Dian团队2017年冬季招新·通宵测试题

恭喜大家顺利通过Dian团队2017年冬季招新的简历筛选、笔试和面试考核，来到通宵测试的现场。在这里希望大家解放思想和双手，敢于尝试和突破，敲出激动人心的奇迹！

一、题目概述

对图片中的文字识别是一个很有意思的研究，在今晚的通宵测试中，需要同学们完成一个对手写阿拉伯数字进行识别的任务。

二、环境说明

编程语言：C语言
说明：编程环境不限，但只能使用C标准库，C语言标准最高为C11，所有代码需要独立完成。

三、任务详述

题目包中的digits目录下，包含了两个目录，

```
.
├── testDigits
└── trainingDigits
```

其中`trainingDigits`为训练集，`testDigits`为测试集，两个集合的数据均为txt格式的文本文件，每一个文件为一个样本。文件名如：`0_173.txt`，文件名中，下划线前的数字`0`为这个样本的标签，下划线后的数字`173`为样本序号。
建立模型，通过训练集的学习，来对测试集中的样本做测试，并评估模型的准确率。
接下来我们会一步步完成今晚的最终目标。

1、编写函数，打开一个目录，获取该目录下的所有文件名。

如：

```
int read_file_list(char *path);
```

函数接收的参数为一个目录的路径，函数打印该目录下所有文件的文件名，返回值为该目录下文

件的数量（不含包含目录文件）。

2、编写函数，读取单个样本文件，将文件的标签和数据读取到你自己定义的数据结构中。打印该文件的内容。

如：

```
void read_a_single_file(char *f_path, your_data_struct);
```

该函数读取`2_79.txt`，将数据保存至你自定义的数据结构中，打印数据：

```
00000001111111000000000000000000
00000011111111100000000000000000
00000011111111110000000000000000
00000011111111111000000000000000
00000011111111111100000000000000
00000011111111111110000000000000
00000011111000111111000000000000
00000011111000111111000000000000
00000011100000111111000000000000
00000000000000011111000000000000
00000000000000011111000000000000
00000000000000011111000000000000
00000000000000011111000000000000
00000000000000011111000000000000
00000000000000011111000000000000
00000000000000011111000000000000
00000000000000011111000000000000
00000000000000011111000000000000
00000000000000011111000000000000
00000000000000011111000000000000
00000000000000011111000000000000
00000000000000011111000000000000
00000000000001111111000000000000
00000000000011111111100000000000
00000000111111111111111111100000
00000001111111111111111111110000
00000001111111111111111111110000
```

```
00000001111111111111111111110000
00000011111111111111111111100000
00000001111111111111111111000000
00000000111111110000000000000000
```

3、在步骤1、2的基础上扩展。

在步骤1、2的基础上，编写函数，将`trainingDigits`目录下所有的数据以及他们的标签读入内存，用你自己的方式将他们管理起来。

4、编写将你读入的数据处理成向量，计算两个向量间的间隔。

如：

```
float cal_distance(your_data_structA, your_data_structB);
```

该函数可以计算两个向量间的距离，计算向量距离的公式为：

$$distance = \|u - v\|_2 = \sqrt{(u_1 - v_1)^2 + \ldots + (u_n - v_n)^2}$$

5、利用步骤2、4的成果，测试`testDigits`目录下的单个样本

- 利用步骤2实现的功能，读入`testDigits`目录下的单个样本的单个样本，逐一计算它与`trainingDigits`目录下每个数据的向量间隔。
- 你的函数接收一个参数`k`，找出前k个与测试样本间隔最小的训练样本。
- 统计这些样本的标签，取k个标签中最多的哪一种，作为测试样本的标签。
- 观察这个标签与测试样本的文件名中的标签是否一致。

6、评估该方法的效果

- 调整上述参数`k`对`testDigits`目录下所有的样本进行测试，统计正确率。
- 比较不同k值下的正确率。

7、识别bmp图片中的数字

在完成1-6步骤的基础上进行扩展：支持bmp图片的读入，利用你的模型判断图片中的手写数字。

说明：

- bmp图片来源自行选择
- 解析bmp图片的方法请通过网络学习
- 将图片处理成模型可接收的输入的方法请自行设计

8、扩展内容

使用其它的模型来完成手写数字的识别。如：

- 逻辑回归
- 决策树
- 支持向量机
- …

说明：

> 尝试了某个模型，即使没有实现，也可以在答辩中说明你对这个模型的理解。

四、其它说明

- 尽量不要跳跃上述步骤，一步步完成任务，注意保存你的阶段性成果。
- 注意代码的规范和充分的注释，高质量的代码会提高成绩。
- 遇到困难请利用网络查找解决办法，不懂的知识可以上网学习。
- 题目描述不清的地方可以询问在场的Dian团队队员
- 独立完成编码工作，不要在线求助
- 抄袭作弊的行为很容易被发现并取消资格

五、成果验收

- 将代码跟目录以你的名字的全拼（无空格，全小写）提交。
- 准备一个五分钟的答辩PPT，演示你的通宵测试成果。