信用卡违约率分析

三个目标：

1、创建各种分类器，包括 SVM、决策树、KNN 以及随机森林分类器；

2、掌握 GridSearchCV 工具，优化算法模型的参数；

3、使用 Pipeline 管道机制进行流水线作业。（准备过程，数据规范化、数据降维等）

一、构建随机森林（RF）分类器

sklearn 里的 RandomForestClassifier() 构造随机森林模型，常用构造参数如下：

n_estimators	随机森林里决策树的个数，默认是10
criterion	决策树分裂的标准，默认是基尼指数（CART算法），也可以选择entropy（ID3算法）
max_depth	决策树的最大深度，默认是None，也就是不限制决策树的深度。也可以设置一个整数，限制决策树的最大深度。
n_jobs	拟合和预测的时候CPU的核数，默认是1，也可以是整数，如果是-1则代表CPU的核数

创建后，使用 fit 函数拟合， predict 函数预测。

二、使用 GridSearchCV 工具对模型参数进行调优

说明：分类算法需要经常调节参数（对应上面的构造参数），目的是得到更好的分类结果。

GridSearchCV，是 Python 的参数自动搜索模块。

输入想要调优的参数以及参数的取值范围，就会把所有的情况都跑一遍，得出最优的参数结果。代码如下：

```
# 引入 GridSearchCV 模块
from sklearn.model_selection import GridSearchCV
# 构造参数的自动搜索模块
GridSearchCV(estimator, param_grid, cv=None, scoring=None)
```

常用参数如下：

estimator	代表我们想要采用的分类器，比如随机森林、决策树、SVM、KNN等
param_grid	代表我们想要优化的参数及取值，输入的是字典或者列表的形式
cv	交叉验证的折数，默认为None，代表使用三折交叉验证。也可以为整数，代表的是交叉验证的折数
scoring	准确度的评价标准，默认为None，也就是需要使用score函数。也可以设置具体的评价标准，比如accuray，f1等

构造完 GridSearchCV 后，使用 fit 函数拟合训练， predict 函数预测，这时预测采用的是最优参数情况下的分类器。

！注意:

本练习使用的分类器为随机森林。

随机森林每建一个树都有两次随机抽样：样本随机抽样和属性随机抽样。

本质是为了防止过拟合，会导致GridSearchCV寻找最优参数每次运行结果不一样。

使用随机数的地方可以指定随机种子

三、使用 Pipeline 管道机制进行流水线作业

说明：Python 有一种 Pipeline 管道机制。管道机制就是让我们把每一步都按顺序列下来，从而创建 Pipeline 流水线作业。每一步都采用 ('名称', 步骤) 的方式来表示。

做分类的步骤：先对数据进行规范化处理，再对数据降维，最后使用分类器分类。

具体即为：

1、先采用 StandardScaler 方法对数据规范化，

即采用数据规范化为均值为 0，方差为 1 的正态分布；

2、再采用 PCA 方法对数据进行降维；

3、最后采用随机森林进行分类。

代码如下:

```
from sklearn.model_selection import GridSearchCV
pipeline = Pipeline([
        ('scaler', StandardScaler()),
        ('pca', PCA()),
```

```
        ('randomforestclassifier', RandomForestClassifier())
])
```

信用卡违约率的项目：
数据集是台湾某银行 2005 年 4 月到 9 月的信用卡数据，数据集一共包括 25 个字段，具体含义如下：

字段	含义
ID	客户ID
LIMIT_BAL	可透支金额
SEX	性别，男：1，女：2
EDUCATION	教育程度，研究生：1，本科：2，高中：3，其他：4
MARRIAGE	婚姻，已婚：1，单身：2，其他：3
AGE	年龄
PAY_0	2005年9月客户还款情况
PAY_2	2005年8月客户还款情况
PAY_3	2005年7月客户还款情况
PAY_4	2005年6月客户还款情况
PAY_5	2005年5月客户还款情况
PAY_6	2005年4月客户还款情况
BILL_AMT1	2005年9月客户每月账单金额
BILL_AMT2	2005年8月客户每月账单金额
BILL_AMT3	2005年7月客户每月账单金额
BILL_AMT4	2005年6月客户每月账单金额
BILL_AMT5	2005年5月客户每月账单金额
BILL_AMT6	2005年4月客户每月账单金额
PAY_AMT1	2005年9月客户每月还款金额
PAY_AMT2	2005年8月客户每月还款金额
PAY_AMT3	2005年7月客户每月还款金额
PAY_AMT4	2005年6月客户每月还款金额
PAY_AMT5	2005年5月客户每月还款金额
PAY_AMT6	2005年4月客户每月还款金额
default.payment.next.month	下个月是否违约，违约：1，守约：0

项目流程分析如下：

1、载入数据；

2、探索数据，划分数据集，测试集与训练集；

3、使用 Pipeline 管道机制，将数据规范化设置为第一步，分类为第二步；

4、构造各种分类器并使用GridSearchCV对具体的分类器进行参数调优。