Project of MATP6600/ISYE6780

(Due mid-night on Dec-7-2018)

Problem description

Logistic regression is one popular model used in statistics and machine learning. As discussed at the beginning of this semester, it can be derived by the maximum log likelihood. Given training data $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ with $y_i \in \{+1, -1\}$ for each $i = 1, \ldots, N$, the model can be formulated as

$$\min_{\mathbf{w},b} f(\mathbf{w}, b) := \frac{1}{N}\sum_{i=1}^{N} \log\left(1 + \exp[-y_i(\mathbf{w}^\top \mathbf{x}_i + b)]\right) + \frac{\lambda_1}{2}\|\mathbf{w}\|_2^2 + \frac{\lambda_2}{2} b^2, \tag{1}$$

where $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ are strong convexity constants. Note that one can set both of λ_1 and λ_2 to zero, but choosing positive λ_1 and λ_2 can guarantee positive definiteness of the Hessian matrix. Suppose $(\bar{\mathbf{w}}, \bar{b})$ is a solution of (1). Then a new data point $\mathbf{x}$ can be classified as positive if $\mathbf{w}^\top \mathbf{x} + b \geq 0$ and negative if $\mathbf{w}^\top \mathbf{x} + b < 0$.

Requirements

Among the following three items, you are required to do the first two. If you also do the third one correctly and efficiently, you can earn up to 100% bonus credits. Write a report to include your code and all results you obtain. Send your report and also the source code (in a single .zip file) to the email address `optimization.rpi@gmail.com`

1. Write two solvers with input $(\mathbf{X}, \mathbf{y}, \lambda_1, \lambda_2)$ where the i-th column of $\mathbf{X}$ is the i-th sample data, the i-th component of $\mathbf{y}$ is the corresponding label, and $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ are the two parameters in the model (1). One solver is the steepest gradient descent, and another solver is the Newton's method. For either of the solvers, any appropriate termination rule is accepted.

2. On the two provided data sets and setting $\lambda_1 = \lambda_2 = 0.001$, compare the two solvers with the same termination rule as follows:

$$\|\nabla f(\mathbf{w}^k, b^k)\| \leq \varepsilon \max\left(1, \|(\mathbf{w}^k, b^k)\|\right), \tag{2}$$

where k is the iteration number, and $\varepsilon > 0$ is the relative stopping tolerance. For $\varepsilon \in \{10^{-2}, 10^{-4}, 10^{-6}\}$, report the total iteration numbers, the total running time for each solver, and their classification accuracy on the testing data. Also report your observations.

3. In item 2, change the values of λ_1 and λ_2 and see how they affect the classification accuracy. Develop another solver by the coordinate gradient descent method or the stochastic gradient method. Compare the third solver to the previous two solvers on the two provided data sets with the same settings as in item 2. Report their total iteration numbers and running time, and also report your observations.