 57 列为 57 种属性值。

（2） 任务描述

（1）将数据中前 3000 项作为训练集，后 1000 项作为测试集，使用 logistic 回归进行二分类，实现语言要求为 Python;

（2）在使用梯度下降法时，调整学习率的固定值，有能力的同学可以学习并使用动态调整学习率的方法，探究不同学习率的选择对训练误差收敛速度的影响，绘制 misclassification rate 曲线进行比较并分析。

（3） 评测标准

（1） 要求计算出准确率。

（2） 要求画出训练和测试 loss 曲线。

（3） 要求调整多个学习率和正则化参数后给出上面的结果。

题目5 分类器性能对比

（1） 数据集

MNIST 手写数字数据集是具有 60,000 个示例的训练集和 10,000 个示例的测试集。它是 NIST 提供的更大集合的子集。数字已经过尺寸标准化并以固定尺寸的图像为中心。对于想要在真实数据上尝试学习技术和模式识别方法，同时在预处理和格式化方面花费最少的人来说，它是一个很好的数据库。

数据集链接：

http://yann.lecun.com/exdb/mnist/

（2） 任务描述

在之前的实验中，我们尝试用了 knn 对 MNIST 数据集进行分类，该任务需要同学用 knn,svm,Logistic Regression,Naive Bayes 对 MNIST 进行分类，并对分类效果（准确率，时间等等）进行比较。

尝试引入非线性特征观察对效果的影响，如 (x1,x2)->(x1,x2,x1^2,x2^2,x1x2)，kernel 函数等等方法。

题目6 基于朴素贝叶斯分类器的语音性别识别

（1） 题目背景

用朴素贝叶斯分类器进行数字手写体识别(基于 MINIST 数据集)，因此在这里用朴素贝叶斯在语音上做一个小应用——分辨声音是男性还是女性。具体题目可以参考 https://www.kaggle.com/primaryobjects/voicegender

（2） 数据集

数据集可自行在 https://www.kaggle.com/primaryobjects/voicegender 下载或附件。这个数据集是基于对男女语音段进行合理的声音预处理而得到的语音特征(并不包含原始语音段)。集合中共有 3168 条数据，男女各 1584 条，每条数据可视作一个长度为 21 的一维数组。其中前 20 个数值是这条语音的 20 个特征值，这些特征值包括了语音信号的长度、基频、标准差、频带中值点/一分位频率/三分位频率等；最后一个数值是性别标记。元数据集中直接以字符串,即 male 和 female 进行标注。使用 7：3 划分数据集。

（3） 任务描述

通过朴素贝叶斯方法，可以先对所有特征值做统计，并且通过连续性参数估计（高斯分布）方法得到参数。之后使用预测函数预测测试集。

（4） 评测标准

要求得到 2*2 预测情况

男声正确率	男声错误率
女声正确率	女声错误率

题目7 自选

个人根据本学期的学习内容和自学内容，自选