实验任务书－2020

一、熟悉scikit-learn数据挖掘包（https://scikit-learn.org/stable/），包括数据预处理、分类算法、聚类算法、性能评价metric等，特别是如下分类算法的使用：

（1）线性判别分析（LDA）
（2）支持向量机（Support Vector Machine）
（3）最近邻分类器（Nearest Neighbors）
（4）朴素贝叶斯（Naive Bayes）
（5）决策树C4.5（Decision Trees）
（6）分类与回归树CART
（7）集成学习（Ensemble Methods）中的随机森林（Random Forest）
（8）集成学习中的Adaboost
（9）集成学习中的Gradient Tree Boosting
（10）半监督学习（Semi-Supervised Learning）中的标签传播（LabelPropagation、LabelSpreading）

注：半监督学习，顾名思义是介于分类（监督学习）与聚类（无监督学习）之间的一种学习范式。给定很少一部分样本的类标签，怎么样利用少部分具有类标签的数据来提高聚类的准确率是其研究主题。其中基于图的标签传播（Label Propagation）算法是有影响的算法之一。

二、熟悉深度学习平台keras（https://keras.io/）。

Keras（最新发布版本2.3.0）是深度学习平台Tensorflow与Theano的进一步API封装，简单易用。需要注意的是Theano将会退出历史，因此要以Tensorflow 2.0为backend。

三、实际操作

1、数据集：UCI机器学习数据库（http://archive.ics.uci.edu/ml/index.php）的分类任务（Classification Task），截止到2020-4-12日，共有366个数据集。

2、个人使用的数据集为：学号除以数据集数目366的余数。

3、个人使用的算法按如下方法确定：

1. 算法字符串：LMNBCTRAGP，其中：

L：线性判别分析LDA；M：支持向量机；N：最近邻分类器；B：朴素贝叶斯；C：决策树C4.5；T：分类与回归树；R：随机森林；A：Adaboost；G：Gradient Tree Boosting；P：标签传播。

2. 姓名字符串的确定：姓名三个中文字符的拼音首字母（只有2个的再取离第2个首字母最近的下一个字母（如果第2个首字母是Z，则下一个字母为A），4个及以上的取前3个）。

3. 字母编码如下：

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

U	V	W	X	Y	Z
21	22	23	24	25	26

4. 使用算法的确定：使用最小平方误差和准则，确定算法字符串中的三个连续字母（不回绕）组成的算法字符子串作为实验的三个算法。

例如：张国荣对应的姓名首字母字符串为ZGR，其对应的编码是26，7，18，而算法子字符串MNB对应的是13，14，2，其误差为$(26-13)^2+(7-14)^2+(18-2)^2$。

5. 除上述算法外，每个人还必做深度学习（用keras平台）。

四、注意事项及相关说明：

1. 语言为Python，以前没有使用过的正好通过这次实验进行熟悉（python为机器学习的最主流语言，具有最大的开源社区，Tensorflow 2.0、Keras用户最为广泛）。
2. 对于深度网络而言，若数据集过小，则极易过拟合，因此若自己数据集过小，可以选下一个大一些的数据集，并在文档中加以注释说明。
3. 评价方法：大一点的数据集采用10-折交叉验证（10-fold cross validation），小一点的数据集（如200以下）采用留一测试。
4. 标签传播算法须实验2个：LabelPropagation和LabelSpreading，其中使用的标签样本为样本总数的5%和10%。因为标签传播使用的带有标签的样本较少，评价准确率可用聚类的评价标准：ARI或NMI。
5. 提交内容：实验报告、源代码，以及导出为Excel文件的数据集，以便我能验证。其中实验报告部分，要将4个算法的结果列在一张表中进行对比。以班级为单位由班长统一提交（以学号+姓名+数据集名+算法字符串命名）。
6. 提交时间：第12周周日晚上12：00以前。实验占总评成绩的30%，逾期未交，此部分成绩为0分。

7、若发现抄袭，抄袭者与被抄者均计0分。