Low-Rank Decompositions of Matrices

Objectives

- Define low-rank decomposition (matrix factorization)

- Explore applications

Matrices represent many types of information 2

1) Features in classification or modeling

2) User ratings

3) Collections of documents

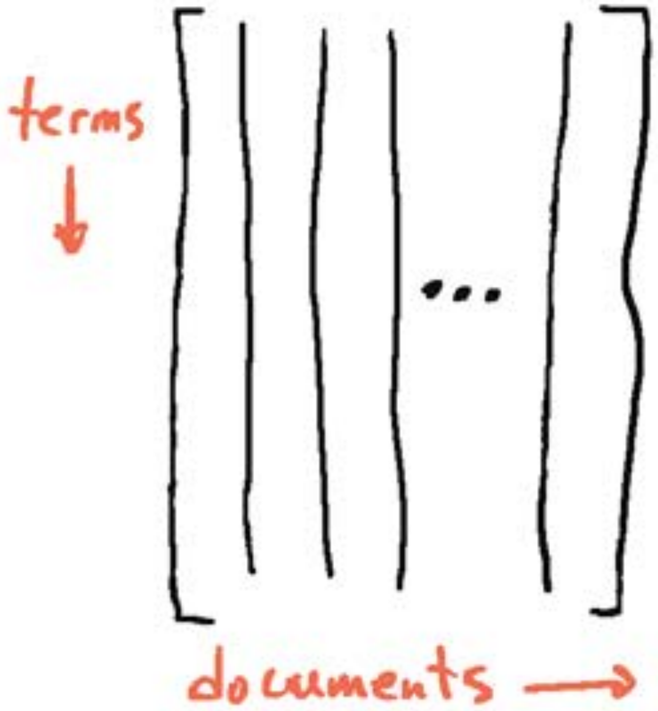
Bag of words model

apple
car
tree
May
blossom
road
agency
politics
happy
currency

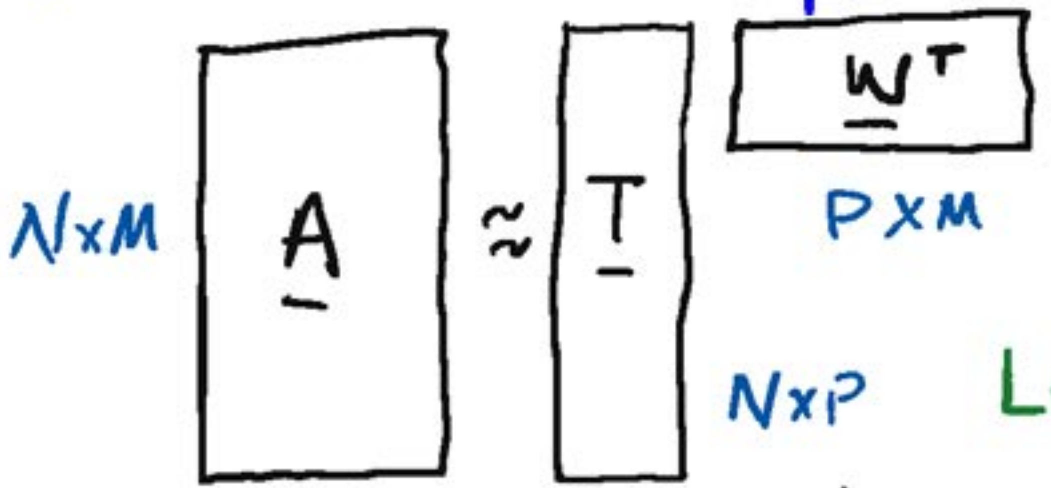
"Apple trees blossom in May."

$$
\begin{array}{l|c}
\text{word} & \text{frequency} \\
\hline
\text{agency} & 0 \\
\text{apple} & 1 \\
\text{blossom} & 1 \\
\text{car} & 0 \\
\text{currency} & 0 \\
\text{happy} & 0 \\
\text{May} & 1 \\
\text{politics} & 0 \\
\text{road} & 0 \\
\text{tree} & 1
\end{array}
$$

Term-document matrix



Low-rank decompositions emphasize patterns

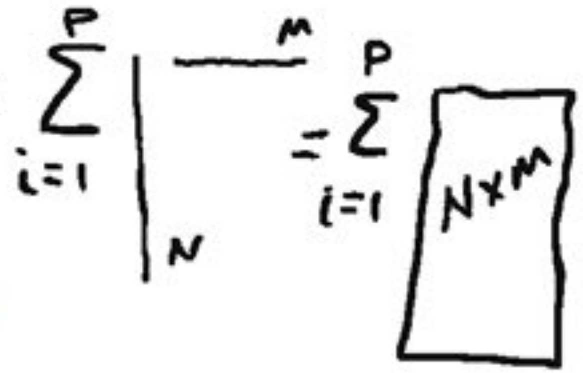


☆ $P<N,\ P<M \quad \operatorname{rank}\underline{T} = \operatorname{rank}\underline{W} = P$

$\Rightarrow \operatorname{rank}(\underline{T}\,\underline{W}^T) = P$

Let $\underline{T} = [\underline{t}_1\ \underline{t}_2 \cdots \underline{t}_P]$, $\underline{W} = [\underline{w}_1 \cdots \underline{w}_P]$

$$\underline{T}\,\underline{W}^T = \underbrace{\left[\underline{t}_1\ \underline{t}_2 \cdots \underline{t}_P\right]}_{\text{column patterns}} \underbrace{\begin{bmatrix} -\underline{w}_1^T- \\ -\underline{w}_2^T- \\ \vdots \\ -\underline{w}_P^T- \end{bmatrix}}_{\text{row patterns}} = \sum_{i=1}^{P} \underbrace{\underline{t}_i\,\underline{w}_i^T}_{\text{rank-one patterns}} = \sum_{i=1}^{P}$$



Finding patterns — $\xrightarrow{\quad}$ 3 ways

1) $\min_{\underline{T},\underline{W}} \|\underline{A} - \underline{T}\,\underline{W}^T\|$

singular value decomposition

2) $\underline{A} \approx \underline{T}\,\underline{W}^T,\ \underline{T},\underline{W} \geq 0$

nonnegative matrix factorization

3) $\underline{A} \approx \underline{T}\,\underline{W}^T$

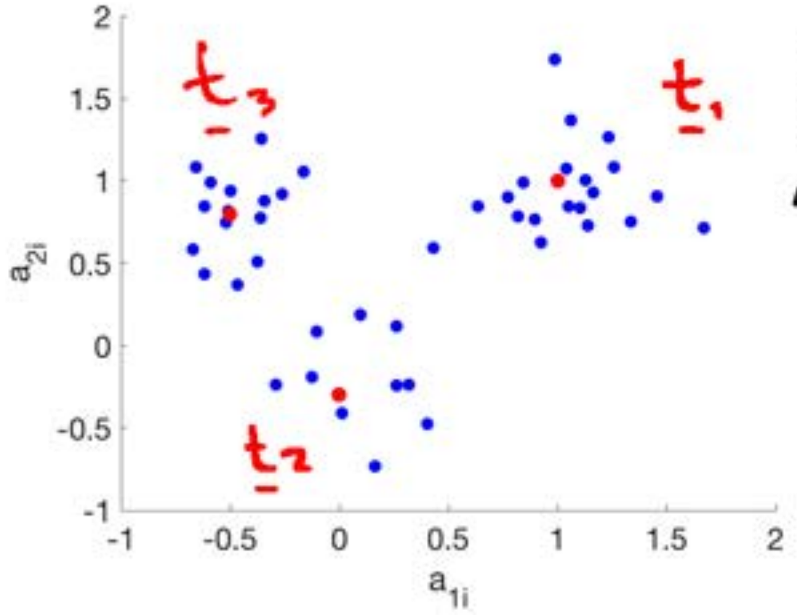
each col $\underline{W}^T$ all 0 w. single 1

clustering

Clustering groups similar columns

$$\begin{bmatrix} \underline{a}_1 & \underline{a}_2 & \underline{a}_3 & \underline{a}_4 \cdots \underline{a}_m \end{bmatrix} \approx \begin{bmatrix} \underline{t}_1 & \underline{t}_2 & \underline{t}_3 \end{bmatrix} \begin{bmatrix} 0 & 1 & 1 & 0 & & 0 \\ 1 & 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & 1 & & 0 \end{bmatrix}$$

$$\Rightarrow \underline{a}_1 \approx \underline{t}_2,\ \underline{a}_m \approx \underline{t}_2,\ \underline{a}_2 \approx \underline{t}_1,\ \underline{a}_3 \approx \underline{t}_1,\ \underline{a}_4 \approx \underline{t}_3 \cdots$$



N=2
P=3
M=45

Group similar documents, customers, products, etc

Many algorithms –

k-means

Low rank models "complete" missing data

$$\begin{array}{r|c|} & \text{Jill} \\ \text{Star Trek} & 8 \\ \text{Pride + Prejudice} & 3 \\ \text{The Martian} & 7 \\ \text{Sense + Sensibility} & 4 \\ \text{Empire Strikes Back} & ? \end{array}$$

Suppose $\underline{a} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} w_1 + \begin{bmatrix} 1 \\ -1 \\ 1 \\ -1 \\ 1 \end{bmatrix} w_2$

Use known ratings to solve w_1, w_2

$$\begin{bmatrix} 1 & 1 \\ 1 & -1 \\ 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} 8 \\ 3 \\ 7 \\ 4 \end{bmatrix} \Rightarrow \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} 5.5 \\ 2 \end{bmatrix}$$

Predict ratings using w_1, w_2

$$\underline{\hat{a}} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} 5.5 + \begin{bmatrix} 1 \\ -1 \\ 1 \\ -1 \\ 1 \end{bmatrix} 2 = \begin{bmatrix} 7.5 \\ 3.5 \\ 7.5 \\ 3.5 \\ 7.5 \end{bmatrix}$$

Use of "patterns" can mitigate noise

Noisy data $\underline{A}_m = \underline{A}_t + \underline{E}$

strong patterns

no dominant patterns

Low-rank model $\underline{\hat{A}}_m = \underline{T}\,\underline{W}^T$ can be closer to $\underline{A}_t$ than $\underline{A}_m$

$X^T = t^T W^T$

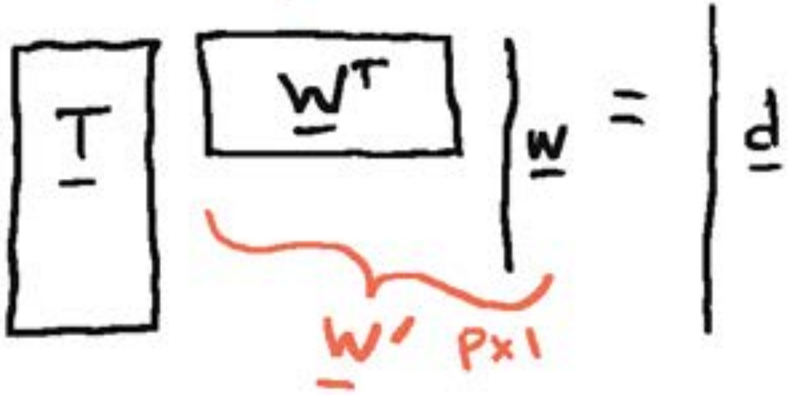
$t^T = x^T W (W^T W)^{-1}$

Low rank classifier / model fit

$\underline{\hat{A}}_m \underline{w} = \underline{d}$

$T W^T w = d$

$T w' = d$



$\underline{T}\,\underline{w}' = \underline{d}$

transformed features $\underline{x}_i'^T$

$x_i'^T = \underline{x}_i^T \underline{W} (\underline{W}^T \underline{W})^{-1}$

new data

投影到低维

New feature $\underline{x}^T$:

1) $\underline{x}'^T = \underline{x}^T \underline{W} (\underline{W}^T \underline{W})^{-1}$

2) $\hat{d} = \text{sign}(\underline{x}'^T \underline{w}')$