

# 实验一 数据探索与预处理

## 【实验目的】

1. 了解 Anaconda 的使用与设置。
2. 掌握数据探索和数据清洗的基本方法。

## 【实验学时】

建议 2 学时

## 【实验环境配置】

- 1、Windows 环境
- 2、Anaconda
- 3、Pandas
- 4、Matplotlib, seaborn

## 【实验原理】

数据挖掘中数据预处理的实现：

- 1、基于 Pandas 和 sklearn 库的数据预处理的实现
- 2、基于 Matplotlib 和 seaborn 库的数据可视化

## 【实验步骤】

1. 打开 Anaconda 的 jupyter notebook，创建实验一，明确标题和步骤。
2. 导入 pandas 库和 os 库，导入实验数据 horse-colic.data，并构造 dataframe 形式的完整数据集，对导入的实验数据进行数据清洗，包括填充缺失值等。
3. 对导入的实验数据完成实验数据的数据探索，展现每个维度数据的概率分布情况，分别绘制直方图、散点图，折线图和箱型图，要求在一张图中包括多张不同维度数据的子图来进行展示。

## 【思考题】

- 1、数据探索的意思是什么？数据预处理的意思是什么？两者有何联系？