

**马的疝病分析**

学院：计算机学院

专业：计算机科学与技术

学号：2120161056

姓名：王帅

1. 问题描述

疝病是描述马胃肠痛的术语，这种病不一定源自马的胃肠问题，其他问题也可能引发马疝病。所给数据集是医院检测的一些指标。数据提供共368个样本，27个特征。

1.1数据处理要求

* 对标称属性，给出每个可能取值的频数。
* 数值属性，给出最大、最小、均值、中位数、四分位数及缺失值的个数。
* 针对数值属性，绘制直方图，如mxPH，用qq图检验其分布是否为正态分布。
* 针对数值属性，绘制盒图，对离群值进行识别。

1.2 数据缺失的处理

数据集中有30%的值是缺失的，因此需要先处理数据中的缺失值。分别使用下列四种策略对缺失值进行处理:

* 将缺失部分剔除
* 用最高频率值来填补缺失值
* 通过属性的相关关系来填补缺失值
* 通过数据对象之间的相似性来填补缺失值

处理后，可视化地对比新旧数据集。

2. 实验环境

编程语言：Python

开发环境：Pycharm、Anaconda

数据分析包：Pandas

Pandas是python的一个数据分析包，最初由AQR Capital Management于2008年4月开发，并于2009年底开源出来，目前由专注于Python数据包开发的PyData开发team继续开发和维护，属于PyData项目的一部分。Pandas最初被作为金融数据分析工具而开发出来，因此，pandas为时间序列分析提供了很好的支持。 Pandas的名称来自于面板数据（panel data）和python数据分析（data analysis）。

绘图框架：Matplotlib

Matplotlib是一个Python的图形框架，类似于MATLAB和R语言。

3. 关键操作

3.1 导入Pandas数据包及绘图包

使用如下语句进行数据包导入：



3.2 导入数据

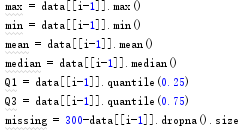
实验中，导入数据为本地CSV文件。使用如下语句进行数据导入：



3.3 数据处理

在pandas中有两类非常重要的数据结构，即序列Series和数据框DataFrame。Series类似于numpy中的一维数组，除了通吃一维数组可用的函数或方法，而且其可通过索引标签的方式获取数据，还具有索引的自动对齐功能；DataFrame类似于numpy中的二维数组，同样可以通用numpy数组的函数和方法，而且还具有其他灵活应用。

pandas模块提供了非常多的描述性统计分析的指标函数，可直接对于数据列进行处理，具体语句如下：



3.4 绘制直方图及盒图

matplotlib库中又直方图及盒须图模块功能，可直接调用，十分方便。具体操作如下：



3.5 缺失值处理

常见的缺失值处理方法有：

删除法：当数据中的某个变量大部分值都是缺失值，可以考虑删除改变量；当缺失值是随机分布的，且缺失的数量并不是很多是，也可以删除这些缺失的观测。

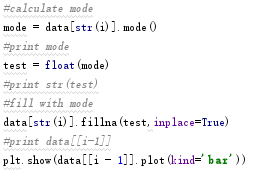
替补法：对于连续型变量，如果变量的分布近似或就是正态分布的话，可以用均值替代那些缺失值；如果变量是有偏的，可以使用中位数来代替那些缺失值；对于离散型变量，我们一般用众数去替换那些存在缺失的观测。

插补法：插补法是基于蒙特卡洛模拟法，结合线性模型、广义线性模型、决策树等方法计算出来的预测值替换缺失值。

使用Pandas可以快速实现删除法和替补法，具体操作如下：



上图为删除缺失数据。



上图为使用众数替换缺失值并制作直方图。

4. 实验心得

在此次实验中，第一次接触Python语言用于处理数据。Python语言在使用Pandas数据包之后对于数据的处理方便快捷，绘制图片也简单易行，是一个不错的数据处理工具。由于完成作业的时间有限，只对这些新知识浅尝辄止，在日后的学习中，要多运用多实践，争取能将这些工具融会贯通。