L2: Hadoop MapReduce 与统计建模

Feng Li

feng.li@cufe.edu.cn

School of Statistics and Mathematics
Central University of Finance and Economics

Revised on December 6, 2018

本节知识要点

1. **Hadoop Streaming**
2. 利用 **Hadoop** 估计一个线性回归模型的系数
3. 逻辑回归模型

Hadoop Streaming

- Hadoop 提供了一个 API （Hadoop Streaming）允许用户使用非 JAVA 语言写 MapReduce 程序。
- Hadoop Streaming 使用 Unix 标准流作为接口在 Hadoop 和你的程序之间交流数据，任何可以读取标准输入 (stdin) 并写入标准输出 (stdout) 的程序都可以编写 MapReduce 程序。
- 完整的 Hadoop Streaming 帮助文档可以从这里获取 https://hadoop.apache.org/docs/r2.7.2/hadoop-streaming/HadoopStreaming.html

Hadoop Streaming

- 演示示例

其他 Hadoop 接口

- 除了以上介绍的 Hadoop Streaming 以外，还有支持 C++ 语言的 Hadoop Pipe 的接口。这里不做过多介绍, 感兴趣的读者可以参考 Hadoop 官方文档 https://hadoop.apache.org/docs/r1.2.1/api/org/apache/hadoop/mapred/pipes/package-summary.html

线性回归模型 I

- 如果你有 10T 的文件，如何在分布式系统上做一个线性模型?
- 假设我们的 $y_{n\times 1}$ 和 $X_{n\times p}$ $(n > p)$ 是分布式存储在一起的.
- 回顾一下线性模型的

$$y = X\beta + \epsilon$$

的最小二乘的解 $\widehat{\beta}$

$$\widehat{\beta} = (X'X)^{-1}X'y$$

- 大数据下的困扰:
  - X 和 y 是分布式存储的。
  - 计算 $X'X$ 和 $X'y$ 是整个计算的关键。
  - 同时我们注意到 $(X'X)_{p\times p}$ 和 $(X'y)_{p\times 1}$ 的维度就会非常的小.

线性回归模型 II

- 可行的解决方案:
  - 让 Hadoop 去计算 $X'X$ 和 $X'y$.
  - 最终结果可以通过简单的组合来实现.
- 技术细节:
  - 首先我们从一个简单的入手:

$$X'y = \begin{bmatrix} x_{1.} \\ x_{2.} \\ \dots \\ x_{k.} \end{bmatrix}' \begin{bmatrix} y_{1.} \\ y_{2.} \\ \dots \\ y_{k.} \end{bmatrix} = \sum_{i=1}^{k} x_{i.}' y_{i.} \tag{1}$$

  - 然后

$$X'X = X' \begin{bmatrix} x_{.1} & x_{.2} & \dots & x_{.l} \end{bmatrix} = \begin{bmatrix} X'x_{.1} & X'x_{.2} & \dots & X'x_{.l} \end{bmatrix} \tag{2}$$

逻辑回归模型 I

- 逻辑回归事实上是大数据行业的基础模型。
- **Bad news**: 逻辑回归系数的估计依赖 Hadoop 并不擅长的迭代算法。
- 最常见的 Hadoop 解决方案是利用随机梯度下降算法 (Stochastic Gradient Descent, SGD) 来随机使用一小部分数据来近似优化算法的梯度。
- Spark 等软件提供和现成的解决方案。

上机实践

- 利用 Hadoop 对数据做一个简单的描述统计
- 实现一个简单的矩阵乘法
- 进阶：实现一个回归参数的估计方法。