Mathematical Introduction to Data Science **December 20, 2012**

Final Project

Instructor: Yuan Yao *Due: Thursday Jan 3rd, 2013*

The final project encourages students to work in small team, but write the final report *independently*. Teammates may share their data, methodology, and results. But in the report, you have to

- *make your own analysis*;
- *give proper credit to the work of your partner.*

1 Topic Models

You can login my server:

```
ssh einstein@162.105.68.237
```

using the password I provided on class. Some data are provided there:

- `/data/twitter7/` contains

  June-Dec 2009 tweets, in `tweets2009-**.txt` files

  `twitter_rv.net` a who-follow-whom network ($9M$), whose id-name correspondence is in file `numeric2screen`
- `/home/einstein/ZHAOXIN/TweetMerged`, Zhao Xin's Singapore tweet data, one line consists of all the tweets of a user
- `/data/nytimes/` contains 50k articles about China in NewYork Times since 1989
- `/data/wikipedia/dbpedia` contains wikipedia articles

The following softwares are helpful

- `/home/einstein/ZHAOXIN/GibbsLDA++-0.2/src`, whose usage can be found at `http://gibbslda.sourceforge.net/`
- if you like matlab, you may try `http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm`

2 Big Matrix Factorization

For those who are interested to do matrix factorization (SVD etc.) with big matrices, the following paper

Divide-and-Conquer Matrix Factorization, by Lester Mackey, Ameet Talwalkar, Michael I. Jordan, version 6, Aug 2012. `http://arxiv.org/abs/1107.0789v6`

summarizes two important class of methods:

- random projection (including column projection)
- subsampling as generalized Nyström Method for SVD

You may implement one or two of these methods on your problem and test their efficiency.

3 Chess Player Rating

For those interested in Chess player rating in the Kaggle competition, the following data and algorithm are for your reference

- `/data/chess/data/` contains the following data

  `training_data.csv`: training data

  `cross_validation_dataset.csv`

  `test_scores.csv`: test data with game scores

  `players.csv` contains 8k chess players in the data

- `/data/chess/R/` contains two algorithms

  `elopp.R` is the kaggle award winner algorithm (MSRE=0.69), whose reference can be found at

  `/data/chess/Reference/`, or the paper: How I won the "Chess Ratings - Elo vs the Rest of the World" Competition, by Yannis Sismanis, 2010, `http://arxiv.org/abs/1012.4571`

  `hodgeRank1.R` is a prototype algorithm for hodgeRank by Dr. Quanwu Xiao.

4 Other topics

You can choose other datasets for the following tasks

- dimensionality reduction and estimation (PCA, MDS, ISOMAP, LLE, Diffusion, Laplacian, LTSA)

- spectral clustering
- RPCA and SPCA

In these tasks, explore the "criteria" to justify whether your method and choice of parameters is good or not.

5 Paper Reading Project

If you prefer to a solid paper reading project, please email me with your interested paper (if not listed on the class) or general interests. A paper presentation will be assigned to you.

Old Datasets

1. *Bird Flu Dataset:* (courtesy of Steve Smale and Cissy) This dataset 162 H5N1 (bird flu) virus sequences discovered around the world:

   `http://www.math.pku.edu.cn/teachers/yaoy/data/birdflu_seq162.txt`

   Locations of such virus discovered are reported with latitude and longitude coordinates on the globe:

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

   Pairwise geodesic distances between these 162 sites are constructed as

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

   A kernel-induced l_2-distances between 162 virus sequences are given in

   `http://www.math.pku.edu.cn/teachers/yaoy/data/birdflu_l2dist.txt`

2. *Chinese Medicine Dataset:* (courtesy of Zhi Geng) This dataset contains a 2885-by-965 matrix X with 2885 diseases (mostly flu) and 965 Chinese herbal prescription drugs collected from traditional Chinese medicine literature. Each element contains either 0 or 1 indicating if the herbal drug is used for the disease. Variable `desease2885` is a 1-by-2885 cell collecting the description of deseases and `herb965` contains the names of those herbal drugs.

   `http://www.math.pku.edu.cn/teachers/yaoy/data/cmed965.mat`

3. *Journey to the West:* This dataset contains a 302-by-408 matrix X with 302 characters and 408 scenes collected from one hundred chapters in the classic novel by WU, Cheng-En. Each element contains either 0 or 1 indicating if the character appears in the scene.

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

   The construction of this matrix from original data can be found in Matlab file:

   `http://www.math.pku.edu.cn/teachers/yaoy/data/xiyouji/readData.m`

4. *Hand-written Digits:* The website

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

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

5. *S&P500 Prices:* This dataset contains a data matrix $X \in \mathcal{R}^{n\times p}$ of $n = 1258$ consecutive observation days and $p = 452$ daily closing stock prices, and the cell variable "stock" collects the names, codes, and the affiliated industrial sectors of the 452 stocks.

   `http://www.math.pku.edu.cn/teachers/yaoy/data/snp452-data.mat`

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

6. *SNPs of World-wide Populations:* This dataset contains a data matrix $X \in \mathcal{R}^{p\times n}$ of about $n = 650,000$ columns of SNPs (Single Nucleid Polymorphisms) and $p = 1064$ rows of peoples around the world. Each element is of three choices, 0 (for 'AA'), 1 (for 'AC'), 2 (for 'CC'), and some missing values marked by 9.

   `http://www.math.pku.edu.cn/teachers/yaoy/data/ceph_hgdp_minor_code_XNA.txt.zip`

   Moreover, the following file contains the region where each people comes from, as well as two variables `ind1` and`ind2` such that $X($`ind1`, `ind2`$)$ removes all missing values.

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

   Some results by PCA can be found in the following paper, Supplementary Information.

   `http://www.sciencemag.org/content/319/5866/1100.abstract`

   Attention: this last dataset is relatively big with about 2GB size.