期末大作业要求

2015 年 6 月 3 日

从主页www.math.pku.edu.cn/teachers/yaoy/Spring2015/下载一组数据，或者选取一组自己感兴趣的数据（需要事先将提议发邮件给任课老师以获得许可）。然后对其作相关的统计学分析，给出相应的结论，并整理成一份完整的报告。报告中建议包含以下内容：

0. 成员分工 (每个小组不超过 4 人)
1. 问题描述
2. 统计建模
3. 编程环境
4. 输出结果
5. 总结分析
6. 参考文献

分析数据时需要使用编程软件，如 R，MATLAB，SPSS，SAS 等，输出结果需要以图表呈现，并做简单的分析。

提交电子版报告及程序代码至cheney@pku.edu.cn。每个小组只需要提交一份即可，报告须为 pdf 格式，程序代码需要有简单的注释。

邮件标题　期末大作业：张三/李四

截止日期　6 月 17 日上午 8 点

额外的三个例子

1 Heart PCI Operation Effect Prediction

The following data, provided by Dr. Jinwen Wang at Anzhen Hospital,

http://www.math.pku.edu.cn/teachers/yaoy/data/heartData_20140401.xlsx

contains 2581 patients with 73 measurements (inputs) as well as a response variable indicating if after the heart operation there is a null-reflux state. This is a classification problem, with a challenge from the large amount of missing values. Sheet 3 and 4 in the file contains some explanation of the data and variables.

The problems are listed here:

1. The inputs (covariates) are of three kinds, measurements upon check-in, measurements before PCI operation, and measurements in PCI operations. For doctors, it is desired to find a prediction model based on measurements before the operation (including check-in). Sheet 2 in the file contains only such measurements.

2. It is also an interesting problem how to predict the effect based on all measurements, with lots of missing values. Sheet 1 contains the full measurements. There are some good work by previous students, which are listed here for your reference:

   The following two reports by LU, Yu and WANG, Qing, are probably inspiring to you.

   http://www.math.pku.edu.cn/teachers/yaoy/reference/LuYu_201303_BigHeart.pdf

   http://www.math.pku.edu.cn/teachers/yaoy/reference/WangQing_201303_BigHeart.pdf

   The following report by MIAO, Wang and LI, Yanfang, pioneers in missing value treatment.

   http://www.math.pku.edu.cn/teachers/yaoy/reference/MiaoLi2013S_project01.pdf

In the final project, it is desired to take only those measurements upon check-in to predict the probability of non-reflux (non-reflow) after PCI operations. An interpretable model adds a big value! You may compare with your first warm-up project to show your improvements.

2 Beer Popularity and Rating

The following data, provided by Mr. Richard (sun.richard@yahoo.com) from Shanghai,

http://www.math.pku.edu.cn/teachers/yaoy/data/Beers_20140514.xlsx

contains 877 brands (rows) of beers in Chinese market, with a few attributes about ingradients, alcoholicity, price (and unit price), reviewers count, mean scores, and as well as sources of reviewers (e.g. amazon, jd, yhd etc.). Two questions are interesting to explore such data

1. What factors are highly correlated with the popularity of beers indicated by reviewers count?

2. What factors accounts for the mean rating scores? Why are those beers lowly rated?

Note that the data does not contain lots of attributes, so think about your goal before you take a try.

3 Keyword Pricing (Regression)

The following data, collected by Prof. Hansheng Wang in Guanghua Business School at PKU,

http://www.math.pku.edu.cn/teachers/yaoy/math2010_spring/Keyword/SE.csv

contains two columns: the first column is a list of keywords; the second column is the profit value (positive for earning and negative for loss). Figure 1 gives some example.

'乌鲁木齐-阿克苏-机票'	14.1200
'乌鲁木齐阿克苏飞机票价'	9.0600
'乌鲁木齐到阿克苏-机票'	-1.1800
'乌鲁木齐到阿克苏打折机票'	-0.4800
'乌鲁木齐到阿克苏机票'	31.9400

图 1: Keywords and profit value

The purpose is to predict the profit value based on features extracted from the keywords, which might be linguistic, geographic, and any new features in your creation. Since the profit values are of real numbers, this problem is regarded as a regression problem by default.

A reference can be found in Mr. Jiaqi Zhu's bachelor thesis work:

http://www.math.pku.edu.cn/teachers/yaoy/reference/Thesis_ZHUJiaqi.pdf