RGallery: A Package for 3 questions in Stochastic Average Gradient Project

Eric Xin Zhou

February 18, 2015

1 Easy level

Q1 : Use glmnet to fit an L2-regularized logistic regression model. Use the system.time function to record how much time it takes for several data set sizes, and make a plot that shows how execution time depends on the data set size.

1.1 Data simulation setup for L2

Given that test 1 need us provide different data set of different size to record how much time **glmnet** takes for different data size.

I generate Gaussian data with N observation and p predictors. with each pair of predictors $X_j, X_{j'}$ has the same population correlation ρ. If N and ρ are determined. We generate the observed data Y by adding several gaussian noise.

$$Y = \sum_{j=1}^{p} X_j \beta_j + kZ \tag{1}$$

If Y is a $N \times 1$ column vector, then $X_j, X_{j'}$ are all $N \times 1$ column vectors, so $\mathbf{X}$ is a $N \times p$ matrix and β is a $p \times 1$ column vector.

Z represents noise of observation, and k is chosen so that we can control signal-to-noise ratio to 3.0.

In generation model, we also should simulate the coefficient vector β, we define that

$$\beta_j = (-1)^j \exp\left(\frac{-2(j-1)}{20}\right) \tag{2}$$

This guarantee that the coefficients are constructed to have alternating signs and to be exponential descreasing.

And in logistical regression model, what we observation is $\mathcal{G} = \{1, 2\}$, therefore, the logistic regression model represents the probability we observed $\{1, 2\}$.

$$
\begin{aligned}
Pr(G=1|x) &= \frac{1}{1+e^{-(\beta_0 x_0+\ldots+\beta_p x_p)}} \\
Pr(G=2|x) &= \frac{e^{-(\beta_0 x_0+\ldots+\beta_p x_p)}}{1+e^{-(\beta_0 x_0+\ldots+\beta_p x_p)}}
\end{aligned} \tag{3}
$$

So we can get the column $\log\left(\frac{Pr(G=1|X)}{Pr(G=2|X)}\right) = \sum_{j=1}^{p} x_j\beta_j$. In this model, if $Pr(G=1|x) > Pr(G=2|x)$, then response is actual 1, otherwise responese is 2.

However, in real observation, noise will be introduced into this regression model. As the result, we should add an $Z \sim \mathcal{N}(0,1)$ into the original LR model.

So, the response Y we generate comes from Eqn.(4)

$$
y = \sum_{j=1}^{p} X_j\beta_j + kZ \tag{4}
$$

And we also can tell that $y \to \beta_0 x_0 + \ldots + \beta_p x_p$, therefore, we can determine $Pr(G=1|x) = \frac{1}{1+e^{-y}}$. And then we will define our observation Y by binomial model which generate Y with $Pr(Y=1) = p$ and $Pr(Y=2) = 1-p$.