LSI using SVD

Task

- SVD
- Examples on SVD
- LSI using SVD
- Code of LSI

Singular Value Decomposition

For an $m \times n$ matrix $\mathbf{A}$ of rank r there exists a factorization (Singular Value Decomposition = **SVD**) as follows:

$$A = U \Sigma V^T$$

$m \times m$ | $m \times n$ | V is $n \times n$

The columns of ***U*** are orthogonal eigenvectors of $\boldsymbol{AA^T}$.

The columns of ***V*** are orthogonal eigenvectors of $\boldsymbol{A^TA}$.

Eigenvalues $\lambda_1 \ldots \lambda_r$ of $\boldsymbol{AA^T}$ are the eigenvalues of $\boldsymbol{A^TA}$.

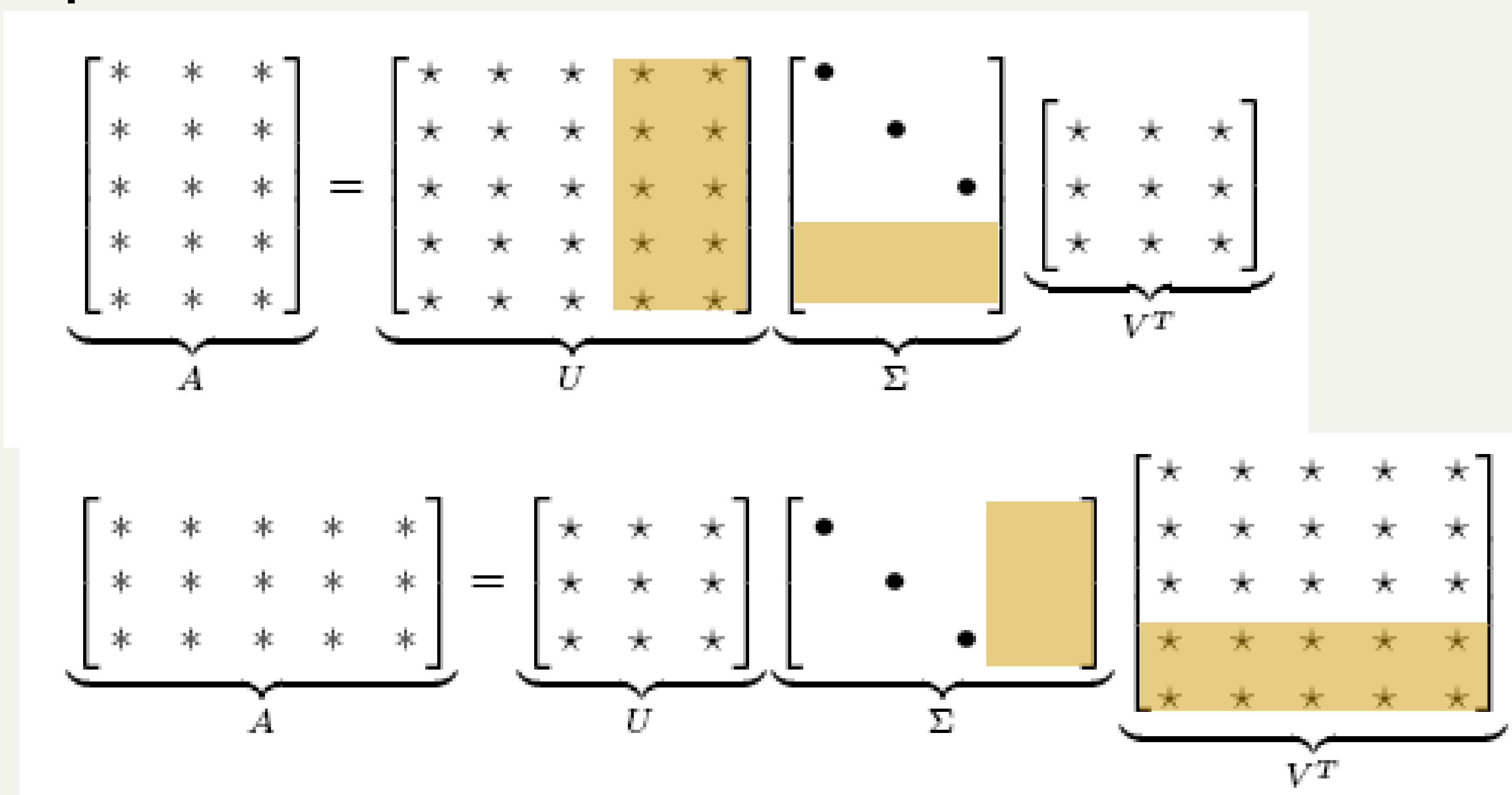
$$\sigma_i = \sqrt{\lambda_i}$$

$$\Sigma = diag(\sigma_1 \ldots \sigma_r)$$

Singular values.

Singular Value Decomposition

- Illustration of SVD dimensions and sparseness



Steps t o Compute SVD

1) Compute **U** matrix
   1) Compute AA^T
   2) Compute Eigen Values for AA^T and arrange in decreasing order.
   3) For each Eigen value compute Eigen vector and normalize it.
   4) Place each Eigen vector in U in order of decreasing Eigen values.
2) Compute V matrix

$$\mathbf{V} = \mathbf{A}^{\mathrm{T}}\mathbf{U}\mathbf{S}^{-1}$$

SVD example

Let $A = \begin{bmatrix} 1 & -1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$

Thus m=3, n=2. Its SVD is

$$\begin{bmatrix} 0 & 2/\sqrt{6} & 1/\sqrt{3} \\ 1/\sqrt{2} & -1/\sqrt{6} & 1/\sqrt{3} \\ 1/\sqrt{2} & 1/\sqrt{6} & -1/\sqrt{3} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \sqrt{3} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix}$$

Typically, the singular values arranged in decreasing order.

Low-rank Approximation

- SVD can be used to compute optimal **low-rank approximations**.
- Approximation problem: Find $\boldsymbol{A_k}$ of rank $\boldsymbol{k}$ such that

$$A_k = \min_{X:rank(X)=k} \|A - X\|_F$$

Frobenius norm

$$\|\mathsf{A}\|_F \equiv \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2}.$$

A_k and X are both $m \times n$ matrices.

Typically, want $k << r$.

LSI using SVD

- Consider 3 documents and Query:
- d1: Shipment of gold damaged in a fire.
- d2: Delivery of silver arrived in a silver truck.
- d3: Shipment of gold arrived in a truck.
- q: gold silver truck

$$
\begin{array}{l}
\text{Terms} \\
\text{a} \\
\text{arrived} \\
\text{damaged} \\
\text{delivery} \\
\text{fire} \\
\text{gold} \\
\text{in} \\
\text{of} \\
\text{shipment} \\
\text{silver} \\
\text{truck}
\end{array}
\qquad
\mathbf{A} =
\begin{array}{c}
\begin{array}{ccc} d1 & d2 & d3 \end{array} \\
\begin{bmatrix}
1 & 1 & 1 \\
0 & 1 & 1 \\
1 & 0 & 0 \\
0 & 1 & 0 \\
1 & 0 & 0 \\
1 & 0 & 1 \\
1 & 1 & 1 \\
1 & 1 & 1 \\
1 & 0 & 1 \\
0 & 2 & 0 \\
0 & 1 & 1
\end{bmatrix}
\end{array}
\qquad
\mathbf{q} =
\begin{array}{c}
q \\
\begin{bmatrix}
0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1
\end{bmatrix}
\end{array}
$$

Figure 2. Term-document matrix and query matrix example.

$$
U = \begin{bmatrix}
-0.4201 & 0.0748 & -0.0460 \\
-0.2995 & -0.2001 & 0.4078 \\
-0.1206 & 0.2749 & -0.4530 \\
-0.1576 & -0.3046 & -0.2006 \\
-0.1206 & 0.2749 & -0.4538 \\
-0.2626 & 0.3794 & 0.1547 \\
-0.4201 & 0.0748 & -0.0460 \\
-0.4201 & 0.0748 & -0.0460 \\
-0.2626 & 0.3794 & 0.1547 \\
-0.3151 & -0.6093 & -0.4013 \\
-0.2995 & -0.2001 & 0.4078
\end{bmatrix}
\qquad
S = \begin{bmatrix}
4.0989 & 0.0000 & 0.0000 \\
0.0000 & 2.3616 & 0.0000 \\
0.0000 & 0.0000 & 1.2737
\end{bmatrix}
$$

$$
V = \begin{bmatrix}
-0.4945 & 0.6492 & -0.5780 \\
-0.6458 & -0.7194 & -0.2556 \\
-0.5817 & 0.2469 & 0.7750
\end{bmatrix}
\qquad
V^{T} = \begin{bmatrix}
-0.4945 & 0.6458 & -0.5817 \\
0.6492 & -0.7194 & 0.2469 \\
-0.5780 & -0.2556 & 0.7750
\end{bmatrix}
$$

Figure 3. SVD results from the Bluebit Matrix Calculator.

$$k = 2$$

$$U \approx U_k = \begin{bmatrix} -0.4201 & 0.0748 \\ -0.2995 & -0.2001 \\ -0.1206 & 0.2749 \\ -0.1576 & -0.3046 \\ -0.1206 & 0.2749 \\ -0.2626 & 0.3794 \\ -0.4201 & 0.0748 \\ -0.4201 & 0.0748 \\ -0.2626 & 0.3794 \\ -0.3151 & -0.6093 \\ -0.2995 & -0.2001 \end{bmatrix} \qquad S \approx S_k = \begin{bmatrix} 4.0989 & 0.0000 \\ 0.0000 & 2.3616 \end{bmatrix}$$

$$V \approx V_k = \begin{bmatrix} -0.4945 & 0.6492 \\ -0.6458 & -0.7194 \\ -0.5817 & 0.2469 \end{bmatrix} \qquad V^T \approx V_k^T = \begin{bmatrix} -0.4945 & -0.6458 & -0.5817 \\ 0.6492 & -0.7194 & 0.2469 \end{bmatrix}$$

Figure 4. A Rank 2 Approximation.

Compute Vector for Document and Query in new Vector Space

- As we know $\mathbf{V} = \mathbf{A}^T\mathbf{U}\mathbf{S}^{-1}$
- Similarly we can compute vector for document and query as

$$\mathbf{d} = \mathbf{d}^T\mathbf{U}_k\mathbf{S}_k^{-1} \qquad \mathbf{q} = \mathbf{q}^T\mathbf{U}_k\mathbf{S}_k^{-1}$$

- Compute similarity as

$$\mathbf{sim}(\mathbf{q}, \mathbf{d}) = \mathbf{sim}(\mathbf{q}^T\mathbf{U}_k\mathbf{S}_k^{-1}, \mathbf{d}^T\mathbf{U}_k\mathbf{S}_k^{-1})$$

$$q = q^T U_k S_k^{-1} \qquad k = 2$$

$$q = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} -0.4201 & 0.0748 \\ -0.2995 & -0.2001 \\ -0.1206 & 0.2749 \\ -0.1576 & -0.3046 \\ -0.1206 & 0.2749 \\ -0.2626 & 0.3794 \\ -0.4201 & 0.0748 \\ -0.4201 & 0.0748 \\ -0.2626 & 0.3794 \\ -0.3151 & -0.6093 \\ -0.2995 & -0.2001 \end{bmatrix} \begin{bmatrix} \frac{1}{4.0989} & 0.0000 \\ 0.0000 & \frac{1}{2.3616} \end{bmatrix}$$

$$q = \begin{bmatrix} -0.2140 & -0.1821 \end{bmatrix}$$

Figure 5. Computing the query vector

Code LSI