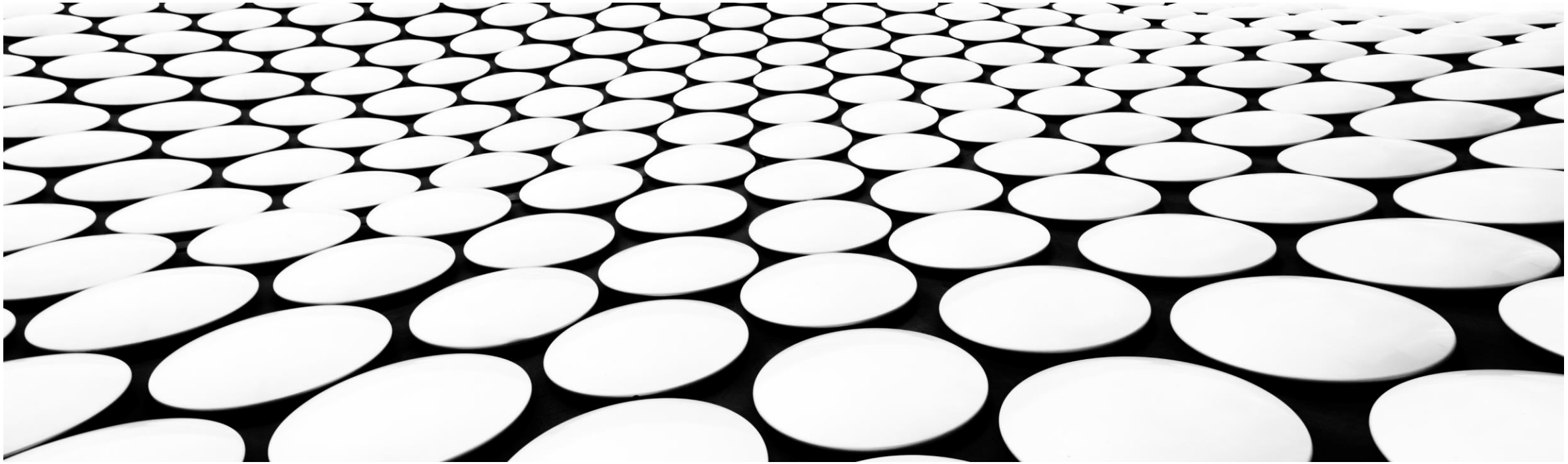
分布式计算

邱怡轩



今天的主题

- 基础分布式算法

分布式算法

常见问题

- 矩阵乘法
- 解线性方程组
- 线性模型
- 岭回归
- Logistic 回归
- 梯度下降法
- 牛顿法
- ……

常见问题

- 矩阵乘法
- 解线性方程组
- 线性模型
- 岭回归
- Logistic 回归
- 梯度下降法
- 牛顿法
- ……

矩阵乘法

矩阵乘法

- 过于简单?

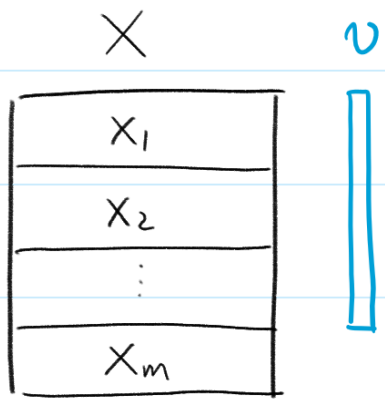
矩阵乘法

- Xv
- $X'X$
- $X'v$

矩阵乘法

- 分布式计算 Xv

① Xv　　$X \in \mathbb{R}^{n \times p}$，$v \in \mathbb{R}^{p}$



$x_i \in \mathbb{R}^{n_i \times p}$　　$x_i v \in \mathbb{R}^{n_i}$

$$Xv = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix} v_{p\times 1} = \begin{pmatrix} x_1 v \\ \vdots \\ x_m v \end{pmatrix}$$

矩阵乘法

- 分布式计算 $X'X$

② $X^{'}X$ $\quad X \in \mathbb{R}^{n \times p}$

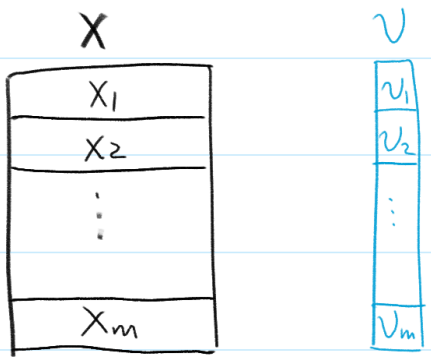
$x_i \in \mathbb{R}^{n_i \times p}$

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_m \end{pmatrix} \qquad X^{'}X = \begin{pmatrix} X_1^{'} & \cdots & X_m^{'} \end{pmatrix} \begin{pmatrix} X_1 \\ \vdots \\ X_m \end{pmatrix} = X_1^{'}X_1 + \cdots + X_m^{'}X_m$$

矩阵乘法

- 分布式计算 $X'Y$

③ $X'v$ $\quad X \in \mathbb{R}^{n \times p}$, $v \in \mathbb{R}^{n}$



$X_i \in \mathbb{R}^{n_i \times p}$ $\quad v_i \in \mathbb{R}^{n_i}$ $\quad X_i' v_i \in \mathbb{R}^{p}$

$$X'v = \begin{pmatrix} X_1' & \cdots & X_m' \end{pmatrix} \begin{pmatrix} v_1 \\ \vdots \\ v_m \end{pmatrix} = X_1'v_1 + \cdots + X_m'v_m$$

实现

- lec7-matprod.ipynb

线性回归

线性回归

- 考虑回归问题 $y = \beta_0 + \beta' x + \varepsilon$
- $n \gg p$
- 回归系数估计值的表达式为

$$\hat{\beta} = (X'X)^{-1}X'Y$$

- 注意还需考虑截距项

解决思路

- 当 p 不太大时
- $X'X$ 和 $X'Y$ 都可以装进内存

1. 从原始数据生成 RDD
2. 分别计算 $X'X$ 和 $X'Y$
3. 解线性方程组，得到 $\hat{\beta} = (X'X)^{-1}X'Y$

读取数据

- 首先了解数据的存储格式
  - 有没有表头?
  - 数据按什么分隔?
  - Y 和 X 的位置如何?

其他细节

- 要保证 X 和 Y 始终处在同一个 RDD 中
- 将 X 和 Y 作为一个整体进行 RDD 分区
- 添加截距项（如何操作？）

实现

- lec7-regression.ipynb

线性回归

- 当 $n < p$ 时
- $X'X$ 不再可逆
- 最小二乘没有唯一解