海量高维数据与维度约减

七月在线 龙老师

2016年7月23日

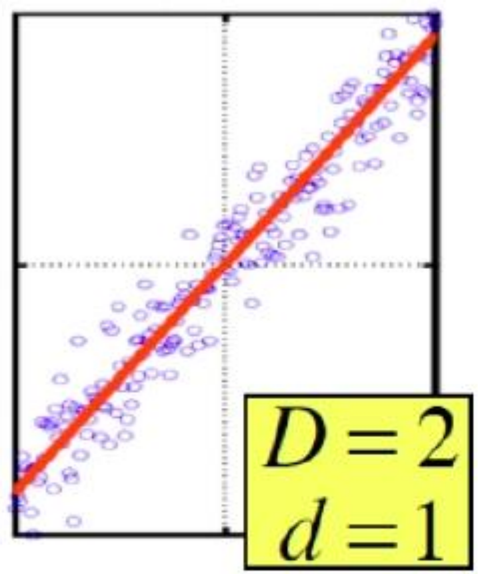
主要内容

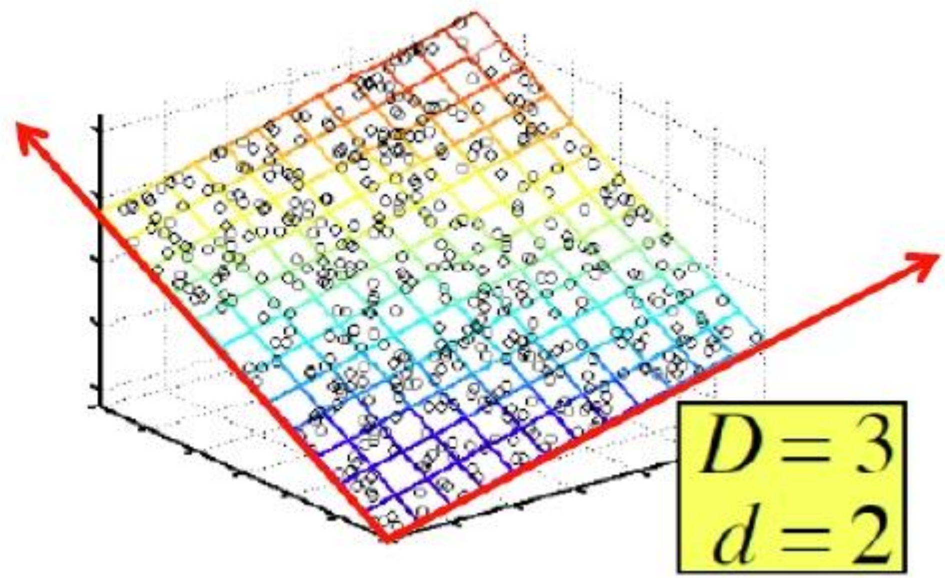
- 为什么要数据降维
- 为什么能数据降维
- SVD
  - 基本概念与性质
  - 怎么用SVD降维
  - 实际案例
- CUR

为什么要降维

- 海量数据太大，不得不降维。
- 可以让你使用简单的模型运算的更快，更容易理解，更容易维护。
- 优质的降维数据可以在使用不是最优的模型参数的情况下得到不错的预测结果，这样你就不必费力去选择最适合的模型和最优的参数了。

为什么能降维





- 假设：数据实际上是存在或者靠近一个低维子空间中。
- 子空间的坐标轴能够最有效地表达这个数据

回忆：矩阵的秩

- 矩阵的秩：矩阵的线性独立的列（行）的个数。

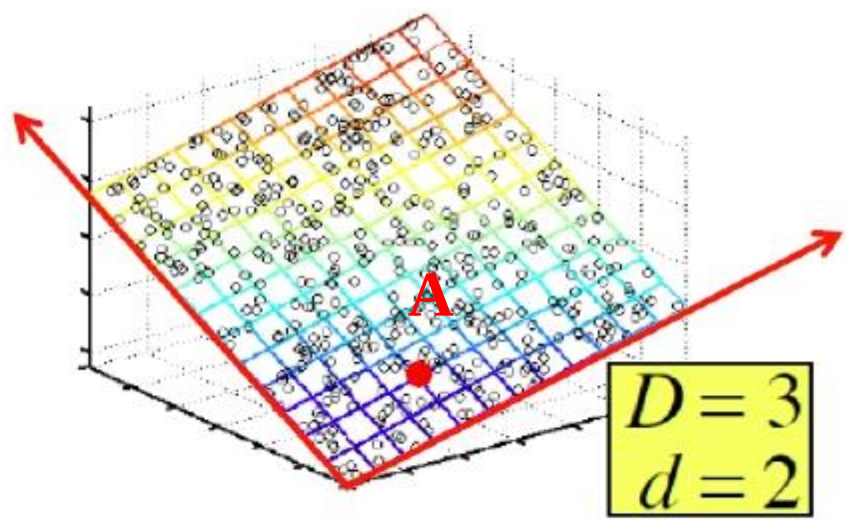
$$\text{矩阵}\ A = \begin{bmatrix} 1 & 2 & 1 \\ -2 & -3 & 1 \\ 3 & 5 & 0 \end{bmatrix}\ \text{的秩}\ r=2$$

- 为什么关心矩阵的秩？
  - 我们可以把A用两个新的基向量表示：
    - [1 2 1] [-2 -3 1]
  - 则相应的坐标就变为: [1 0] [0 1] [1 1]

秩其实就是最小的维度

- 观察三维视图：
  - 把矩阵的每一行作为一点在三维空间的坐标

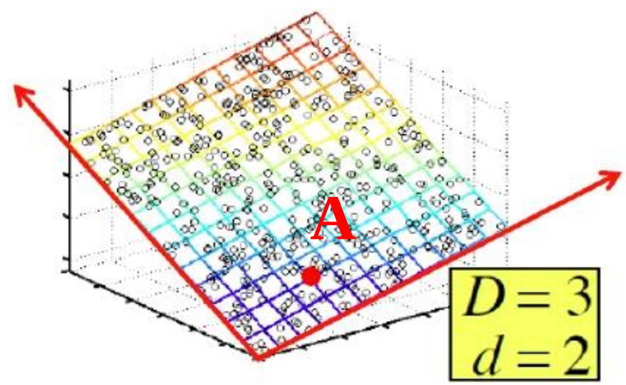
$$\begin{array}{c c} \begin{bmatrix} 1 & 2 & 1 \\ -2 & -3 & 1 \\ 3 & 5 & 0 \end{bmatrix} & \begin{array}{c} \mathbf{A} \\ \mathbf{B} \\ \mathbf{C} \end{array} \end{array}$$



秩其实就是最小的维度

- 我们可以重新采用一个坐标系，这个坐标系是以原矩阵的前两行作为坐标基底的。
  - 老坐标基底：[1 0 0] [0 1 0] [0 0 1]
  - 新坐标基底：**[1 2 1] [-2 -3 1]**
- 则ABC三点的新坐标分别为：
  - **A:** [1 0],
  - **B**: [0 1],
  - **C**: [1 1]

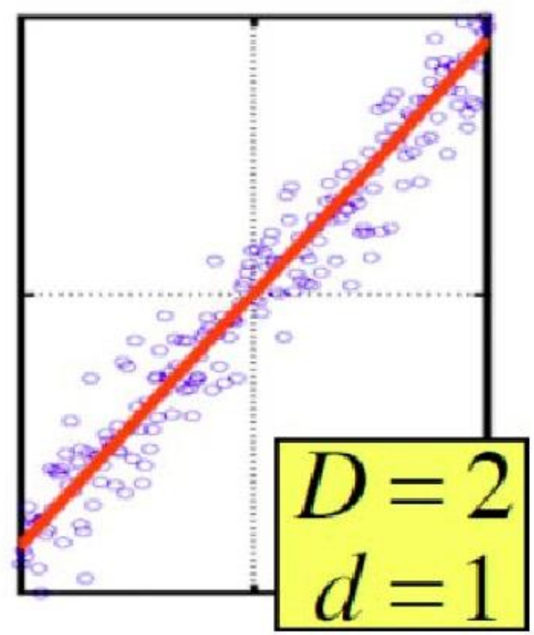
$$\begin{array}{cc} \begin{bmatrix} 1 & 2 & 1 \\ -2 & -3 & 1 \\ 3 & 5 & 0 \end{bmatrix} & \begin{array}{c} \mathbf{A} \\ \mathbf{B} \\ \mathbf{C} \end{array} \end{array}$$



- 注意：我们已经把ABC三点降维了

降维的关键

- 降维的关键就是找到能够表达数据的最少维度，用最少的坐标轴表示数据。
  - 右图的点在二维空间中
  - 但大量聚集在红线附近
  - 所以就可以用红线所代表的一维坐标来表示。
  - 当然这样做有一点误差



SVD

$$A_{[m \times n]} = U_{[m \times r]} \Sigma_{[r \times r]} (V_{[n \times r]})^T$$

A: 输入矩阵

m x n (e.g., m 篇文章, n 个词语)

U: 左奇异向量矩阵

m x r (m 篇文章, r 个主题)

Σ: 奇异值矩阵

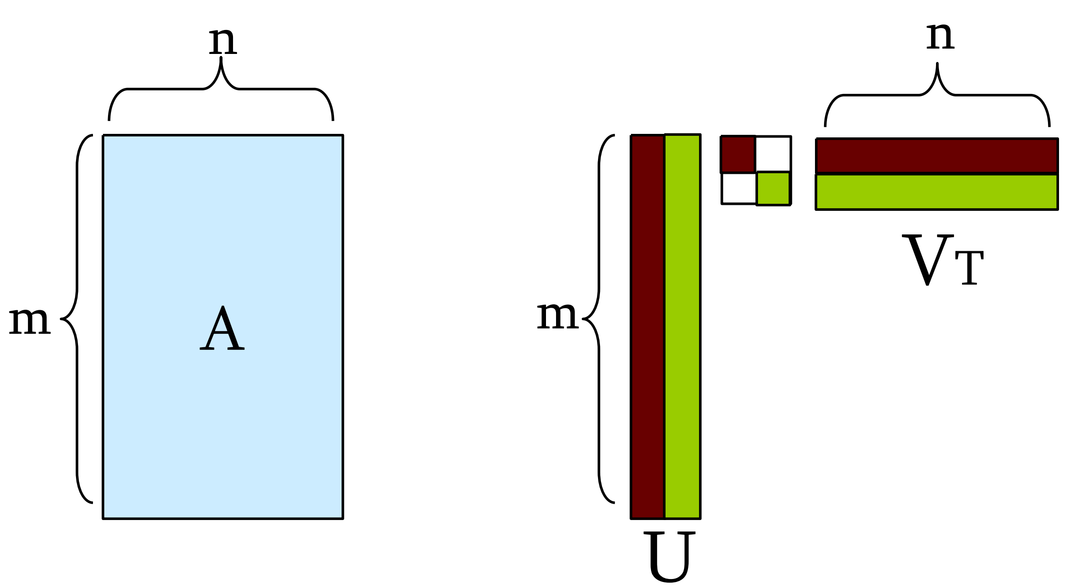
r x r 对角阵 (每个主题的重要性) (r : 矩阵**A**的秩)

V:右奇异向量矩阵

n x r (n 个词语, r 个主题)

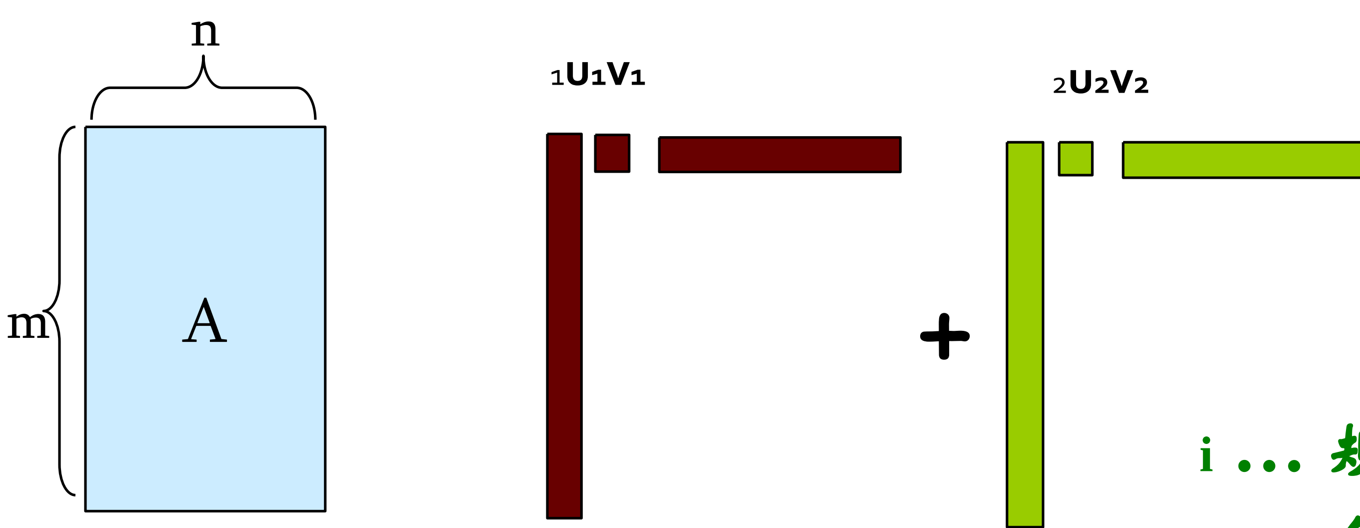
奇异值分解

$$\mathbf{A} \approx \mathbf{U}\Sigma\mathbf{V}^T = \sum_i \sigma_i \mathbf{u}_i \circ \mathbf{v}_i^{\mathrm{T}}$$



奇异值分解

$$\mathbf{A} \approx \mathbf{U}\Sigma\mathbf{V}^T = \sum_i \sigma_i \mathbf{u}_i \circ \mathbf{v}_i^{\mathrm{T}}$$



i … 规模

ui … 向量

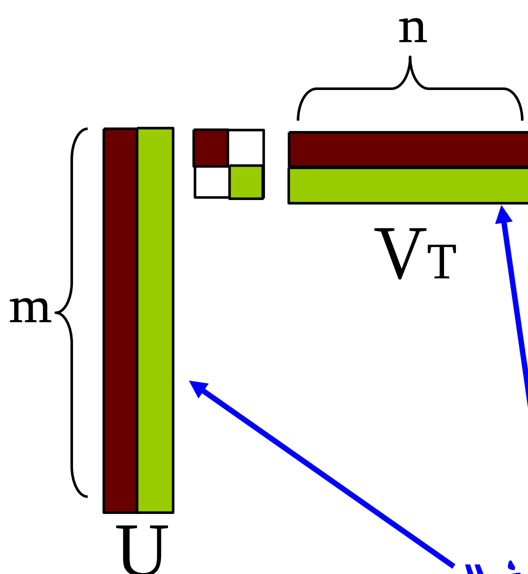
Vi … 向量

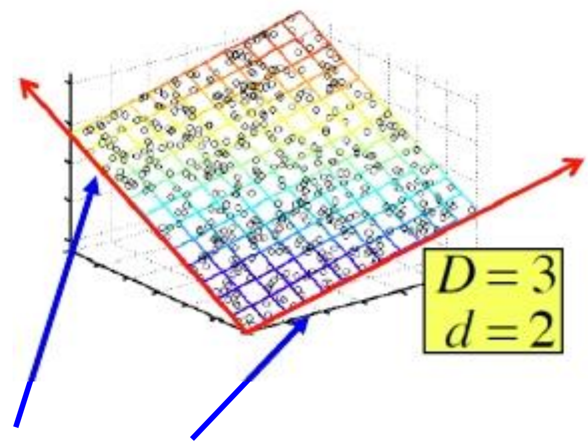
SVD的性质

- 对于一个实数矩阵，总能够拆解成三个矩阵相乘$A = U \sum \cdot V_T$
- 这三个矩阵满足如下性质：
  - U、∑、V 是唯一的
  - U、V的列是单位标准正交基
  - ∑是对角阵，对角上每一个值是奇异值，是正数，并且按降序排序。

SVD案例：用户看电影

$$
\begin{array}{c|c}
 & \begin{matrix} \text{黑客帝国} & \text{异形} & \text{冲出宁静号} & \text{卡萨布兰卡} & \text{天使爱美丽} \end{matrix} \\
\begin{matrix} \text{科幻} \\ \\ \\ \\ \text{言情} \\ \\ \\ \end{matrix} &
\begin{bmatrix}
1 & 1 & 1 & 0 & 0 \\
3 & 3 & 3 & 0 & 0 \\
4 & 4 & 4 & 0 & 0 \\
5 & 5 & 5 & 0 & 0 \\
0 & 2 & 0 & 4 & 4 \\
0 & 0 & 0 & 5 & 5 \\
0 & 1 & 0 & 2 & 2
\end{bmatrix}
\end{array}
=
$$





"主题"
隐维度
隐因子

SVD案例：用户看电影

$$
\begin{array}{c|c}
 & \begin{array}{ccccc} \text{黑客帝国} & \text{异形} & \text{冲出宁静号} & \text{卡萨布兰卡} & \text{天使爱美丽} \end{array} \\
\begin{array}{c} \text{科幻} \\ \\ \\ \\ \text{言情} \\ \\ \\ \end{array} &
\begin{bmatrix}
\mathbf{1} & \mathbf{1} & \mathbf{1} & 0 & 0 \\
\mathbf{3} & \mathbf{3} & \mathbf{3} & 0 & 0 \\
\mathbf{4} & \mathbf{4} & \mathbf{4} & 0 & 0 \\
\mathbf{5} & \mathbf{5} & \mathbf{5} & 0 & 0 \\
0 & \mathbf{2} & 0 & \mathbf{4} & \mathbf{4} \\
0 & 0 & 0 & \mathbf{5} & \mathbf{5} \\
0 & \mathbf{1} & 0 & \mathbf{2} & \mathbf{2}
\end{bmatrix}
\end{array}
=
\begin{bmatrix}
\mathbf{0.13} & 0.02 & -0.01 \\
\mathbf{0.41} & 0.07 & -0.03 \\
\mathbf{0.55} & 0.09 & -0.04 \\
\mathbf{0.68} & 0.11 & -0.05 \\
0.15 & \mathbf{-0.59} & \mathbf{0.65} \\
0.07 & \mathbf{-0.73} & \mathbf{-0.67} \\
0.07 & \mathbf{-0.29} & \mathbf{0.32}
\end{bmatrix}
\times
\begin{bmatrix}
\mathbf{12.4} & 0 & 0 \\
0 & \mathbf{9.5} & 0 \\
0 & 0 & \mathbf{1.3}
\end{bmatrix}
\times
\begin{bmatrix}
\mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\
0.12 & -0.02 & 0.12 & \mathbf{-0.69} & \mathbf{-0.69} \\
0.40 & \mathbf{-0.80} & 0.40 & 0.09 & 0.09
\end{bmatrix}
$$

SVD案例：用户看电影

$$
\begin{array}{cc}
 & \begin{array}{ccccc} \text{黑客帝国} & \text{异形} & \text{冲出宁静号} & \text{卡萨布兰卡} & \text{天使爱美丽} \end{array} \\
\begin{array}{c} \\ \text{科幻} \\ \\ \\ \\ \text{言情} \\ \\ \end{array} & \begin{bmatrix} \mathbf{1} & \mathbf{1} & \mathbf{1} & 0 & 0 \\ \mathbf{3} & \mathbf{3} & \mathbf{3} & 0 & 0 \\ \mathbf{4} & \mathbf{4} & \mathbf{4} & 0 & 0 \\ \mathbf{5} & \mathbf{5} & \mathbf{5} & 0 & 0 \\ 0 & \mathbf{2} & 0 & \mathbf{4} & \mathbf{4} \\ 0 & 0 & 0 & \mathbf{5} & \mathbf{5} \\ 0 & \mathbf{1} & 0 & \mathbf{2} & \mathbf{2} \end{bmatrix}
\end{array}
=
\begin{array}{c}
\begin{array}{ccc} \text{科幻-主题} & \text{言情-主题} & \end{array} \\
\begin{bmatrix} \mathbf{0.13} & 0.02 & -0.01 \\ \mathbf{0.41} & 0.07 & -0.03 \\ \mathbf{0.55} & 0.09 & -0.04 \\ \mathbf{0.68} & 0.11 & -0.05 \\ 0.15 & \mathbf{-0.59} & \mathbf{0.65} \\ 0.07 & \mathbf{-0.73} & \mathbf{-0.67} \\ 0.07 & \mathbf{-0.29} & \mathbf{0.32} \end{bmatrix}
\end{array}
\times
\begin{bmatrix} \mathbf{12.4} & 0 & 0 \\ 0 & \mathbf{9.5} & 0 \\ 0 & 0 & \mathbf{1.3} \end{bmatrix}
\times
\begin{bmatrix} \mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\ 0.12 & -0.02 & 0.12 & \mathbf{-0.69} & \mathbf{-0.69} \\ 0.40 & \mathbf{-0.80} & 0.40 & 0.09 & 0.09 \end{bmatrix}
$$

SVD案例：用户看电影

*U*是"用户-主题"相似矩阵

科幻-主题

言情-主题

$$
\begin{array}{c}
 \\ 科幻 \\ \\ \\ \\ 言情 \\ \\ \\
\end{array}
\begin{array}{c}
\begin{array}{ccccc} 黑客帝国 & 异形 & 冲出宁静号 & 卡萨布兰卡 & 天使爱美丽 \end{array} \\
\left[\begin{array}{ccccc}
\mathbf{1} & \mathbf{1} & \mathbf{1} & 0 & 0 \\
\mathbf{3} & \mathbf{3} & \mathbf{3} & 0 & 0 \\
\mathbf{4} & \mathbf{4} & \mathbf{4} & 0 & 0 \\
\mathbf{5} & \mathbf{5} & \mathbf{5} & 0 & 0 \\
0 & \mathbf{2} & 0 & \mathbf{4} & \mathbf{4} \\
0 & 0 & 0 & \mathbf{5} & \mathbf{5} \\
0 & \mathbf{1} & 0 & \mathbf{2} & \mathbf{2}
\end{array}\right]
\end{array}
=
\left[\begin{array}{ccc}
\mathbf{0.13} & 0.02 & -0.01 \\
\mathbf{0.41} & 0.07 & -0.03 \\
\mathbf{0.55} & 0.09 & -0.04 \\
\mathbf{0.68} & 0.11 & -0.05 \\
0.15 & \mathbf{-0.59} & \mathbf{0.65} \\
0.07 & \mathbf{-0.73} & \mathbf{-0.67} \\
0.07 & \mathbf{-0.29} & \mathbf{0.32}
\end{array}\right]
\times
\left[\begin{array}{ccc}
\mathbf{12.4} & 0 & 0 \\
0 & \mathbf{9.5} & 0 \\
0 & 0 & \mathbf{1.3}
\end{array}\right]
\times
\left[\begin{array}{ccccc}
\mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\
0.12 & -0.02 & 0.12 & \mathbf{-0.69} & \mathbf{-0.69} \\
0.40 & \mathbf{-0.80} & 0.40 & 0.09 & 0.09
\end{array}\right]
$$

SVD案例：用户看电影

科幻-主题

科幻-主题的“强度”

$$
\begin{array}{c|ccccc}
 & \text{黑客帝国} & \text{异形} & \text{冲出宁静号} & \text{卡萨布兰卡} & \text{天使爱美丽} \\
\hline
\text{科幻} & 1 & 1 & 1 & 0 & 0 \\
\text{科幻} & 3 & 3 & 3 & 0 & 0 \\
\text{科幻} & 4 & 4 & 4 & 0 & 0 \\
\text{科幻} & 5 & 5 & 5 & 0 & 0 \\
\text{言情} & 0 & 2 & 0 & 4 & 4 \\
\text{言情} & 0 & 0 & 0 & 5 & 5 \\
\text{言情} & 0 & 1 & 0 & 2 & 2
\end{array}
$$

$$
\begin{bmatrix}
1 & 1 & 1 & 0 & 0 \\
3 & 3 & 3 & 0 & 0 \\
4 & 4 & 4 & 0 & 0 \\
5 & 5 & 5 & 0 & 0 \\
0 & 2 & 0 & 4 & 4 \\
0 & 0 & 0 & 5 & 5 \\
0 & 1 & 0 & 2 & 2
\end{bmatrix}
=
\begin{bmatrix}
\mathbf{0.13} & 0.02 & -0.01 \\
\mathbf{0.41} & 0.07 & -0.03 \\
\mathbf{0.55} & 0.09 & -0.04 \\
\mathbf{0.68} & 0.11 & -0.05 \\
0.15 & \mathbf{-0.59} & \mathbf{0.65} \\
0.07 & \mathbf{-0.73} & \mathbf{-0.67} \\
0.07 & \mathbf{-0.29} & \mathbf{0.32}
\end{bmatrix}
\times
\begin{bmatrix}
\mathbf{12.4} & 0 & 0 \\
0 & \mathbf{9.5} & 0 \\
0 & 0 & \mathbf{1.3}
\end{bmatrix}
\times
\begin{bmatrix}
\mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\
0.12 & -0.02 & 0.12 & \mathbf{-0.69} & \mathbf{-0.69} \\
0.40 & \mathbf{-0.80} & 0.40 & 0.09 & 0.09
\end{bmatrix}
$$

SVD案例：用户看电影

科幻-主题

V是"电影-主题"相似矩阵

$$
\begin{array}{c}
\\ \text{科幻} \\ \\ \\ \text{言情} \\ \\
\end{array}
\begin{array}{c}
\begin{array}{ccccc} \text{黑客帝国} & \text{异形} & \text{冲出宁静号} & \text{卡萨布兰卡} & \text{天使爱美丽} \end{array} \\
\begin{bmatrix}
1 & 1 & 1 & 0 & 0 \\
3 & 3 & 3 & 0 & 0 \\
4 & 4 & 4 & 0 & 0 \\
5 & 5 & 5 & 0 & 0 \\
0 & 2 & 0 & 4 & 4 \\
0 & 0 & 0 & 5 & 5 \\
0 & 1 & 0 & 2 & 2
\end{bmatrix}
\end{array}
=
\begin{bmatrix}
\mathbf{0.13} & 0.02 & -0.01 \\
\mathbf{0.41} & 0.07 & -0.03 \\
\mathbf{0.55} & 0.09 & -0.04 \\
\mathbf{0.68} & 0.11 & -0.05 \\
0.15 & \mathbf{-0.59} & \mathbf{0.65} \\
0.07 & \mathbf{-0.73} & \mathbf{-0.67} \\
0.07 & \mathbf{-0.29} & \mathbf{0.32}
\end{bmatrix}
\times
\begin{bmatrix}
\mathbf{12.4} & 0 & 0 \\
0 & \mathbf{9.5} & 0 \\
0 & 0 & \mathbf{1.3}
\end{bmatrix}
\times
\begin{bmatrix}
\mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\
0.12 & -0.02 & 0.12 & \mathbf{-0.69} & \mathbf{-0.69} \\
0.40 & \mathbf{-0.80} & 0.40 & 0.09 & 0.09
\end{bmatrix}
$$

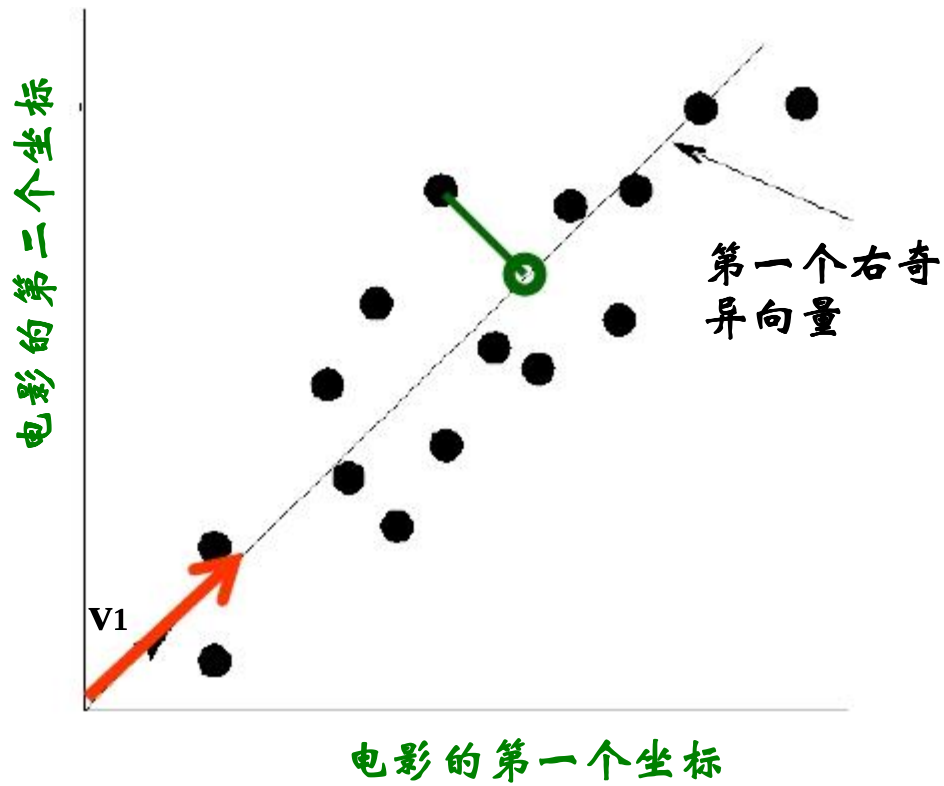
科幻-主题

SVD的深入理解

- “电影”、“用户”和“主题”
  - U:“用户-主题”相似矩阵
  - V:“电影-主题”相似矩阵
  - Σ：其对角元素是每一个主题的“强度”

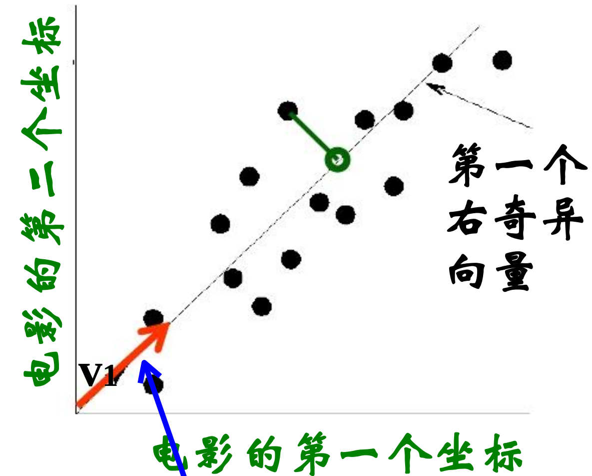
SVD进行降维

- SVD能够给出“最好”的结果
  - 所谓“最好”就是使得平方误差（投影）最小



SVD进行降维

- $A = U \sum V_T$
- U: "用户-主题"相似矩阵
- V: "电影-主题"相似矩阵

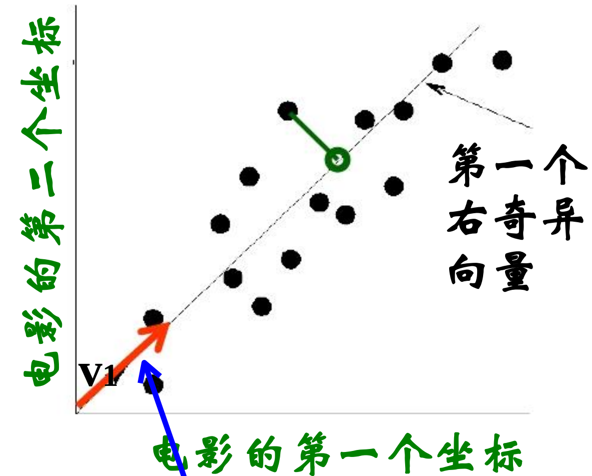


$$\begin{bmatrix} \mathbf{1} & \mathbf{1} & \mathbf{1} & 0 & 0 \\ \mathbf{3} & \mathbf{3} & \mathbf{3} & 0 & 0 \\ \mathbf{4} & \mathbf{4} & \mathbf{4} & 0 & 0 \\ \mathbf{5} & \mathbf{5} & \mathbf{5} & 0 & 0 \\ 0 & \mathbf{2} & 0 & \mathbf{4} & \mathbf{4} \\ 0 & 0 & 0 & \mathbf{5} & \mathbf{5} \\ 0 & \mathbf{1} & 0 & \mathbf{2} & \mathbf{2} \end{bmatrix} = \begin{bmatrix} \mathbf{0.13} & 0.02 & -0.01 \\ \mathbf{0.41} & 0.07 & -0.03 \\ \mathbf{0.55} & 0.09 & -0.04 \\ \mathbf{0.68} & 0.11 & -0.05 \\ 0.15 & \mathbf{-0.59} & \mathbf{0.65} \\ 0.07 & \mathbf{-0.73} & \mathbf{-0.67} \\ 0.07 & \mathbf{-0.29} & \mathbf{0.32} \end{bmatrix} \times \begin{bmatrix} \mathbf{12.4} & 0 & 0 \\ 0 & \mathbf{9.5} & 0 \\ 0 & 0 & \mathbf{1.3} \end{bmatrix} \times \begin{bmatrix} \boxed{\mathbf{0.56} \quad \mathbf{0.59} \quad \mathbf{0.56} \quad 0.09 \quad 0.09} \\ 0.12 \quad -0.02 \quad 0.12 \quad \mathbf{-0.69} \quad \mathbf{-0.69} \\ 0.40 \quad \mathbf{-0.80} \quad 0.40 \quad 0.09 \quad 0.09 \end{bmatrix}$$

SVD进行降维

- $A = U \sum \cdot V_T$

V1坐标轴的方差（“扩散”）



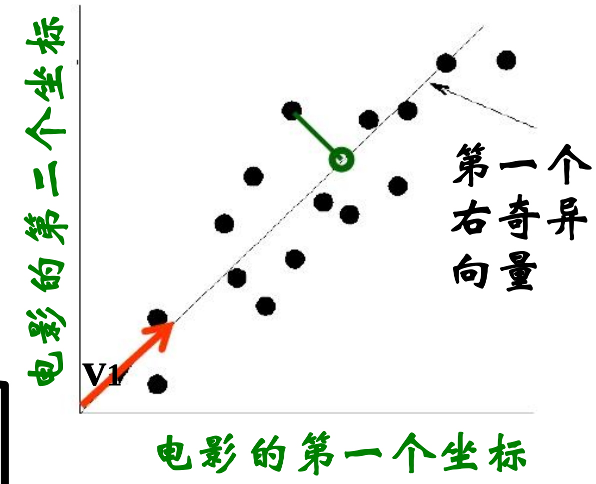
$$
\begin{bmatrix}
\mathbf{1} & \mathbf{1} & \mathbf{1} & 0 & 0 \\
\mathbf{3} & \mathbf{3} & \mathbf{3} & 0 & 0 \\
\mathbf{4} & \mathbf{4} & \mathbf{4} & 0 & 0 \\
\mathbf{5} & \mathbf{5} & \mathbf{5} & 0 & 0 \\
0 & \mathbf{2} & 0 & \mathbf{4} & \mathbf{4} \\
0 & 0 & 0 & \mathbf{5} & \mathbf{5} \\
0 & \mathbf{1} & 0 & \mathbf{2} & \mathbf{2}
\end{bmatrix}
=
\begin{bmatrix}
\mathbf{0.13} & 0.02 & -0.01 \\
\mathbf{0.41} & 0.07 & -0.03 \\
\mathbf{0.55} & 0.09 & -0.04 \\
\mathbf{0.68} & 0.11 & -0.05 \\
0.15 & \mathbf{-0.59} & \mathbf{0.65} \\
0.07 & \mathbf{-0.73} & \mathbf{-0.67} \\
0.07 & \mathbf{-0.29} & \mathbf{0.32}
\end{bmatrix}
\times
\begin{bmatrix}
\mathbf{12.4} & 0 & 0 \\
0 & \mathbf{9.5} & 0 \\
0 & 0 & \mathbf{1.3}
\end{bmatrix}
\times
\begin{bmatrix}
\mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\
0.12 & -0.02 & 0.12 & \mathbf{-0.69} & \mathbf{-0.69} \\
0.40 & \mathbf{-0.80} & 0.40 & 0.09 & 0.09
\end{bmatrix}
$$

SVD进行降维

- $A = U \ \sum \ \cdot V_T$

用户坐标在“科幻-主题”坐标轴上的投影$(U\Sigma)^T$:

$$
\begin{bmatrix}
\mathbf{1} & \mathbf{1} & \mathbf{1} & 0 & 0 \\
\mathbf{3} & \mathbf{3} & \mathbf{3} & 0 & 0 \\
\mathbf{4} & \mathbf{4} & \mathbf{4} & 0 & 0 \\
\mathbf{5} & \mathbf{5} & \mathbf{5} & 0 & 0 \\
0 & \mathbf{2} & 0 & \mathbf{4} & \mathbf{4} \\
0 & 0 & 0 & \mathbf{5} & \mathbf{5} \\
0 & \mathbf{1} & 0 & \mathbf{2} & \mathbf{2}
\end{bmatrix}
\qquad
\begin{bmatrix}
\mathbf{0.13} & 0.02 & -0.01 \\
\mathbf{0.41} & 0.07 & -0.03 \\
\mathbf{0.55} & 0.09 & -0.04 \\
\mathbf{0.68} & 0.11 & -0.05 \\
0.15 & \mathbf{-0.59} & \mathbf{0.65} \\
0.07 & \mathbf{-0.73} & \mathbf{-0.67} \\
0.07 & \mathbf{-0.29} & \mathbf{0.32}
\end{bmatrix}
$$



SVD进行降维

- Q:SVD怎么降维？能够精确到什么程度？

$$
\begin{bmatrix} \mathbf{1} & \mathbf{1} & \mathbf{1} & 0 & 0 \\ \mathbf{3} & \mathbf{3} & \mathbf{3} & 0 & 0 \\ \mathbf{4} & \mathbf{4} & \mathbf{4} & 0 & 0 \\ \mathbf{5} & \mathbf{5} & \mathbf{5} & 0 & 0 \\ 0 & \mathbf{2} & 0 & \mathbf{4} & \mathbf{4} \\ 0 & 0 & 0 & \mathbf{5} & \mathbf{5} \\ 0 & \mathbf{1} & 0 & \mathbf{2} & \mathbf{2} \end{bmatrix}
=
\begin{bmatrix} \mathbf{0.13} & 0.02 & -0.01 \\ \mathbf{0.41} & 0.07 & -0.03 \\ \mathbf{0.55} & 0.09 & -0.04 \\ \mathbf{0.68} & 0.11 & -0.05 \\ 0.15 & \mathbf{-0.59} & \mathbf{0.65} \\ 0.07 & \mathbf{-0.73} & \mathbf{-0.67} \\ 0.07 & \mathbf{-0.29} & \mathbf{0.32} \end{bmatrix}
\times
\begin{bmatrix} \mathbf{12.4} & 0 & 0 \\ 0 & \mathbf{9.5} & 0 \\ 0 & 0 & \mathbf{1.3} \end{bmatrix}
\times
\begin{bmatrix} \mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\ 0.12 & -0.02 & 0.12 & \mathbf{-0.69} & \mathbf{-0.69} \\ 0.40 & \mathbf{-0.80} & 0.40 & 0.09 & 0.09 \end{bmatrix}
$$

SVD进行降维

- Q:SVD怎么降维？能够精确到什么程度？
- A:把最小的奇异值设为0

$$
\begin{bmatrix}
\mathbf{1} & \mathbf{1} & \mathbf{1} & 0 & 0 \\
\mathbf{3} & \mathbf{3} & \mathbf{3} & 0 & 0 \\
\mathbf{4} & \mathbf{4} & \mathbf{4} & 0 & 0 \\
\mathbf{5} & \mathbf{5} & \mathbf{5} & 0 & 0 \\
0 & \mathbf{2} & 0 & \mathbf{4} & \mathbf{4} \\
0 & 0 & 0 & \mathbf{5} & \mathbf{5} \\
0 & \mathbf{1} & 0 & \mathbf{2} & \mathbf{2}
\end{bmatrix}
=
\begin{bmatrix}
\mathbf{0.13} & 0.02 & -0.01 \\
\mathbf{0.41} & 0.07 & -0.03 \\
\mathbf{0.55} & 0.09 & -0.04 \\
\mathbf{0.68} & 0.11 & -0.05 \\
0.15 & \mathbf{-0.59} & \mathbf{0.65} \\
0.07 & \mathbf{-0.73} & \mathbf{-0.67} \\
0.07 & \mathbf{-0.29} & \mathbf{0.32}
\end{bmatrix}
\times
\begin{bmatrix}
\mathbf{12.4} & 0 & 0 \\
0 & \mathbf{9.5} & 0 \\
0 & 0 & \cancel{\mathbf{1.3}}
\end{bmatrix}
\times
\begin{bmatrix}
\mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\
0.12 & -0.02 & 0.12 & \mathbf{-0.69} & \mathbf{-0.69} \\
0.40 & \mathbf{-0.80} & 0.40 & 0.09 & 0.09
\end{bmatrix}
$$

SVD进行降维

- Q:SVD怎么降维？能够精确到什么程度？
- A:把最小的奇异值设为0

$$\begin{bmatrix} \mathbf{1} & \mathbf{1} & \mathbf{1} & 0 & 0 \\ \mathbf{3} & \mathbf{3} & \mathbf{3} & 0 & 0 \\ \mathbf{4} & \mathbf{4} & \mathbf{4} & 0 & 0 \\ \mathbf{5} & \mathbf{5} & \mathbf{5} & 0 & 0 \\ 0 & \mathbf{2} & 0 & \mathbf{4} & \mathbf{4} \\ 0 & 0 & 0 & \mathbf{5} & \mathbf{5} \\ 0 & \mathbf{1} & 0 & \mathbf{2} & \mathbf{2} \end{bmatrix} \approx \begin{bmatrix} \mathbf{0.13} & 0.02 & -0.01 \\ \mathbf{0.41} & 0.07 & -0.03 \\ \mathbf{0.55} & 0.09 & -0.04 \\ \mathbf{0.68} & 0.11 & -0.05 \\ 0.15 & \mathbf{-0.59} & \mathbf{0.65} \\ 0.07 & \mathbf{-0.73} & \mathbf{-0.67} \\ 0.07 & \mathbf{-0.29} & \mathbf{0.32} \end{bmatrix} \times \begin{bmatrix} \mathbf{12.4} & 0 & 0 \\ 0 & \mathbf{9.5} & 0 \\ 0 & 0 & \cancel{\mathbf{1.3}} \end{bmatrix} \times \begin{bmatrix} \mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\ 0.12 & -0.02 & 0.12 & \mathbf{-0.69} & \mathbf{-0.69} \\ 0.40 & \mathbf{-0.80} & 0.40 & 0.09 & 0.09 \end{bmatrix}$$

SVD进行降维

- Q:SVD怎么降维？能够精确到什么程度？
- A:把最小的奇异值设为0

$$
\begin{bmatrix}
\mathbf{1} & \mathbf{1} & \mathbf{1} & 0 & 0 \\
\mathbf{3} & \mathbf{3} & \mathbf{3} & 0 & 0 \\
\mathbf{4} & \mathbf{4} & \mathbf{4} & 0 & 0 \\
\mathbf{5} & \mathbf{5} & \mathbf{5} & 0 & 0 \\
0 & \mathbf{2} & 0 & \mathbf{4} & \mathbf{4} \\
0 & 0 & 0 & \mathbf{5} & \mathbf{5} \\
0 & \mathbf{1} & 0 & \mathbf{2} & \mathbf{2}
\end{bmatrix}
\approx
\begin{bmatrix}
\mathbf{0.13} & 0.02 & \cancel{-0.01} \\
\mathbf{0.41} & 0.07 & \cancel{-0.03} \\
\mathbf{0.55} & 0.09 & \cancel{-0.04} \\
\mathbf{0.68} & 0.11 & \cancel{-0.05} \\
0.15 & \mathbf{-0.59} & \cancel{\mathbf{0.65}} \\
0.07 & \mathbf{-0.73} & \cancel{\mathbf{-0.67}} \\
0.07 & \mathbf{-0.29} & \cancel{\mathbf{0.32}}
\end{bmatrix}
\times
\begin{bmatrix}
\mathbf{12.4} & 0 & 0 \\
0 & \mathbf{9.5} & 0 \\
0 & 0 & \cancel{\mathbf{1.3}}
\end{bmatrix}
\times
\begin{bmatrix}
\mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\
0.12 & -0.02 & 0.12 & \mathbf{-0.69} & \mathbf{-0.69} \\
\cancel{0.40} & \cancel{\mathbf{-0.80}} & \cancel{0.40} & \cancel{0.09} & \cancel{0.09}
\end{bmatrix}
$$

SVD进行降维

- Q:SVD怎么降维？能够精确到什么程度？
- A:把最小的奇异值设为0

$$\begin{bmatrix} \mathbf{1} & \mathbf{1} & \mathbf{1} & 0 & 0 \\ \mathbf{3} & \mathbf{3} & \mathbf{3} & 0 & 0 \\ \mathbf{4} & \mathbf{4} & \mathbf{4} & 0 & 0 \\ \mathbf{5} & \mathbf{5} & \mathbf{5} & 0 & 0 \\ 0 & \mathbf{2} & 0 & \mathbf{4} & \mathbf{4} \\ 0 & 0 & 0 & \mathbf{5} & \mathbf{5} \\ 0 & \mathbf{1} & 0 & \mathbf{2} & \mathbf{2} \end{bmatrix} \approx \begin{bmatrix} \mathbf{0.13} & 0.02 \\ \mathbf{0.41} & 0.07 \\ \mathbf{0.55} & 0.09 \\ \mathbf{0.68} & 0.11 \\ 0.15 & \mathbf{-0.59} \\ 0.07 & \mathbf{-0.73} \\ 0.07 & \mathbf{-0.29} \end{bmatrix} \times \begin{bmatrix} \mathbf{12.4} & 0 \\ 0 & \mathbf{9.5} \end{bmatrix} \times \begin{bmatrix} \mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\ 0.12 & -0.02 & 0.12 & \mathbf{-0.69} & \mathbf{-0.69} \end{bmatrix}$$

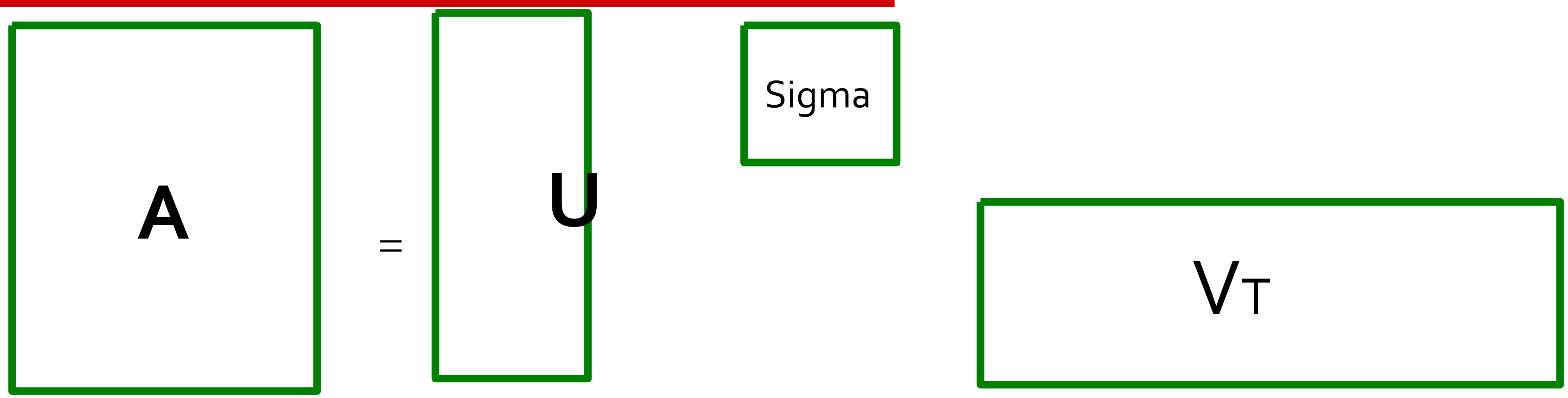
SVD进行降维

- Q:SVD怎么降维？能够精确到什么程度？
- A:把最小的奇异值设为0

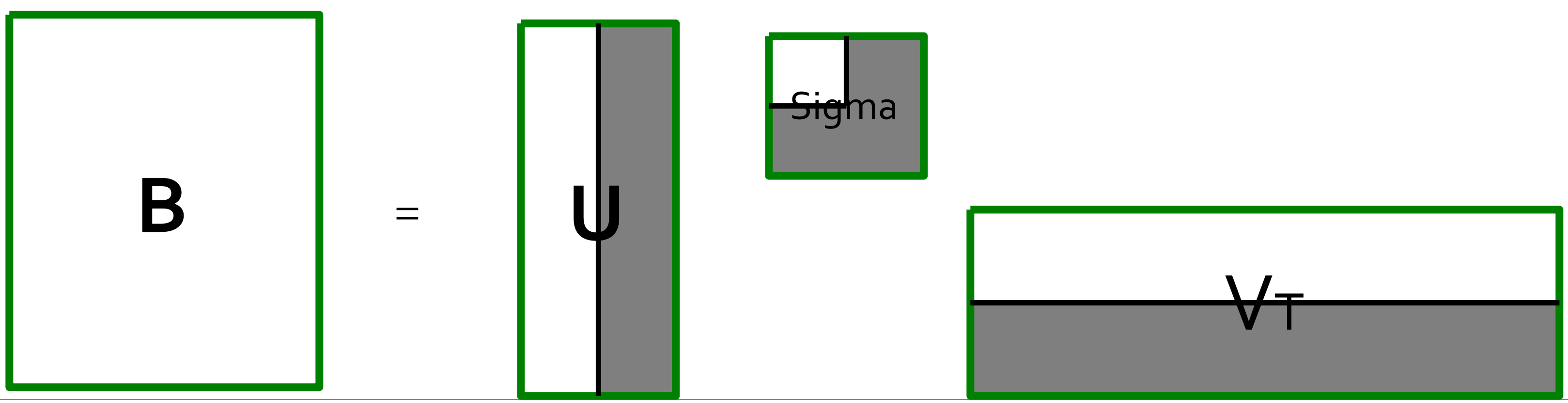
$$\begin{bmatrix} \mathbf{1} & \mathbf{1} & \mathbf{1} & 0 & 0 \\ \mathbf{3} & \mathbf{3} & \mathbf{3} & 0 & 0 \\ \mathbf{4} & \mathbf{4} & \mathbf{4} & 0 & 0 \\ \mathbf{5} & \mathbf{5} & \mathbf{5} & 0 & 0 \\ 0 & \mathbf{2} & 0 & \mathbf{4} & \mathbf{4} \\ 0 & 0 & 0 & \mathbf{5} & \mathbf{5} \\ 0 & \mathbf{1} & 0 & \mathbf{2} & \mathbf{2} \end{bmatrix} \approx \begin{bmatrix} \mathbf{0.92} & \mathbf{0.95} & \mathbf{0.92} & 0.01 & 0.01 \\ \mathbf{2.91} & \mathbf{3.01} & \mathbf{2.91} & -0.01 & -0.01 \\ \mathbf{3.90} & \mathbf{4.04} & \mathbf{3.90} & 0.01 & 0.01 \\ \mathbf{4.82} & \mathbf{5.00} & \mathbf{4.82} & 0.03 & 0.03 \\ 0.70 & \mathbf{0.53} & 0.70 & \mathbf{4.11} & \mathbf{4.11} \\ -0.69 & 1.34 & -0.69 & \mathbf{4.78} & \mathbf{4.78} \\ 0.32 & \mathbf{0.23} & 0.32 & \mathbf{2.01} & \mathbf{2.01} \end{bmatrix}$$

弗罗宾尼斯范数 $\|M\|_F = \sqrt{\Sigma_{ij} M_{ij}^2}$ $\|A-B\|_F = \sqrt{\Sigma_{ij} (A_{ij}-B_{ij})^2}$ 很小

SVD:最好的低秩近似

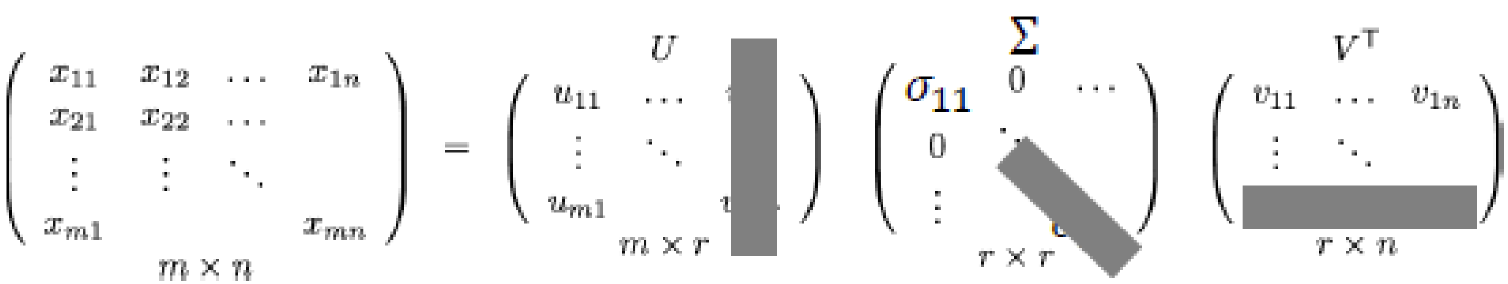


B 是 A 最好的近似: $\|A-B\|_F = \sqrt{\Sigma_{ij}\,(A_{ij}-B_{ij})^2}$ 很小



SVD:最好的低秩近似

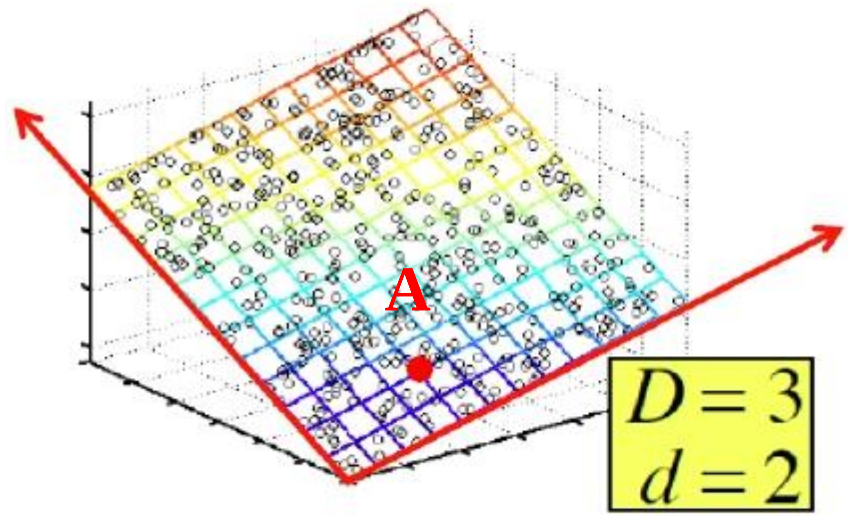
- 理论证明：
  - $A=U\Sigma V^T$ 和 $B=USV^T$，其中S是r*r的对角矩阵，
  - $s_i=\sigma_i\ (i=1\ldots k)$ ，剩下的 $s_i=0$
  - 于是B是rank(B)=k的情况下对A的最好的近似。
- 所谓“最好”是指：
  - B是 $\|A-B\|_F$ 在rank(B)=k情况下取最小值的解



保留多少奇异值

- 80-90%的能量 $=\sum_i \sigma_i^2$
  - ——一般经验值

$$\begin{array}{cc}\begin{bmatrix}1 & 2 & 1\\ -2 & -3 & 1\\ 3 & 5 & 0\end{bmatrix} & \begin{matrix}\mathbf{A}\\ \mathbf{B}\\ \mathbf{C}\end{matrix}\end{array}$$



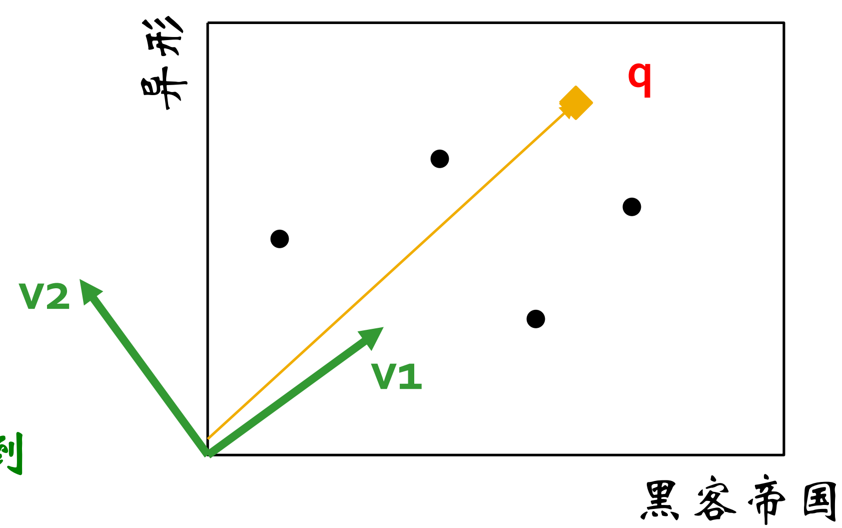
SVD:计算复杂度

- 完全计算复杂度：
  - $O(nm^2)$或$O(n^2m)$，取最小的。
- 但是我们可以减少计算量：
  - 如果只需要奇异值
  - 或者只需要前k个奇异向量
  - 或者矩阵是稀疏矩阵
- 一般用开源线性运算算法包：
  - LINPACK, SciPy,Matlab, SPlus, Mathematica

案例：给用户推荐电影

- Q:找到潜在喜欢“黑客帝国”电影的用户
- A:将用户的搜索统计映射到“主题空间”
  - ——怎么做？

$$q=\begin{array}{ccccc} \text{黑客帝国} & \text{异形} & \text{冲出宁静号} & \text{卡萨布兰卡} & \text{天使爱美丽} \\ \end{array}$$
$$q=\begin{bmatrix} 5 & 0 & 0 & 0 & 0 \end{bmatrix}$$

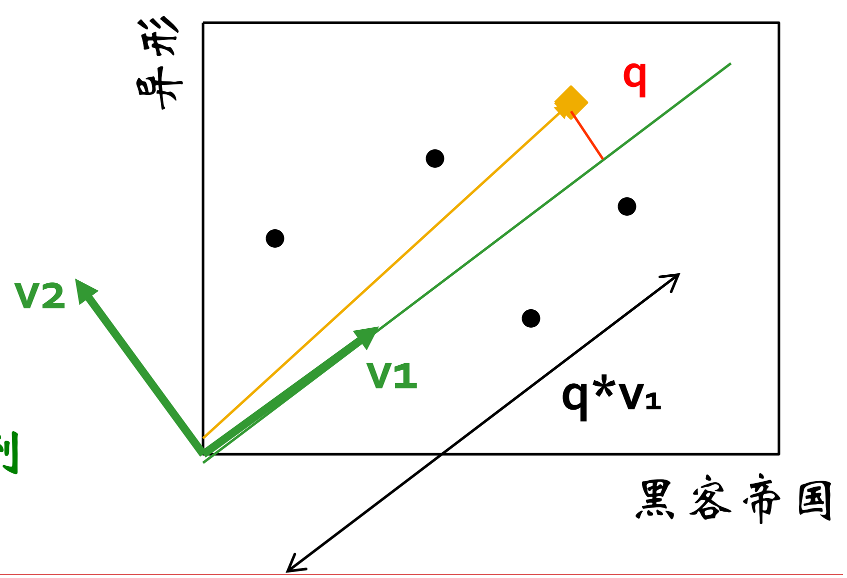


与主题向量的基底做内积，投影到“主题空间”中。

案例：给用户推荐电影

- Q:找到潜在喜欢“黑客帝国”电影的用户
- A:将用户的搜索统计映射到“主题空间”
  - ——怎么做？

$$
q = \begin{array}{ccccc} \text{黑客帝国} & \text{异形} & \text{冲出宁静号} & \text{卡萨布兰卡} & \text{天使爱美丽} \\ \end{array}
$$
$$
q = \begin{bmatrix} 5 & 0 & 0 & 0 & 0 \end{bmatrix}
$$



与主题向量的基底做内积，投影到“主题空间”中。

案例：给用户推荐电影

- 也就是说，我们有
- $q_{主题}=qV$

$$
q=\begin{array}{ccccc} \text{黑客帝国} & \text{异形} & \text{冲出宁静号} & \text{卡萨布兰卡} & \text{天使爱美丽} \\ \end{array}
$$

$$
q=\begin{bmatrix} 5 & 0 & 0 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} 0.56 & 0.12 \\ 0.59 & -0.02 \\ 0.56 & 0.12 \\ 0.09 & -0.69 \\ 0.09 & -0.69 \end{bmatrix} = \begin{bmatrix} 2.8 & 0.6 \end{bmatrix}
$$

科幻-主题

电影-主题近似转换矩阵(V)

案例：给用户推荐电影

- 对于喜欢“异形”、“冲出宁静号”的用户d
- $d_{主题}=dV$

$$
\mathbf{d}= \begin{array}{ccccc} \text{黑客帝国} & \text{异形} & \text{冲出宁静号} & \text{卡萨布兰卡} & \text{天使爱美丽} \\ \end{array}
$$

$$
\begin{bmatrix} 0 & 4 & 5 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} 0.56 & 0.12 \\ 0.59 & -0.02 \\ 0.56 & 0.12 \\ 0.09 & -0.69 \\ 0.09 & -0.69 \end{bmatrix} = \begin{array}{cc} \text{科幻-主题} & \\ \end{array} \begin{bmatrix} 5.2 & 0.4 \end{bmatrix}
$$

电影-主题近似转换矩阵(V)

案例：给用户推荐电影

- 观察：用户d与用户q是相似的，虽然他们在原始坐标下是正交的！

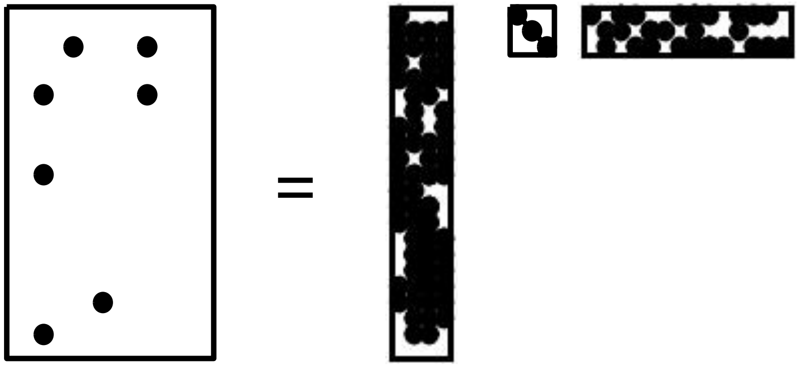
$$
\begin{array}{cccccc}
 & \text{黑客帝国} & \text{异形} & \text{冲出宁静号} & \text{卡萨布兰卡} & \text{天使爱美丽}
\end{array}
$$

$$q = \begin{bmatrix} 0 & 4 & 5 & 0 & 0 \end{bmatrix} \dashrightarrow \begin{array}{cc} \text{科幻-主题} & \\ \downarrow & \\ \end{array}$$
$$\begin{bmatrix} 2.8 & 0.6 \end{bmatrix}$$

$$d = \begin{bmatrix} 5 & 0 & 0 & 0 & 0 \end{bmatrix} \dashrightarrow \begin{bmatrix} 5.2 & 0.4 \end{bmatrix}$$

SVD降维的特点

- 奇异向量
  - 每一个奇异向量是所有输入矩阵的行向量或列向量的线性组合
- 奇异向量是稠密的

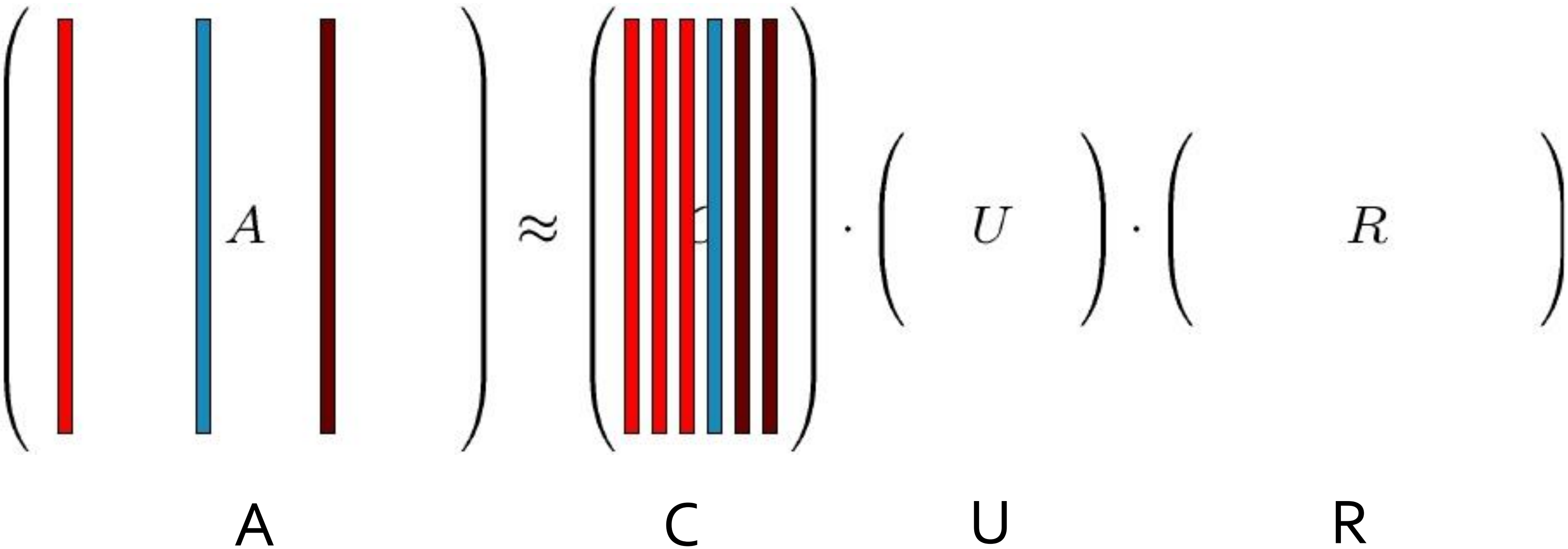


SVD代码

```
>>> import numpy as np
>>> from scipy import linalg
>>> A = np.array([[1,2,3],[4,5,6]])
>>> A
array([[1, 2, 3],
       [4, 5, 6]])
>>> M,N = A.shape
>>> U,s,Vh = linalg.svd(A)
>>> Sig = linalg.diagsvd(s,M,N)
>>> U, Vh = U, Vh
>>> U
array([[-0.3863177 , -0.92236578],
       [-0.92236578,  0.3863177 ]])
>>> Sig
array([[ 9.508032  ,  0.        ,  0.        ],
       [ 0.        ,  0.77286964,  0.        ]])
>>> Vh
array([[-0.42866713, -0.56630692, -0.7039467 ],
       [ 0.80596391,  0.11238241, -0.58119908],
       [ 0.40824829, -0.81649658,  0.40824829]])
>>> U.dot(Sig.dot(Vh)) #check computation
array([[ 1.,  2.,  3.],
       [ 4.,  5.,  6.]])
```

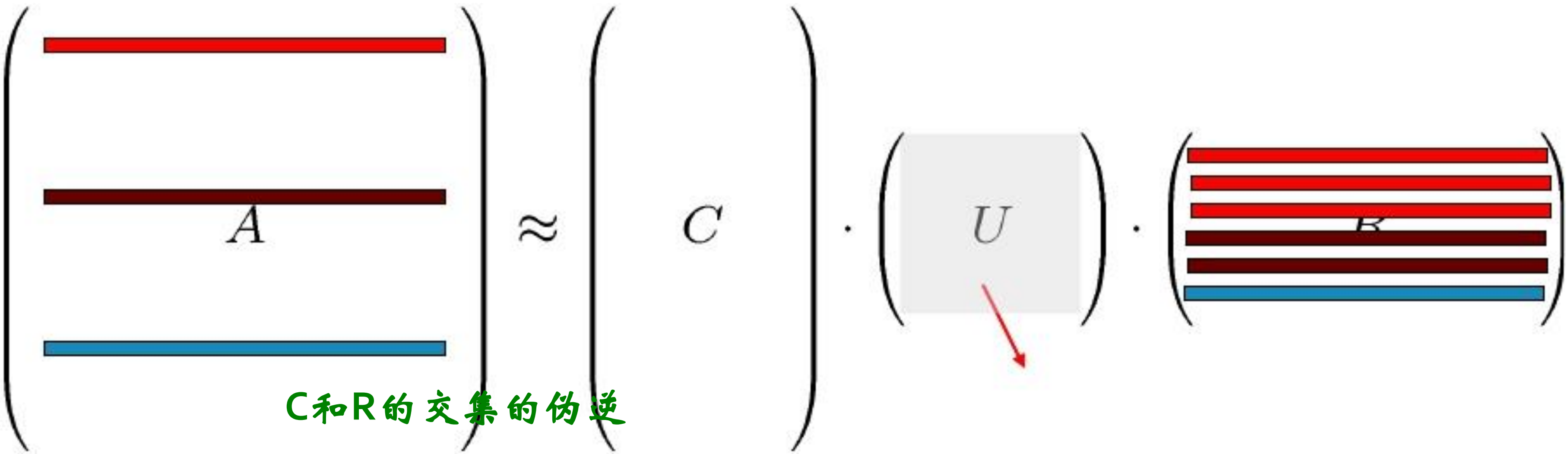
CUR分解简介

- 将矩阵A表示成C、U、R三个矩阵相乘
  - 使得 $\|A-C\cdot U\cdot R\|_F$ 变得很小。



CUR分解简介

- 将矩阵A表示成C、U、R三个矩阵相乘
  - 使得 $\|A\text{-}C\cdot U\cdot R\|_F$ 变得很小。



A C U R

SVD vs. CUR

稀疏且小

$$\text{SVD:}\quad A = U\ \Sigma\ V^T$$

大却稀疏

大且稠密

稠密却小

$$\text{CUR:}\quad A = C\ U\ R$$

大却稀疏

大却稀疏

感谢大家！

恳请大家批评指正！