STA 250 Lecture 8 note

Chia-Pei Chen

I. Types of Big Data

- Large n and not large p
  - We focus on this. Example: Linear regression with 100m observations and 500 covariates.
- Large p and not large n
- Large p and large n
- Complex (non-rectangular) data.
  - Example: Brain images.

II. Scale to Big Data

- Assume lower dimension: sparsity, conditional independence. (If correlated, prefer to do joint distribution)
- Fast algorithm: parallelize, linear time
- Avoid to fit full data: consensus Monte Carlo, bag of little boostraps. (Focus on bag of little boostraps)
- Complex (non-rectangular) data. [**Note**: hundred megabytes in size can cause R shutdown.]

III. Alternative Methods for Big Data

(i) File-backed data structure: avoid reading data in the memory and store on disk. Ex: bigmemory: easy to use.

(ii) Databases:

  - Relational database (SQL): Rigid and relational structure

– NoSQL database (CouchDB):Less structure and functionality

(iii) Distributed file system

– ex: Hadoop distributed file system (HDFS): across multiple machines
– Pros:duplicate data, stronger compute power and speed up
– Cons:hard to interact with data

We focus on (i) and (iii).

IV. Example of Big logistic regression

Goal:Find stand errors for parameter estimates and how to work with big data by "bigmemory"

(i) Use bigmemory concept: read some arbitrary lines instead of full file

(ii) Find CI's or SE's for $\hat{\beta}$: We can use bootstrap For the logistic regression problem, using $B = 500$:

1. Let $\hat{F}$ denote the true probability distribution of the data (i.e., placing mass 1/6000000 at each of the 6000000 data points)
2. Take a random sample of size 6000000 from $\hat{F}$ (with replacement). Call this a "bootstrap dataset", X_j^* for $j = 1, \cdots, 500$.
3. For each of the 500 bootstrap datasets, compute the estimate $\hat{\beta}_j^*$.
4. Use the standard deviation of $\{\hat{\beta}_1^*, \cdots, \hat{\beta}_{500}^*\}$ to approximate $SD(\hat{\beta})$.

Traditional Bootstrap (resample with replacement again and again) takes longer time. We can consider using the following algorithm.

(iii) The Bag of Little Bootstraps

Sample s subsets of size $b < n$ and the resample n points from those.
For estimating SD($\widehat{\beta}$):

(1) Let $\widehat{F}$ denote the empirical probability distribution of the data (i.e., placing mass $1/n$ at each of the n data points)
(2) Select s subsets of size b from the full data (i.e. randomly sample a set of b indices $\{I\}_j = \{i_1, ..., i_b\}$ from $\{1, 2, ..., n\}$ without replacement, and repeat s times)
(3) For each of the s subsets ($j = 1, ..., s$): Repeat the following steps r times ($k = 1, ..., r$):
  * Resample a bootstrap dataset $X_{j,k}^*$ of size n from subset j. (i.e., sample $(n_1, , n_b) \sim Multinomial(n, (1/b, , 1/b))$, where $(n_1, , n_b)$ denotes the number of times each data point of the subset occurs in the bootstrapped dataset.)
  * Compute and store the estimator $\widehat{\theta}_{j,k}$

* Compute the bootstrap SE of $\widehat{\theta}$ based on the r bootstrap data sets for subset j i.e., compute:

$$\xi_j^* = SD\{\widehat{\theta^*}_{j,1}, ..., \widehat{\theta^*}_{j,1}\}$$

(4) Average the s bootstrap SE's, $\xi_1^*, \cdots, \xi_s^*$ to obtain an estimate of $SD(\hat{\theta})$ i.e.,

$$\widehat{SD}(\hat{\theta}) = \frac{1}{s}\sum_{j=1}^{s} \xi_j^*.$$

(5) More to think about:

* How to select s? (Number of subsets)
* How to select b? (Subset sample size)
* How to select r? It is better that r (Number of bootstrap replicates per subset)to be large enough for each of the s subsets ($r > s$). For example, if 500 subjects, then $r = 50$ and $s = 10$ or $s = 5$. Real key is b. From paper $b \approx n^{0.6}$ or $b \approx n^{0.7}$ works well.
* Reducing the unique data points in each data set can help speed things up.
* How to utilize the array job capability of Gauss for BLB?