# 同济大学软件学院 2021年春季数据分析/挖掘课程作业1

1. (a) 用数据举例说明Curse of Dimensionality（维度灾难）是什么、以及维度灾难所带来的问题。

(b) 从Dimension reduction（降维）和Feature selection（特征选择）两个角度，分别选取两种降维和特征选取的方法来对某个数据进行实验，并展现和分析结果。

1. 距离（distance）和相似度（similarity）用于计算数据之间（即输入为两行数据）的相近程度，而相关系数（correlation）、信息论中的熵（entropy）等往往可以用在数据的属性Attribute（即输入为两列数据）上。对此请完成以下两个实验：
2. 实现三种以上的距离和相似度计算，实验对比并分析指出计算结果之间的差异性。
3. 实现三种相关系数和基于信息论的信息度量计算。
4. (a) 数据的预处理中，缺失值（missing value）是经常碰到问题，请找一个带缺失值的数据，对该数据进行相应的预处理，并分析评价该预处理的方法对于数据后续处理的影响。

(b) 数据采样（sampling）也是数据预处理中常用技术，请在某个数据上进行采样处理，并分析评价该方法对于数据后续处理的影响。

注：

1. 本次作业所涉及的数据可自由选择，请前往<http://archive.ics.uci.edu/ml/index.php>自行下载。
2. 代码的实现语言建议Python，其它语言的实现请注明。

提交日期：2021/03/25 23:59，提交至canvas。提交内容要求：

提交文件命名为学号\_姓名\_hw1.zip，其中包括3个子目录，命名方式分别为q1,q2和q3，每个子目录包括以下内容：

1. 源代码文件。
2. README文件，介绍运行环境和运行方式。
3. 实验报告文件，其中报告包括1）数据描述；2）代码运行结果屏幕拷贝；3）讨论分析部分；4）实验结果图表。