Statistical Learning **Deadline: April 22, 2014**

Project 1. Linear Regression and Classification

Instructor: Yuan Yao

Our projects encourage you to work in small teams, with each team consisting no more than THREE persons and submitting one report for each team. In your report, please state clearly *the contributions of each teammate*. Send your submission via email to teaching assistants (statlearning_hw@126.com). Choose one of the following problem; if you would like to pursue some data rather the ones listed below, please write email to me at yuany@math.pku.edu.cn).

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-inl, 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

2 CTR (Click-Through-Rate) Prediction in Bidding Algorithm

Original competition can be found from iPinYou Global Bidding Algorithm Competition at

`http://contest.ipinyou.com/`

where the full data (about 40GB) of 3 seasons can be downloadable at Baidu WebDrive

`http://pan.baidu.com/s/1kTkGUQN`

As part of the data, README file can be read from here:

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

For those who need a server, you may connect to the Linux account `einstein@162.105.68.237` which is public to the students in this class. Remember to make your own directory before starting creation of your own files. For example

1. `ssh einstein@162.105.68.237`
2. `INPUT your password`
3. `mkdir [your own directory]`

More information can be found in class notes at `www.ebanshu.com`.