数据挖掘课程实验

**要求：**

1. 分为实验报告、实验手册和源代码三部分，学生按要求提交相应文档和代码。

2. 实验报告命名：采用学号+姓名+实验相结合，比如：201608010101-王啸-实验报告1.doc，采用实验+学号+姓名相结合，比如：201608010101-王啸-实验手册1.doc

3. 源代码以压缩文件形式提交，比如：201608010101-王啸-实验代码1.rar

**实验一 实验平台及环境安装**

Python 是一个高层次的结合了解释性、编译性、互动性和面向对象的脚本语言。Python 是 FLOSS（自由/开放源码软件）之一。Python 的设计具有很强的可读性，相比其他语言经常使用英文关键字，其他语言的一些标点符号，它具有比其他语言更有特色语法结构。Python的最大的优势之一是丰富的库，跨平台的，在UNIX，Windows和Macintosh兼容很好。

**实验要求：**在Windows平台下安装、配置python环境和相关软件。具体如下：

1. 在Windows平台上搭建Python平台，并安装Python环境工具anaconda.

2. 掌握Anaconda下的Python环境安装，创建名称为gnn的python3.7环境.

3. 熟练安装pycharm和jupyter notebook。

4. 安装 PyTorch 和其他依赖库，这里使用 PyTorch2.2.2

5. 掌握pip和conda命令安装常用软件包。比如numpy、pandas、tensorflow、 h5py、mygene matplotlib、seaborn、umap-learn等。

**实验二 数据降维与可视化**

数据降维是指将高维数据映射到低维空间的过程。在现实生活中，很多数据集都是高维的，每个样本包含着大量特征。然而，高维数据不仅对计算资源要求较高，而且容易造成“维数灾难”，即在高维空间中，数据样本的稀疏性和分布规律难以理解。数据降维的目的是保留数据集的主要结构和信息，同时减少特征的维数，从而更好地进行数据分析和可视化。

**实验要求：**

在Windows平台下gnn环境中，进行数据降维与可视化。具体如下：

1. 熟悉基本的数据预处理方法，对数据进行初步降维，降维到500-1000之内，降维方法可以自由选择。

2. 熟练掌握无监督数据降维方法，比如PCA，ICA、UMap等

3. 在不同的维度下面对数据进行数据分布分析及可视化比较。

3. 实现数据的可视化，并进行适当的对比分析。

􀒻􂢙􀑻􄰤􀨟􃜭􁆈􀵘􂲴􀞣􃚄􁡆􃘵􃚄􃌫􀇄􂈄􀾲􄙊􄗷􄈳􁸕􀭶􀵪􄟼􄺮􁇒􀒠􂲴􀑌􃾯􀨁􂧠􀋈

􀑚􃇇􂌅􄘋􃹼􀞣􃚄􃿴􀡉􁥆􁧈􀋈􁒦􂈄􄖳􀕈􃕪􂛩