A Mathematical Introduction to Data Analysis **28 April, 2017**

Final Project

Instructor: Yuan Yao *Due: midnight Sunday 21 May, 2017*

1 Final Project Requirement

This project aims to exercise the tools in the class based on the real world datasets. But you are encouraged to explore the datasets with any techniques you feel reasonable. In the below, we list some candidate datasets for your reference.

1. Pick up ONE (or more if you like) favorite dataset below to work. If you would like to work on a different problem outside the candidates we proposed, please email course instructor about your proposal.

2. Team work: we encourage you to form small team, up to THREE persons per group, to work on the same problem. Each team just submit ONE report, *with a clear remark on each person's contribution*. The report can be in the format of a *technical report within 8 pages*, e.g. NIPS conference style

   https://nips.cc/Conferences/2016/PaperInformation/StyleFiles

   or of a *poster*, e.g.

   http://math.stanford.edu/~yuany/publications/poster_CleaveBioCPH2017_ForReview.pptx

3. In the report, (1) design or raise your scientific problems (a good problem is sometimes more important than solving it); (2) show your main results with a careful analysis supporting the results toward answering your problems. Remember: scientific analysis and reasoning are more important than merely the performance results. Supplementary material such as source codes (and/or personal datasets) 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 paper version no later than the deadline, to the following address (datascience.hw@gmail.com) with Title: Math 6380: Final.

2 Ranking with Pairwise Comparisons

2.1 Drug Sensitivity Ranking in Kaggle in class contest

This contest is announced at

http://https://inclass.kaggle.com/c/drugsensitivity-3

with the invitation link at https://kaggle.com/join/math6380project3. This is a URL that can be shared around and re-used. Anyone who visits this link (and satisfies any other invitation rules that have been set) will be able to participate in the competition.

Given cell line k with genetic feature vector $X(k)$ and pairwise sensitivity comparison of drug i and j on k, $y(k, i, j)$, a basic generalized linear model can be

$$Prob(y(k, i, j) = 1) = F(\beta_0(i) - \beta_0(j) + \langle X(k), \beta_1(i) - \beta_1(j)\rangle). \tag{1}$$

In fact when $F = Id$, it leads to the following linear regression problem:

$$\min_{\beta_0,\beta_1\in\mathbb{R}^n} \quad \frac{1}{2}\sum_{(k,i,j)} (y(k, i, j) - (\beta_0(i) - \beta_0(j) + \langle X(k), \beta_1(i) - \beta_1(j)\rangle))^2 + P_\lambda(\beta) \tag{2}$$

For example, sparsity assumption on β_1 can be associated with LASSO penalty $P_\lambda(\beta) = \lambda\|\beta\|_1$.

Moreover, when $F(x) = e^x/(e^x + e^{-x})$ with binary response $y \in \{\pm 1\}$, it leads to the logistic regression, e.g. using the Maximum Likelihood principle one can find an estimate by

$$\min_{\beta_0,\beta_1\in\mathbb{R}^n} \quad \sum_{(k,i,j)} \log(1 + \exp(-y(k, i, j)f((k, i, j), \beta))) + P_\lambda(\beta) \tag{3}$$

where $f((k, i, j), \beta) = \beta_0(i) - \beta_0(j) + \langle X(k), \beta_1(i) - \beta_1(j)\rangle$.

A summary of generalized linear models in pairwise ranking problem, when β_1 is not presented, can be found in the following paper and references therein

http://math.stanford.edu/~yuany/publications/TMM12-final.pdf

2.2 WorldCollege Ranking

WorldCollege dataset is composed of 261 colleges. Using the Allourideas crowdsourcing platform, a total of 340 distinct annotators from various countries (e.g., USA, Canada, Spain, France, Japan) are shown randomly with pairs of these colleges, and asked to decide which of the two universities is more attractive to attend. The following dataset contains a total of 8,823 pairwise comparisons.

http://math.stanford.edu/~yuany/course/data/college.csv

This dataset has been used for various studies, e.g. Qianqian Xu, Jiechao Xiong, Xiaochun Cao, and Yuan Yao. False Discovery Rate Control and Statistical Quality Assessment of Annotators in Crowdsourced Ranking, ICML 2016, in http://arxiv.org/pdf/1605.05860v1.pdf

2.3 Human Age Ranking

The following dataset is kindly provided by Qianqian Xu, CAS, for the exploration on class.

The dataset is contained in the following zip file.

http://math.stanford.edu/~yuany/course/data/age.zip

where you may find

1. `readme.txt`: description of data
2. `Agedata.mat`: data file collected
3. `Groundtruth.mat`: Groundtruth
4. `30 images.zip`: 30 human age images

The basic problem is to *predict the human age*, using all the information collected so far. A simple sub-problem is rank aggregation of ages from pairwise comparisons. If you are interested, you can try some generalized linear models (Qianqian Xu, Qingming Huang, Tingting Jiang, Bowei Yan, Weisi Lin, and Yuan Yao. HodgeRank on Random Graphs for Subjective Video Quality Assessment. IEEE Transactions on Multimedia, 14(3):844-857, 2012, http://math.stanford.edu/~yuany/publications/TMM12-final.pdf) on this dataset, such as uniform model, Bradley-Terry model, Thurstone-Mosteller model, and Angular transform model. Compare maximum likelihood estimators and least square ones.

3 Identification of Raphael's paintings from the forgeries

The following data, provided by Prof. Yang WANG from HKUST,

https://drive.google.com/folderview?id=0B-yDtwSjhaSCZ2FqN3AxQ3NJNTA&usp=sharing

contains a 28 digital paintings of Raphael or forgeries. Note that there are both jpeg and tiff files, so be careful with the bit depth in digitization. The following file

https://docs.google.com/document/d/1tMaaSIrYwNFZZ2cEJdx1DfFscIfERd5Dp2U7K1ekjTI/edit

contains the labels of such paintings, which are

1 Maybe Raphael - Disputed

2 Raphael

3 Raphael

4 Raphael

5	Raphael
6	Raphael
7	Maybe Raphael - Disputed
8	Raphael
9	Raphael
10	Maybe Raphael - Disputed
11	Not Raphael
12	Not Raphael
13	Not Raphael
14	Not Raphael
15	Not Raphael
16	Not Raphael
17	Not Raphael
18	Not Raphael
19	Not Raphael
20	My Drawing (Raphael?)
21	Raphael
22	Raphael
23	Maybe Raphael - Disputed
24	Raphael
25	Maybe Raphael - Disputed
26	Maybe Raphael - Disputed
27	Raphael
28	Raphael

Can you exploit the known Raphael vs. Not Raphael data to predict the identity of those 6 disputed paintings (maybe Raphael)? The following student poster report seems a good exploration

http://math.stanford.edu/~yuany/course/2015.fall/poster/Raphael_LI%2CYue_1300010601.pdf

The following paper by Haixia Liu, Raymond Chan, and me studies Van Gogh's paintings which might be a reference for you:

http://dx.doi.org/10.1016/j.acha.2015.11.005

4 Co-appearance data in novels: Dream of Red Mansion and Journey to the West

A 374-by-475 binary matrix of character-event can be found at the course website, in .XLS, .CSV, .RData, and .MAT formats. For example the RData format is found at

http://math.stanford.edu/~yuany/course/data/dream.RData

with a readme file:

http://math.stanford.edu/~yuany/course/data/dream.Rd

as well as the .txt file which is readable by R command `read.table()`,

http://math.stanford.edu/~yuany/course/data/HongLouMeng374.txt

http://math.stanford.edu/~yuany/course/data/readme.m

Thanks to Ms. WAN, Mengting, who helps clean the data and kindly shares her BS thesis for your reference

http://math.stanford.edu/~yuany/report/WANMengTing2013_HLM.pdf

Moreover you may find a similar matrix of 302-by-408 for the Journey to the West (by Chen-En Wu) at:

http://math.stanford.edu/~yuany/course/data/west.RData

whose matlab format is saved at

http://math.stanford.edu/~yuany/course/data/xiyouji.mat

5 Jiashun Jin's data on Coauthorship and Citation Networks for Statisticians

Thanks to Prof. Jiashun Jin at CMU, who provides his collection of citation and coauthor data for statisticians. The data set covers all papers between 2003 and the first quarter of 2012 from the Annals of Statistics, Journal of the American Statistical Association, Biometrika and Journal of the Royal Statistical Society Series B. The paper corrections and errata are not included. There are 3607 authors and 3248 papers in total. The zipped data file (14M) can be found at

http://math.stanford.edu/~yuany/course/data/jiashun/Jiashun.zip

with an explanation file

http://math.stanford.edu/~yuany/course/data/jiashun/ReadMe.txt

With the aid of Mr. LI, Xiao, a subset consisting 35 COPSS award winners (https://en.wikipedia.org/wiki/COPSS_Presidents%27_Award) up to 2015, is contained in the following file

http://math.stanford.edu/~yuany/course/data/copss.txt

An example was given in the following article, A Tutorial of Libra: R Package of Linearized Bregman Algorithms in High Dimensional Statistics, downloaded at

http://math.stanford.edu/~yuany/course/reference/Libra_Tutorial_springer.pdf

The citation of this dataset is: *P. Ji and J. Jin. Coauthorship and citation networks for statisticians. Ann. Appl. Stat. Volume 10, Number 4 (2016), 1779-1812*, (http://projecteuclid.org/current/euclid.aoas)

6 NIPS paper datasets

NIPS is one of the major machine learning conferences. The following datasets collect NIPS papers:

6.1 NIPS papers (1987-2016)

The following website:

https://www.kaggle.com/benhamner/nips-papers

collects titles, authors, abstracts, and extracted text for all NIPS papers during 1987-2016. In particular the file `paper_authors.csv` contains a sparse matrix of paper coauthors.

6.2 NIPS words (1987-2015)

The following website:

https://archive.ics.uci.edu/ml/datasets/NIPS+Conference+Papers+1987-2015

collects the distribution of words in the full text of the NIPS conference papers published from 1987 to 2015. The dataset is in the form of a 11463 x 5812 matrix of word counts, containing 11463 words and 5811 NIPS conference papers (the first column contains the list of words). Each column contains the number of times each word appears in the corresponding document. The names of the columns give information about each document and its timestamp in the following format: `Xyear_paperID`.

7 Drug Sensitivity Datasets

7.1 One Drug data by Prof. Jiguang WANG@HKUST

The following dataset, provided by Prof. Jiguang WANG and Dr. Biaobin JIANG, contains 642 cancer cell line samples in response to one drug, Afatinib, with 60 binary features as gene mutation status. Column 1 is the sample ID, column 2 is the cell line names, column 3 is the real valued response as drug sensitivity measured by logarithmic IC50, and all the remaining 60 columns are the binary features/predictors. IC50 values are the dosage amounts of the drug such that after a period of time (say 24 hours) the experimental cancer cell lines are killed by 50%. Therefore the lower is the IC50, the more sensitive is the cancer line to the drug. Usually negative logarithmic IC50s are called *sensitive* while positive logarithmic IC50s are called *resistant*.

`http://math.stanford.edu/~yuany/course/data/OneDrug/OneDrug_train.csv`

A random subset of 100 cell line responses are withdrawn to the test sample. The basic problem is to predict the responses of these test samples based on their binary predictors/features. A sample prediction file is provided in

`http://math.stanford.edu/~yuany/course/data/OneDrug/OneDrug_submission.csv`

A Kaggle in-class contest is in the following website.

`https://inclass.kaggle.com/c/drugsensitivity-2`

The invitation link is: `https://kaggle.com/join/math6380project2`. This is a URL that can be shared around and re-used. Anyone who visits this link (and satisfies any other invitation rules that have been set) will be able to participate in the competition.

7.2 Combinatorial Drug 20 Efficacy Data by Prof. Ed Chow@NUS

The following dataset, provided by Prof. Ed K. CHOW at NUS, contains 120 cancer cell line samples in response to a combination of 20 drugs in 4 dosage levels. Column 1 is the sample ID, column 2 through column 21 list the discrete dosage levels of 20 drugs, column 22 is the real valued response as *viability* measured by difference between normalized cells and cancer cells. The higher is the viability, the more effective is the drug combination.

`http://math.stanford.edu/~yuany/course/data/Drug20/DrugEfficacy_train.csv`

A random subset of 20 cell line responses are withdrawn to the test sample. The basic problem is to predict the responses of these test samples based on their discrete levels of combinatorial drugs. Due to the expensive experiments in combinatorial drug efficacy measurements, the less features are in your models, the better. A sample prediction file is provided in

`http://math.stanford.edu/~yuany/course/data/Drug20/DrugEfficacy_submission.csv`

A Kaggle in-class contest is setup in the following website.

`https://inclass.kaggle.com/c/combodrug20`

The invitation link is: `https://kaggle.com/join/math6380edchow`. This is a URL that can be shared around and re-used. Anyone who visits this link (and satisfies any other invitation rules that have been set) will be able to participate in the competition.

7.3 Cleave Dataset

The following dataset is kindly provided by Cleave Co. Ltd. USA, for the exploration on class. **Please keep its use only in this class and any publication will be subject to the approval of Cleave.**

The dataset is contained in the following zip file (73M).

http://math.stanford.edu/~yuany/course/data/cleave.zip

where you may find

1. `data explanation.pptx`: description of data in pptx
2. `data for Yuan Yao.xlsx`: data file
3. `Gene set collection 1 for Yuan Yao.txt`: gene set collection
4. `Gene set collection 2 for Yuan Yao.txt`: gene set collection
5. `reference`: a folder contains a survey paper on 40+ machine learning algorithms as well as some source codes – *Nature Biotechnology 32, 1202–1212 (2014)* (http://www.nature.com/nbt/journal/v32/n12/full/nbt.2877.html)

The basic problem is to predict the drug response `IC50 within 72 hours`, using all the information collected so far, introduced by Ms. Lijing Wang with slides

http://math.stanford.edu/~yuany/course/2016.spring/cleave_lijing.pdf

as well as our CPH'2017 poster

http://math.stanford.edu/~yuany/publications/poster_CleaveBioCPH2017_ForReview.pdf

where the crucial discovery is that recursive variable selection by LASSO is more effective than one-stage LASSO.

In the mini-project 1, we launched a Kaggle in-class contest with this dataset at the following website,

https://inclass.kaggle.com/c/drugsensitivity

which is closed right now. But you may still see the leader board and submit one result for evaluations.

8 Heart PCI Operation Effect Prediction

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

http://math.stanford.edu/~yuany/course/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.

   The following two reports by LV, Yuan and LI, Xiao, respectively, might be interesting to you:

   http://math.stanford.edu/~yuany/course/reference/MSThesis.LvYuan.pdf

   http://arxiv.org/abs/1511.04656

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

   http://math.stanford.edu/~yuany/course/reference/MiaoLi2013S_project01.pdf

Since we did not separate a subset as test data, you may use k-fold cross validation (e.g. $k = 5$) to evaluate your models. In the mini-project 2, you may start from the prediction with all measurements, playing with various ways of missing value fill-in and feature selections. An interpretable model is preferred!

9 Finance Data

The following data contains 1258-by-452 matrix with closed prices of 452 stocks in SNP'500 for workdays in 4 years.

http://math.stanford.edu/~yuany/course/data/snp452-data.mat

or in R:

http://math.stanford.edu/~yuany/course/data/snp500.Rda

10 Hand-written Digits

The website

http://www-stat.stanford.edu/~tibs/ElemStatLearn/datasets/zip.digits/

contains images of 10 handwritten digits (‘0’,...,‘9’);

11 Animal Sleeping Data

The following data contains animal sleeping hours together with other features:

http://math.stanford.edu/~yuany/course/data/sleep1.csv

12 US Crime Data

The following data contains crime rates in 59 US cities during 1970-1992:

http://math.stanford.edu/~yuany/course/data/crime.zip