分布式存储与计算 第四次作业

提交时间：2022 年 11 月 29 日 18:30

考虑如下的线性回归模型:

$$y_i = \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i, \qquad i = 1, \dots, N$$

其中$x_{ij} \sim N(0,1)$和 y 分别是解释变量和响应变量，而$\epsilon_i \sim N(0,0.04)$是模型误差。

当$j = 1,2,3$时，$\beta_j = 2$；当$j > 3$时，$\beta_j = 0$. $N = 1000, p = 200$.

1. 试编写 Scala 程序求上述模型的 LASSO 估计量:

$$\hat{\beta}(\lambda) = \arg\min_{\beta} \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1$$

   其中λ是调试参数，$\|\cdot\|_1$表示L_1模。

   在本次作业中，我们取$\lambda = 0.85 \times 10^{-12}$. $\beta = (\beta_1, \beta_2, \cdots, \beta_p)$的初值取它的最小二乘估计量。最大迭代次数为 6 次。

2. 将问题 1 中的代码使用 MapReduce 方法进行分布式改进。