# 《数据仓库与数据挖掘》实验指导书

**实验项目一：**Python语法和数据挖掘应用

**实验目的**：

1. 学习python语法，理解sklearn库的调用方法；
2. 熟悉Python语言，掌握典型的sklearn方法。

**实验任务：**

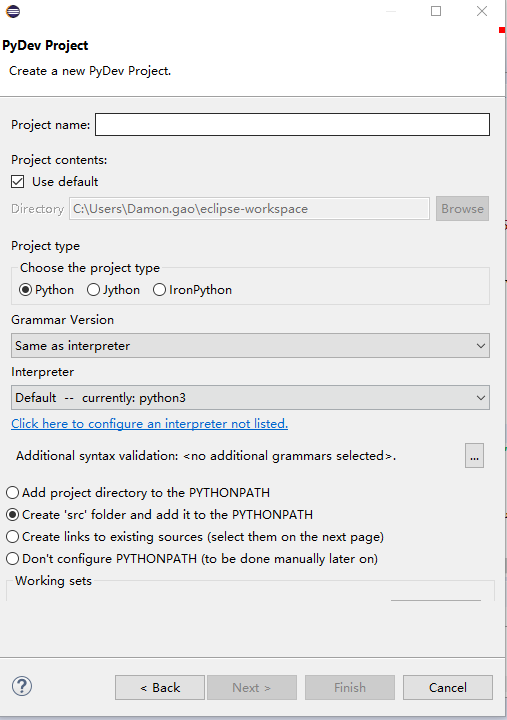
1. 掌握Python中对各种包的导入；
2. 使用python语言中numpy包进行数组的生成和数组切分；
3. 使用python语言中numpy包生成数组进行排序；
4. 使用python语言中numpy包生成数组进行拼接实验；
5. 掌握sklearn包中函数的导入和使用，为下次实验打好基础。

**实验方式**：程序设计

**实验场地与设备**：实验室4074、计算机

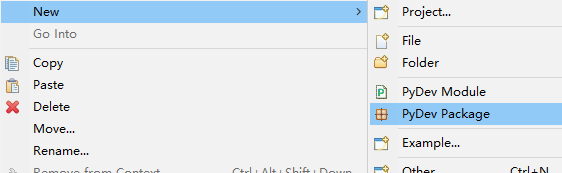
**实验项目创建流程：**

**第一步：运行**eclipse—点击左上角“file”，选择“new”，找到“other”—pydev project，会出现如下图的复选框：

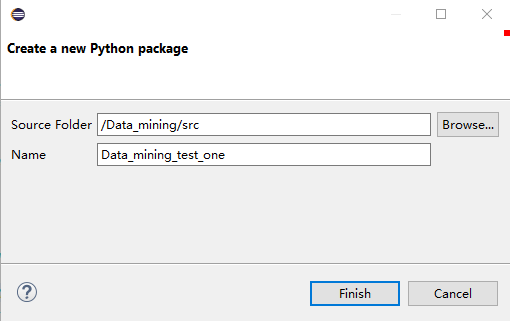


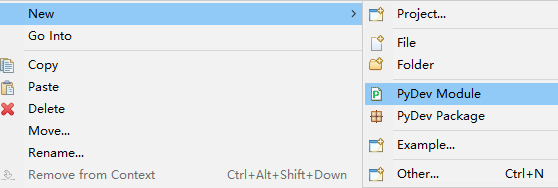
根据自己的习惯新建project name项目名称， 其他的选项在实验环境下可以选择默认（在自己电脑上可以根据自己的情况做调整）。在这里我将项目名设置为Data\_mining，然后点击finish。这是我们在桌面左边就会看到刚才我们新建的空项目。

第二步：在打开空项目，我们会看到两个文件夹，其中一个“src”，另外一个是python的版本信息。选择“src”，右键然后选择“new”，接着选择PyDev\_Package。如下图所示：

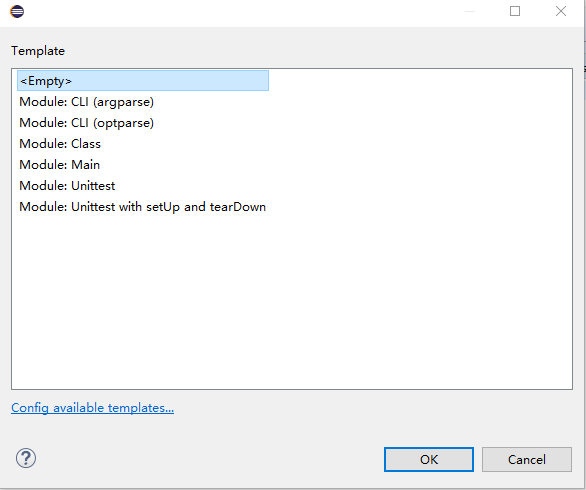


接下来会弹出如下窗口：在name文本框中，输入当前package的名字（示例：Data\_mining\_test\_one）,然后点击finish。

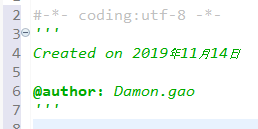


第三步：点击打开刚才新建的package，然后右键选择“new”，选择pydev\_moudle

，会弹出和上面类似的窗口，然后在name文本框中输入当前模块的名字，接下来点finish，然后会弹出如下窗口，选择<Empty>，然后ok。



第四步：在刚才新建的模块上面输入#-\*- coding:utf-8 -\*-（目的是为了可以在注释汇总显示中文字符）



下面就可正式开始编辑我们的实验代码了。

**实验项目二：**分类算法的实现

**实验目的**：

1. 理解不同分类的算法的原理，掌握决策树、贝叶斯分类和SVM的应用。
2. 对给定数据集进行整理，利用分类算法对给定数据集进行分类

**实验任务：**

1. 导入sklearn包中自带的鸢尾花数据集；
2. 对数据集做预处理；
3. 用数据集中的训练部分训练svm分类器；
4. 用构建好的svm模型对测试部分的数据做预测；
5. 将预测的标签和其真实的标签作对比，计算模型的准确率。

**实验方式**：程序设计

实验场地与设备：实验室4074，计算机。

**实验项目三：关联分析应用**（支撑教学目标1，2，3）

**实验目的**：

1. 对给定数据集进行整理，利用apriori， FP-Growth算法对给定数据集进行分类；
2. 理解不同分类的算法的原理，掌握apriori， FP-Growth 算法的应用。

**实验内容**：

1. 掌握apriori算法原理；
2. 掌握FP-Growth算法原理；
3. 调用sklearn中的数据集进行相关算法的实验。

**实验方式**：程序设计

**实验场地与设备**：实验室4074，计算机。

**实验项目四：**聚类算法的实现

**实验目的：**

1. 掌握聚类算法的思想，尤其是k-means、k中心点算法，掌握利用Python语言进行聚类的方法。
2. 要求学生学习和掌握典型的关聚类发现算法，并利用Python语言实现聚类算法。

**实验内容**：

1. 理解K-means聚类算法的原理；
2. 掌握k-means聚类算法的迭代更新过程；
3. 动手实践k-means算法；
4. 在给定的数据集上使用k-means进行聚类；
5. 调用skearn包中的k-means算法，并对比实验结果。

**实验方式**：程序设计

**实验场地与设备**：实验室4074，计算机。