Mathematical Introduction to Data Science **Dec 25, 2013**

Final Project

Instructor: Yuan Yao *Due: Sunday Jan 12, 2014*

The final project encourages you to work in small teams, with each team consisting no more than THREE persons and submitting one report before midnight of Sunday Jan 12, 2014. In your report, please state clearly *the contributions of each teammate*. Send your submission to me (`yyao04@gmail.com`), cc-ed to teaching assistants.

1 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. SNPs data is a good candidate for example.

2 Chess Player Rating

You can login my server:

```
ssh einstein@162.105.205.92
```

using the password I provided on class.

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.

Note: please create your own folder at /home/einstein/ if you would like to work on the server. Otherwise download the data and codes to your local computer.

3 World College Rating

- `/data/worldcollege` contains the following data

  `export_4271_ideas_20131117.csv`: 261 colleges in the world

  `export_4271_votes_20131128_n5k.csv`: about 5000 votes up to Nov 28, 2013

  `export_4271_non_votes_20131128.csv`: non-votes marked as "I can't decide" with various reasons

Explore this data with Hodge decomposition of paired comparison data. No one has ever tried it yet.

4 Protein Folding

There are two problems related to the following datasets.

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

- prediction of *contact map* by multiple aligned sequences in the same family;
- 3D structure reconstruction based on incomplete MDS with uncertainty.

In the file, you will find PF00013 (PCBP1_HUMAN/281-343, PDB 1WVN), PF00018 (YES_HUMAN/97-144, PDB 2HDA), and PF00254 (O45418_CAEEL/24-118, PDB 1R9H). Data format information can be found at

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

For example, file `http://www.math.pku.edu.cn/teachers/yaoy/data/protein/sequence/PF00018_match.aln` contains 3610 sequences of length 48 for the same family PF00018, where the first sequence is

`----------------ENEIVQVFSIVDESWWSGKLRRNGAEGIFPK`

Here

- `-` denotes the gap,
- other alphabets denotes the Amino Acid code, from 20 characters.

Therefore in total the sequence is coded by 21 characters. Correspondingly file `http://www.math.pku.edu.cn/teachers/yaoy/data/protein/sequence/PF00018_2HDA.pdb` contains the 3D coordinates of alpha-carbons for a particular amino acid sequence in the family, YES_HUMAN/97-144, read as

`VALYDYEARTTEDLSFKKGERFQIINNTEGDWWEARSIATGKNGYIPS`

where the first line in the file is

97 V 0.967 18.470 4.342

Here

- '97': start position 97 in the sequence
- 'V': first character in the sequence
- $[x, y, z]$: 3D coordinates in unit $\mathring{A}$.

Figure 1 gives the 3D representation of its structure.

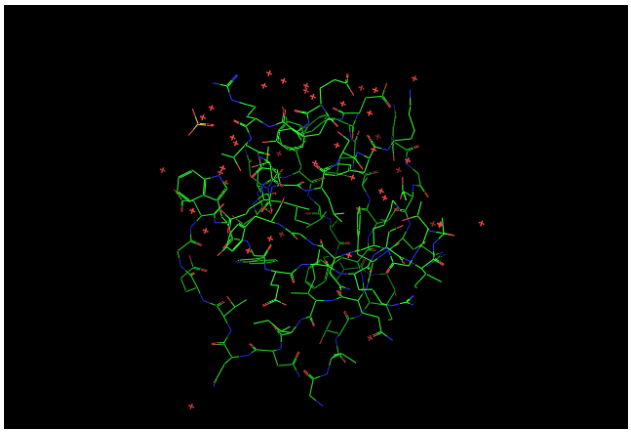

Figure 1: default

Given the 3D coordinates of the amino acids in the sequence, one can computer pairwise distance between amino acids, $[d_{ij}]^{l\times l}$ where l is the sequence length. A *contact map* is defined to be a graph $G_\theta = (V, E)$ consisting l vertices for amino acids such that and edge $(i, j) \in E$ if $d_{ij} \leq \theta$, where the threshold $\theta = 8\mathring{A}$ here.

Non-local contact map $G_{\theta,\tau}$ considers the restricted contact map with only edges (i, j) with i and j are τ-separated way in sequence distance. Here we choose $\tau = 5$, *i.e.* $|i - j| > 5$.

In contact map prediction, we are going to learn a graphical model from multiple aligned sequences, to predict the non-local contact map $G_{\theta=8\mathring{A},\tau=5}$. Performance is evaluated in terms of the fraction of correct predicted non-local contacts (true-positive-rates) among the top k pairs with highest scores, *e.g.* $k = l/5, l/3, l/2, l$, etc. Figure **??**, courtesy by Chendi Huang, gives you a reference on comparing the Directed Information by Morcos and the graphical lasso. For your reference, Chendi's report can be found at

`http://www.math.pku.edu.cn/teachers/yaoy/reference/Huang_protein_report_2013-04-28.pdf`

and a recent paper by Steve Smale's group is found at

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

In 3D structure recovery, we can try to find the 3D coordinates given in .pdb files, assumed that we are given incomplete pairwise distances contaminated by noise. We can construct such data based on information given in .pdb files.

5 iPinyou Global Bidding Algorithm Competition

Original data can be downloaded from iPinYou Global Bidding Algorithm Competition at

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

where Stage III ended last week already. To register the contest and submit your algorithm, you need to fill in your algorithm in a Java interface found at

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

You may still pursue the CTR (click-through-rate) prediction using the data provided by iPinyou. For example, you may pursue online (stochastic gradient descent) logistic regression for CTR, with other variations like linearized Bregman iteration or mirror descent.

6 Spectral Methods

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

This project aims to exercise the tools in the class, such as random projections, robust PCA, sparse PCA, and MDS with uncertainty, etc., based on the real datasets. In the below, we list some candidate datasets for your reference.

1. Choose the dataset in your favorite. In your report, raise a problem that you are interested to attack with the tools you learned. The following problems are for your examples:

   Robust PCA will decompose the data matrices into two parts, a low-rank and a sparse, which can be applied to all the datasets below. What's the meaning of sparse part and what's the meaning of low-rank part in your data?

   Sparse PCA will find PCA whose support are only subsets of the variables. Journey to the West and Chinese Medicine datasets are probably suitable for such a tool, as we would like to find small teams of characters dominating the story, or small sets of essential herb curing most of similar deseases (flu).

   Random Projection is to do dimensionality reduction if one wants to find SVD of X of large dimensionality. This particularly suits the analysis of SNPs data with $650k$ sites, where you can random projects the matrix into a subset of SNPs and see if top eigenvectors are still approximately correct.

   MDS with uncertainty is to reconstruct Euclidean coordinates when pairwise distances are partially observed and contaminated. Can one reconstruct Euclidean coordinates of Bird Flu virus from geodesic distances and sequence distances, respectively, and compare one against another?

2. Try to apply the methods to analyze your data. Analyze the phenomena you observed and explain whether or not the results answer your problems in the first step and *why*. Finally in your report list the kind of *open problems* in your project and possible future directions in your mind.

7 Paper Reading Project

If you prefer to a solid paper reading project, please email me with your interested paper and we can assign some paper to you.

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. *Protein Folding Data:* (courtesy of Steve Smale)

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

3. *A Dream of Red Mansion:* This dataset contains a 376-by-475 matrix X with 376 characters and 475 scenes collected from 120 chapters in the classic novel by CAO, Xueqin. Each element contains either 0 or 1 indicating if the character appears in the scene. The construction of this matrix from original data can be found in Matlab file:

   `http://www.math.pku.edu.cn/teachers/yaoy/data/hongloumeng/readme.m`

   Thanks to Ms. WAN, Mengting (now at UIUC), an update of data matrix consisting 374 characters (two of 376 are repeated) which is readable by R read.table() can be found at

   `http://www.math.pku.edu.cn/teachers/yaoy/data/hongloumeng/HongLouMeng374.txt`

   She also kindly shares her BS thesis for your reference

   `http://www.math.pku.edu.cn/teachers/yaoy/reference/WANMengTing2013_HLM.pdf`

   Among various choices of analysis, with this data matrix X, you may form a weighted graph $W = X * X'$, pursue PCA of X.

4. *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`

5. *Hand-written Digits:* The website

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

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

6. *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`

   You may use PCA to explore the 'invisible hands' of markets.

7. *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 $X(\texttt{ind1}, \texttt{ind2})$ 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.