BIOSTAT M280, Homework 3

Due Tue, Feb 11, 2016

1. (a) Read in the 'longley.dat' with the response (number of people employed) in the first column and six explanatory variables in the other columns (GNP implicit price deflator, Gross National Product, number of unemployed, number of people in the armed forces, 'noninstitutionalized' population $\geq$ 14 years of age, year). Include an intercept in your model.

   (b) Assuming linear model $\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$, compute the regression coefficients $\hat{\boldsymbol{\beta}}$, their standard errors, and variance estimate $\hat{\sigma}^2$ using following methods: QR decomposition, Cholesky decomposition, and sweep operator. Please compute them directly using numerical linear algebra functions; you can use the "black-box" function `lm()` only to check your results. (Hint: `chol2inv()` function computes the inverse of a matrix from its Cholesky factor.)

   (c) Find out which method is the `lm()` function in R using? And which algorithm is being used?

   (d) One popular regularization method is the ridge regression, which estimates regression coefficients by minimizing a penalized least squares criterion

   $$\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X\beta}\|_2^2 + \frac{\lambda}{2}\|\boldsymbol{\beta}\|_2^2.$$

   Show that the ridge solution is given by

   $$\hat{\boldsymbol{\beta}}_\lambda = (\boldsymbol{X}^\intercal\boldsymbol{X} + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{X}^\intercal\boldsymbol{y}.$$

   (e) Show that the ridge estimator is equivalent to the solution of a regular least squares problem with added observations. Compute the ridge regression estimates $\hat{\boldsymbol{\beta}}_\lambda$ at two different values of λ by solving this augmented least squares problem. You can use any method of your choice (QR, Cholesky, or sweep).

2. In this question, you are going to try different numerical methods learnt in class on the vanilla Google PageRank problem.

   (a) Let $\boldsymbol{A} \in \{0,1\}^{n\times n}$ be the connectivity matrix of n web pages with entries

   $$a_{ij} = \begin{cases} 1 & \text{if page } i \text{ links to page } j \\ 0 & \text{otherwise} \end{cases}.$$

   $r_i = \sum_j a_{ij}$ is the out-degree of page i. That is r_i is the number of links on page i. Imagine a random surfer exploring the space of n pages according to the following rules.

   - From a page i with $r_i > 0$
     - with probability p, (s)he randomly chooses a link on page i (uniformly) and follows that link to the next page

  - with probability $1 - p$, (s)he randomly chooses one page from the set of all n pages (uniformly) and proceeds to that page

- From a page i with $r_i = 0$ (a dangling page), (s)he randomly chooses one page from the set of all n pages (uniformly) and proceeds to that page

The process defines a Markov chain on the space of n pages. Write down the transition matrix $\boldsymbol{P}$ of the Markov chain (in matrix/vector notation).

(b) According to standard Markov chain theory, the (random) position of the surfer converges to the stationary distribution $\boldsymbol{x} = (x_1, \ldots, x_n)^T$ of the Markov chain. x_i has the natural interpretation of the proportion of times the surfer visits page i in the long run. Therefore $\boldsymbol{x}$ serves as page ranks; a higher x_i means page i is more visited. It is well-known that $\boldsymbol{x}$ is the left eigenvector corresponding to the top eigenvalue 1 of the transition matrix $\boldsymbol{P}$. That is $\boldsymbol{P}^T\boldsymbol{x} = \boldsymbol{x}$. Therefore $\boldsymbol{x}$ can be solved as an eigen-problem. Show that it can also be cast as solving a linear system. Remember $\boldsymbol{x}$ is a distribution so we normalize it to have $\sum_{i=1}^n x_i = 1$.

(c) Download the `ucla.zip` package from course webpage. Unzip the package, which contains two files `U.txt` and `A.txt`. `U.txt` lists the 500 URL names. `A.txt` is the 500×500 connectivity matrix retrieved on Feb 1, 2016. Read data into R. Compute summary statistics:

- number of pages
- number of edges
- number of dangling nodes
- max in-degree
- max out-degree
- visualize sparsity pattern of $\boldsymbol{A}$

(d) Set the *teleportation* parameter at $p = 0.85$. Try the following methods to solve for $\boldsymbol{x}$ using the `ucla.zip` data.

i. A dense linear system solver such as LU decomposition

ii. A simple iterative linear system solver such as Jacobi or Gauss-Seidel

iii. A dense eigen-solver

iv. A simple iterative eigen-solver such as the power method

(e) List the top 20 ranked URLs you found.

(f) As of Monday Feb 1 2016, there are at least 4.84 billion indexed webpages on internet according to `http://www.worldwidewebsize.com/`. Comment on whether each of these methods may or may not work for the PageRank problem at this scale.