同济大学课程

python数学建模期末随堂测验

2019— 2020学年第2学期

年级 大二 专业 软件工程 学号 1851201 姓名 周子龙 得分

完成时间 以提交时间为准（压缩文件的生成时间）

**请在你的D:\上（其它盘也可），建立目录 ： py+你的学号，如学号 190001，则目录为：D:\py1900001**

**将老师提供的数据文件，全部拷贝到该目录下。**

**请你在该目录下编写你的程序代码（你所需要的自己编写类库，也请拷贝到该目录下），**

**程序文件名，以题号命名，如第一题是 1.py**

**读数据文件的地方，全部采用相对路径，如下格式:**

**A=np.loadtxt(r'data.txt')**

**测验结束后，将将整个目录打包上传到cal网站cal.tongji.edu.cn/IT，也发邮箱** [**pshcong@tongji.edu.cn,**](mailto:pshcong@tongji.edu.cn,测验答案发邮箱) **邮件的标题： 学号+数学建模**

**成绩以正确率、完成时间排序。**

**一、程序**

1. sklearn中提供的手写数字图片库，可以用下面获得其数据

from sklearn.datasets import load\_digits

digits = load\_digits()

X = digits.data # 数字图片特征向量

y = digits.target # y是数字，如 0，1，2，..., 9

X矩阵中有某些列都是0，这样数据对建模是没有意义的，请你用np的数组过滤功能，将这些列过滤掉，使得到的新X，不存在所有元素都是0的列。

2、通过本学期的学习，对线性模型，我们讲过哪些算法？

应该如何判断所建模型的好坏？

如何判断变量间是否相关，判断独立变量数的具体指标有哪些？

对有变量相关的建模问题，有哪些算法可解决？

假如你是个精通数学优化理论的人，且有将算法程序化的能力，请问，对你而言，多层感知器算法MLP和Keras，哪一个更能让你腾飞？原因是什么？

1. 线性模型算法：一元线性回归，多元线性回归，主成分分析，偏最小二乘法（向方差最大化方向旋转）（MLR PCA PCR PLS），此外，一般的神经网络算法也可以求解线性模型。
2. 如果是判别，聚类等模型，可以对原始数据集进行训练集和测试集划分，通过统计学中的误判率、召回率等数据对模型进行判别。由于数据集的划分具有随机性，一般的做法是对训练集和测试集合进行交叉验证；

如果是回归等，可以通过fisher检验、、调整后来对模型进行评估，同时还可以进行多因素方差分析，对模型中参数的有效性进行检验。

1. 可以对变量间求解相关矩阵，来判断是否有关系，此外，还可以通过多个变量进行方差分析进行检验变量的自相关性。一般的判断独立变量数的指标为求解主成分以及特征值，这里有几种求独立变量数的手段：一是选取特征值大于1的成分，二是选取反应样本方差变化80%（人为指定）的主成分数作为独立变量数，三是可以通过碎石图，找到突越变化的主成分，四其实和碎石图一样，但是通过具体的数据来判断，使用前一主成分特征值于后一主成分作比，舍掉小于某一比值后的所有主成分。
2. 一般可以通过主成分分析法首先求取主成分，对数据进行将为后求解。在求解过程中还可以通过方差最大化方向旋转来得到更好的效果。（偏最小二乘算法）
3. 我认为是Keras更能让我腾飞。因为keras首先是基于tensorFlow的，这一强大的后台和优化具有其他平台不可比拟的好处。

此外，keras在编写上具有高度的模块化，以及可以让用户自定许多模型的关键参数，以及自行定制用户自己的模型；类似于损失函数、优化器、等等，这些都可以让用户自行指定，可以更具实际问题，针对化的进行建模。

此外，keras还支持CPU和GPU的切换，可以为较大型的数据处理提供支持。

3. sklearn.datasets中有一套手写数字，共有1797个图片，被分成10类（10个数字0~9），可以通过如下语句加载：

from sklearn.datasets import load\_digits

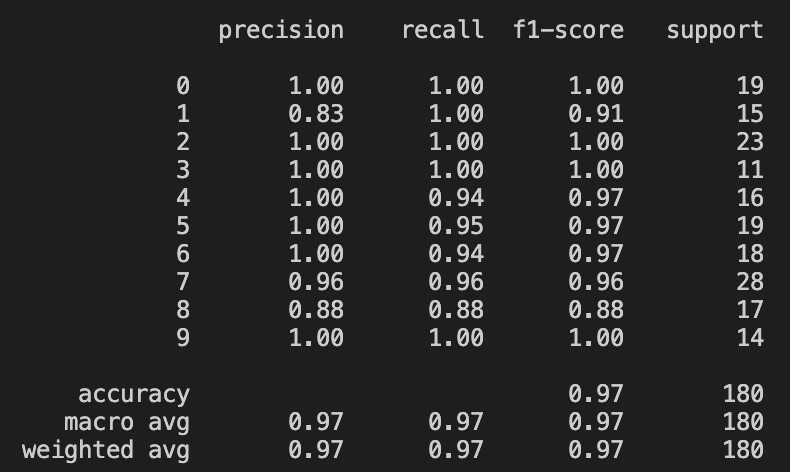
digits = load\_digits()

X = digits.data

y = digits.target

X的每一个样本都是64个元素，请使用keras，讨论合理的网络拓扑结构、优化算法，对数据进行建模，对建立的模型进行评价。

输入层节点个数：64，中间隐藏1层，节点个数为256，输出层节点数为10，表示对10个数组的预测。



进行100次迭代后对预测样本的结果如上。平均准确率达到了97%，可以认为是很有效的。

4、数据文件datatest.txt是一个实验实测信号，现在需要检测信号在哪些位置出现峰值，信号中有轻微的白噪声，所以基线位置有小的波动，数值都比较小，例如都小于0.1。

请你首先谈谈，如果想得到如下的测定峰位置的图，应该使用哪些数学中的理论；

1. 可以使用随机概率模型的方法，具体操作如下：

生成一个在数据取值区间的一个随机数，作为起始位点。

在该位位点半径的区间上继续缺的另外一个随机数，如果新位置的数值大于原位点，则保留，将新位置记录为原位置，否则继续本步骤。

不断的进行上述随机过程，在每一次选取新位置后，降低区间半径的大小，以期望获得收敛。

这样总能找到一个峰值，似的他在半径区间为的位置上为最值。

反复这样的过程，每次求接触一个峰值后，下次就不在考虑该峰值，继续求解其他峰值。

其二，请你编程求解得到如下的图。图例中，蓝色是原图，红色是每个峰的位置，第一个数字是峰的位置（数组的第几个元素），第二个数字，则是x的坐标（即数组元素的取值）。

