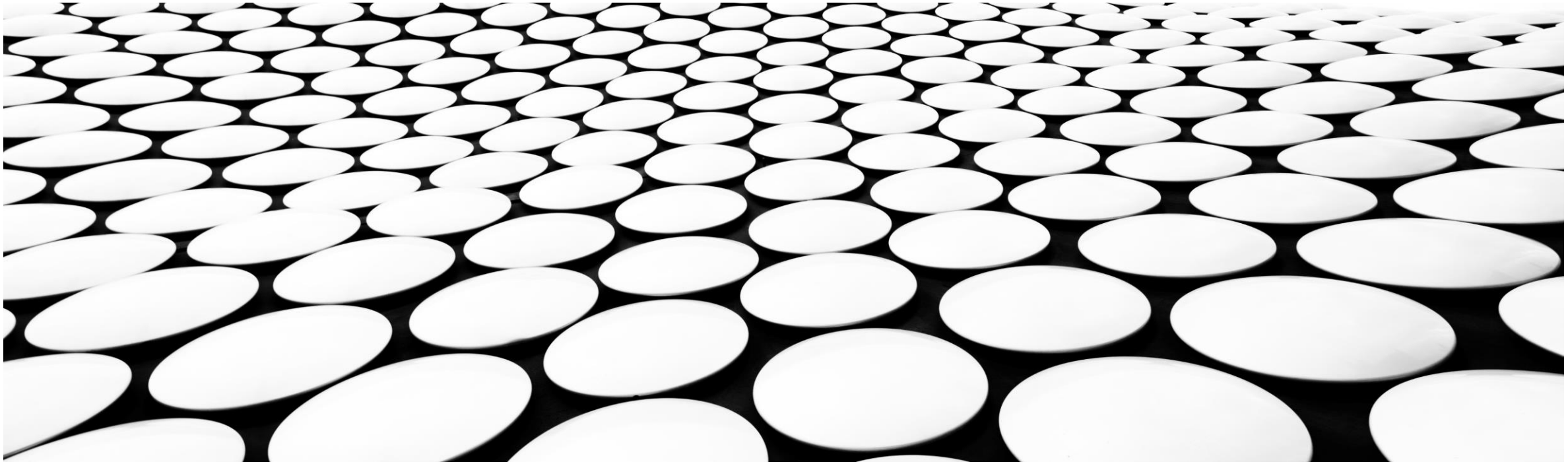
分布式计算

邱怡轩



今天的主题

- 共轭梯度法
- 分布式 Logistic 回归

线性回归

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

岭回归

岭回归

- 依然是回归问题 $y = \beta_0 + \beta' x + \varepsilon$
- 当 $n \ll p$ 时，$X'X$ 不可逆
- 此时在损失函数上加入惩罚项 $\lambda \|\beta\|^2$
- 回归系数估计值的表达式为

$$\hat{\beta}_\lambda = (X'X + \lambda I)^{-1} X'Y$$

- λ 为一个给定的正数

岭回归

- 注意当 $n \ll p$ 时， $X'X + \lambda I$ 是一个高维的矩阵，无法直接解线性方程组
- 引入共轭梯度法

共轭梯度法

线性方程组

- 考虑线性方程组 $Ax = b$
- 假设 A 是正定矩阵
- 正定：A 的特征值都大于0
- 求解 $x = A^{-1}b$ 总共分几步?

线性方程组

- 考虑线性方程组 $Ax = b$
- 假设 A 是正定矩阵
- 正定：A 的特征值都大于0
- 求解 $x = A^{-1}b$ 总共分几步?

- p 步，某种意义上
- p 是 A 的维度（A 是方阵）

共轭梯度法

- 共轭梯度法（Conjugate gradient，CG）是一种解正定线性方程组的方法
- 它有趣的地方在于，可以通过乘法运算 $v \to Av$ 来得到逆运算结果 $A^{-1}b$
- 更有意思的是，数学上可以证明它在 p 步迭代之后就可以得到精确解

共轭梯度法

Target: solve linear equation $Ax = b$. $A_{m\times m}$ is positive definite

Input: A, b, x_0 (initial guess)

$r_0 := b - Ax_0$

$p_0 := r_0$

$k := 0$

Loop until $k = m$

$\alpha_k := \dfrac{r_k' r_k}{p_k' A p_k}$

$x_{k+1} := x_k + \alpha_k p_k$

$r_{k+1} := r_k - \alpha_k A p_k$

If r_{k+1} is sufficently small then exit loop

$\beta_k := \dfrac{r_{k+1}' r_{k+1}}{r_k' r_k}$

$p_{k+1} := r_{k+1} + \beta_k p_k$

$k := k + 1$

End loop

Output: x_{k+1}

https://en.wikipedia.org/wiki/Conjugate_gradient_method

适用范围

- CG 尤其适合矩阵乘法能高效计算的场合
- 稀疏矩阵
- 分布式矩阵

- 但一定要注意验证正定性

正定性

- 哪些矩阵是正定的?

- 特征值均大于0
- 非退化分布的协方差矩阵
- $X'X + \lambda I,\ \lambda > 0$

实现

- lec8-cg.ipynb

应用

- 利用 CG 来求解回归问题
- https://cosx.org/2016/11/conjugate-gradient-for-regression/

岭回归-续

解决思路

1. 从原始数据生成 RDD（与线性回归步骤相同）
2. 计算 $X'Y$
3. 定义运算 $h \to Ah$，其中 $A = X'X + \lambda I$
4. 利用 CG 解线性方程组

CG

- 利用 CG 求解 $Ax = b$ 时，我们只需要定义计算 Ah 的“运算符”即可，其中 h 是任意的向量
- 并不需要真正计算出 A
- 例如对于 $A = X'X + \lambda I$，计算

$$Ah = X'Xh + \lambda h$$

要比计算 A 本身高效得多！

乘法运算

- $A = X'X + \lambda I$
- $Ah = X'Xh + \lambda h$
- $X'Xh$ 可分布式进行

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_m \end{pmatrix} \qquad X \in \mathbb{R}^{n \times p}, \quad X_i \in \mathbb{R}^{n_i \times p}$$

$$X'Xv = \begin{pmatrix} X_1' & X_2' & \cdots & X_m' \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_m \end{pmatrix} v$$

$$= \begin{pmatrix} X_1' & X_2' & \cdots & X_m' \end{pmatrix} \begin{pmatrix} X_1 v \\ X_2 v \\ \vdots \\ X_m v \end{pmatrix} = X_1' X_1 v + \cdots + X_m' X_m v$$

- 计算 $X_i' X_i v = X_i'(X_i v)$ 时应先算 $X_i v$!

实现

- lec7-regression.ipynb

Logistic 回归

Logistic 回归

- 假定 $Y|x \sim Bernoulli(\rho(\beta' x))$
- $\rho(x) = 1/(1 + e^{-x})$，即 Sigmoid 函数
- $\rho(\beta' x)$ 代表 Y 取1的概率

- 给定数据 $(y_i, x_i),\ i = 1, \ldots, n$
- 估计 β

目标函数

- 极大似然准则

$$L(\beta) = -\sum_{i=1}^{n} \{y_i \log \rho_i + (1 - y_i) \log(1 - \rho_i)\}$$

- 其中 $\rho_i = \rho(x_i'\beta)$