R中大规模矩阵的SVD与矩阵补全

第七届中国R语言会议

邱怡轩

CAPITAL OF STATISTICS
PROFESSION, HUMANITY & INTEGRITY

概要

SVD 基本概念

SVD 的计算与实现

矩阵近似与矩阵补全

概要

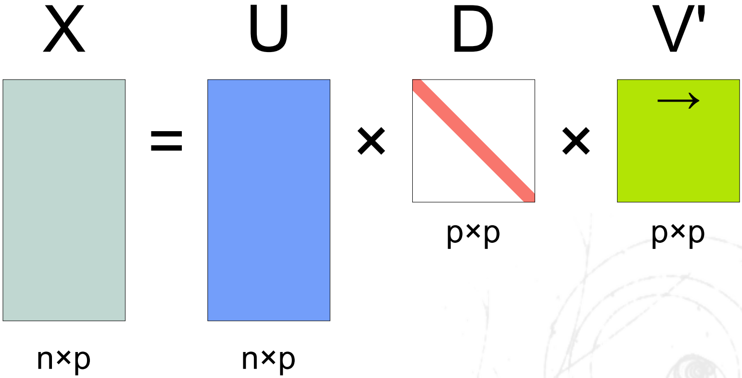
SVD 基本概念

SVD 的计算与实现

矩阵近似与矩阵补全

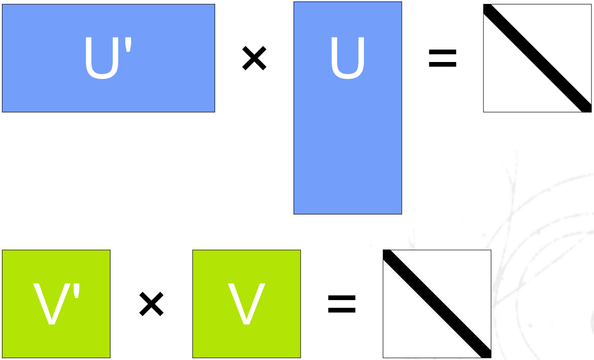
SVD

- SVD = Singular Value Decomposition = 奇异值分解



SVD

- U：左奇异向量按列组成的矩阵
- D：奇异值组成的对角矩阵
- V：右奇异向量按列组成的矩阵



SVD 的作用

- 为什么要研究 SVD？两条主线
- 对传统统计方法的重新理解
    - 主成分分析（PCA）
    - 线性回归
- 更加现代的应用
    - 矩阵近似
    - 矩阵补全

SVD 与 PCA

- PCA：回归和机器学习中常用的降维方法
- 通常的流程
  - 给定数据矩阵 X，假设已进行中心化
  - 计算协方差矩阵 $V = X'X$
  - 对协方差矩阵进行特征值分解 $V = \Gamma\Lambda\Gamma'$，PCA 载荷（系数）保存于 Γ 中
  - 计算 PCA 得分 $S = X\Gamma$
- SVD 可以极大简化这一流程

SVD 与 PCA

- 假设 X 已进行 SVD 分解 $X = UDV'$
- 协方差矩阵 $V = X'X = VDU'UDV = VD^2V'$，所以 $\Gamma = V, \Lambda = D^2$
- PCA 得分 $S = X\Gamma = UDV'V = UD$
- PCA 系数保存在 V 中，得分保存在 UD 中
- 避免了矩阵运算 $X'X$，且通常减小了精度损失

SVD 与回归

- $X = UDV'$ 为数据矩阵，Y 为因变量
- 回归系数

$$\hat{\beta} = (X'X)^{-1}X'Y = VD^{-2}V'VDU'Y = VD^{-1}U'Y$$

- 拟合值

$$\hat{y} = X(X'X)^{-1}X'Y = UDV'VD^{-1}U'Y = UU'Y$$

注意，$U'U = I$ 但 $UU' \neq I$！

概要

SVD 基本概念

SVD 的计算与实现

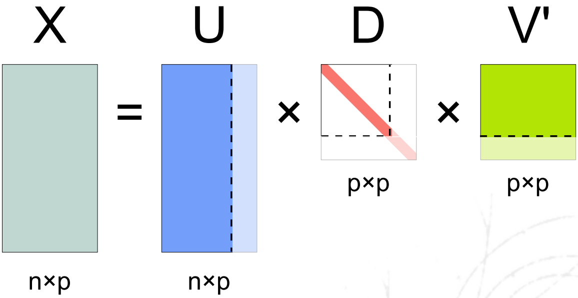
矩阵近似与矩阵补全

计算与实现

- SVD 属于非常底层的运算
- 几乎所有的科学计算软件包（R，Matlab/Octave，Numpy/Scipy，Julia 等等）都提供了 SVD 的相关函数
- R 中为 `svd()`
- 主要的挑战在于矩阵维度非常高时，SVD 的计算负担太大

计算与实现

- 提高计算效率的途径
  - 只计算一部分奇异值/奇异向量（为什么?）



  - 利用稀疏矩阵等特殊的结构

ARPACK/rARPACK

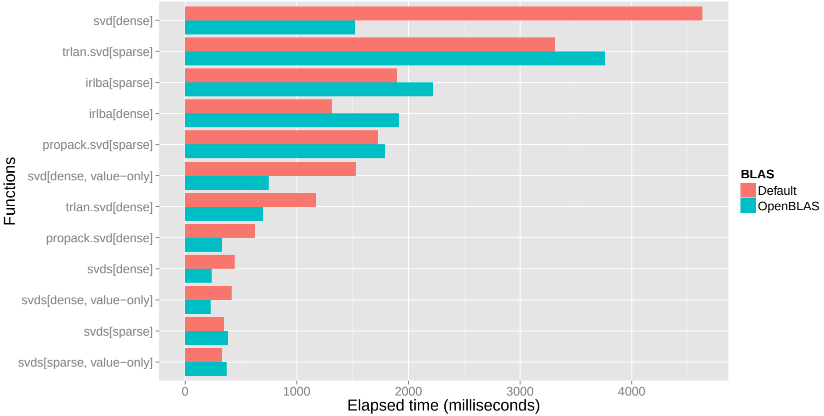
- ARPACK (`http://www.caam.rice.edu/software/ARPACK/`)
  - 一套用 FORTRAN 编写的软件库，用来解决特征值/特征向量问题
  - 只计算满足需求的**一部分**特征值/特征向量
- rARPACK（`http://cran.r-project.org/web/packages/rARPACK/index.html`）
  - R 对 ARPACK 的一个封装
  - 提供函数 `eigs()` 计算部分特征值分解，`svds()` 计算部分 SVD
  - 针对 R 中特殊类型的矩阵（对称矩阵、稀疏矩阵）进行优化
  - 利用 BLAS 加速/并行运算

函数说明

```
svds(x, k, nu, nv)
```

- 参数
  - `x`：进行 SVD 分解的矩阵，可为普通矩阵 (`matrix`)、对称矩阵 (`dsyMatrix`) 或稀疏矩阵 (`dgCMatrix`)
  - `k`：需计算的奇异值数量
  - `nu`，`nv`：需计算的左/右奇异向量数量
- 返回值——列表
  - `u`，`d`，`v`：（部分）奇异值和左/右奇异向量
  - `nconv`：收敛的奇异值数量
  - `niter`：迭代次数

性能



svd[dense]
trlan.svd[sparse]
irlba[sparse]
irlba[dense]
propack.svd[sparse]
svd[dense, value−only]
trlan.svd[dense]
propack.svd[dense]
svds[dense]
svds[dense, value−only]
svds[sparse]
svds[sparse, value−only]
Functions
0
1000
2000
3000
4000
Elapsed time (milliseconds)
BLAS
Default
OpenBLAS

概要

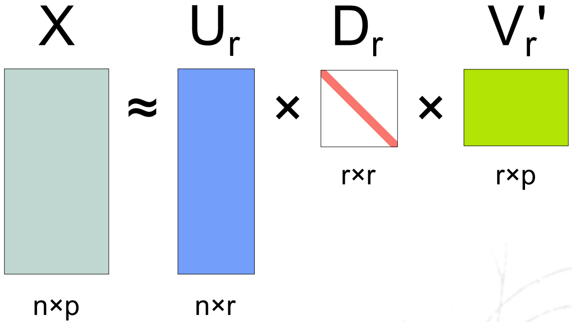
SVD 基本概念

SVD 的计算与实现

矩阵近似与矩阵补全

矩阵近似

- SVD 更大的价值在于其提供了一种对矩阵的近似方法



- 计算前 r 个奇异值/奇异向量配对，$X \approx U_r D_r V_r'$
- 在一定的准则下，这种近似是秩为 r 的矩阵中“最优”的

矩阵近似与降维

- 矩阵的 r 阶近似，等价于使用 r 个主成分对数据进行降维
- 前 r 个主成分的得分矩阵即为 $U_r D_r$
- 可以利用 rARPACK 只计算一部分奇异值的优势，避免计算那些我们不需要的主成分

应用：图像压缩

- 原图（1000×622）

http://www.bjcaca.com/bisai/show.php?pid=33844&bid=40

应用：图像压缩

- $r = 5$（压缩比 1.3%）

应用：图像压缩

- $r = 20$（压缩比 5.2%）

应用：图像压缩

- $r = 50$（压缩比 13%）

应用：图像压缩

- $r = 100$（压缩比 26%）

矩阵补全

- 观测到了矩阵的部分元素，希望对缺失值进行插补
- 推荐系统

	电影 1	电影 2	电影 3	电影 4	……
用户 1	1	?	3	?	……
用户 2	3	5	?	2	……
用户 3	?	4	?	5	……
用户 4	2	2	1	3	……
……	……	……	……	……	……

矩阵补全

- 图片修复

矩阵补全

- 主要原理
  - 矩阵的主要信息保存在低维的结构中
  - 缺失的信息可以通过这些信息进行还原
- Rahul Mazumder, Trevor Hastie and Rob Tibshirani (2010)

$$\min_{Z} \sum_{Observed(i,j)} (X_{ij} - Z_{ij})^2 \quad \text{subject to } \|Z\|_* \leq \tau$$

- 反复计算 SVD 进行迭代，求解恢复后的矩阵 Z

矩阵补全

- 图片修复

R 中实现

- softImpute 软件包
- 给定带缺失值的矩阵 `x`
- `softImpute(x, rank.max, lambda)` 拟合模型
- `complete(x, fit)` 补全矩阵

总结

- SVD 本身是一个强大的矩阵代数工具
- 与统计学中的经典方法有紧密的联系
- 矩阵近似与降维
- 矩阵补全的理论基础