Statistical Learning

Mini-Project 1

Instructor: Yuan Yao *Due: October 12, 2015*

1 Requirement

1. Pick up ONE (or more if you like) favorite problem *below* or *from the datasets in textbook* to attack. If you would like to work on a different problem outside the candidates we proposed, please email course instructors about your proposal. Brave hearts for explorations will be encouraged!

2. Team work: we encourage you to form small team, up to FIVE persons per group, to work on the same problem. Each team just submit ONE report. For example a *poster* report, with a clear remark on each person's contribution. A sample poster file with PKU logo can be found at
`http://www.math.pku.edu.cn/teachers/yaoy/reference/poster_v5.pdf`
whose source LATEX codes can be downloaded at
`http://www.math.pku.edu.cn/teachers/yaoy/reference/pkuposter.zip`

3. In the report, set up your problem to attack and show your results with your analysis. Remember: scientific analysis and reasoning are more important than merely the performance results. Source codes may be submitted through email as a zip file, or as an appendix if it is not large.

4. Submit your report by email or in paper version no later than the deadline, to Teaching Assistants (TA).

2 Animal Species Sleeping Hours Regression

The following dataset contains $n = 51$ species with several features including the average sleeping hours per day.

`http://www.math.pku.edu.cn/teachers/yaoy/SummerSchool_2010/sleep1.csv`

Explore *what affects the sleep an animal needs*? A sample R code can be found at

`http://www.math.pku.edu.cn/teachers/yaoy/Spring2015/Lecture19.R`

You need to create your own design of analysis.

3 Crime Rate

Explore the following dataset about crime rates in 59 US cities during 1970-1992.

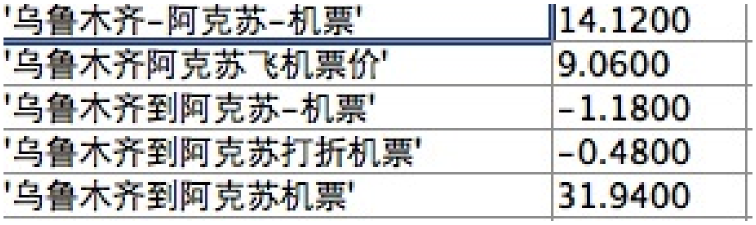
`http://www.math.pku.edu.cn/teachers/yaoy/data/crime.xlsx`

4 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

Figure 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`

5 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.

$$AA^T = A^T A = I$$

$$AA^T = I \Leftrightarrow \|x\|_2 = \|A^T x\|_2, \forall x$$

$$J(A, s) = \|s - A^T x\|_2^2 + \lambda \sqrt{s^2 + \epsilon} + \gamma \|A\|_2^2$$

$$\|A\|_2^2 := trace(A^T A)$$