Linear Algebra

주재걸
고려대학교 컴퓨터학과

Eigendecomposition in Machine Learning

$AA^T = A \; A^T$ (10, 10; 10, 3; 10, 3)

$A =$ 키 혈 몸 / 사1 2 3 / 10, $A^TA = $ 3 10 / 3 10 $= A^TA$ 3 3

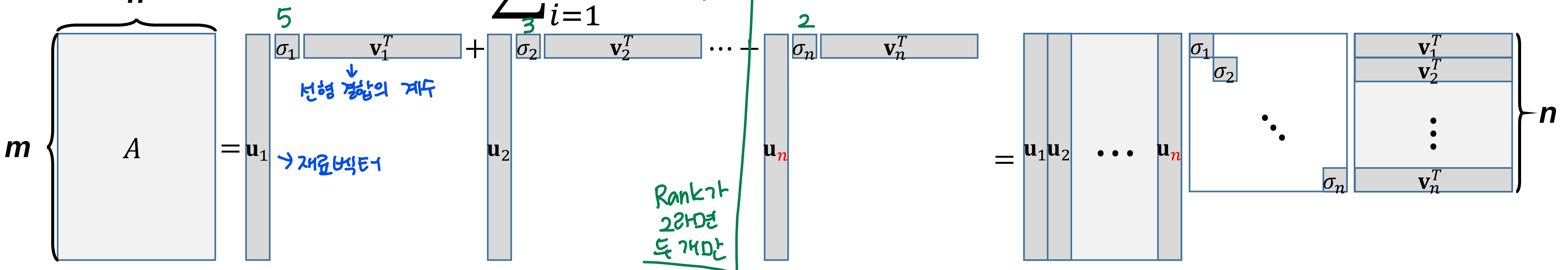
- In machine learning, we often handle symmetric positive (semi-)definite matrix.
- Given a (feature-by-data item) matrix $A \in \mathbb{R}^{m \times n}$,
- $A^T A$ represents a (data item-by-data item) similarity matrix between all pairs of data items, where the similarity is computed as an inner product
- Likewise, AA^T represents a (feature-by-feature) similarity matrix between all pairs of features, indicating a kind of correlations between features.
  - Covariance matrix in principal component analysis
  - Gram matrix in style transfer

Low-Rank Approximation of a Matrix

(feature by data)

- Recall a rectangular matrix $A \in \mathbb{R}^{m \times n}$, its SVD can be represented as the sum of outer products

$$A = U\Sigma V^T = \sum_{i=1}^{n} \sigma_i \mathbf{u}_i \mathbf{v}_i^T , \qquad \text{where } \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$$



- Consider the problem of the best low-rank approximation of A:

$$\hat{A}_r = \arg\min_{A_r} \| \underset{3}{\boxed{A}}{}_5 - \underset{3}{\boxed{A_r}}{}_5 \|_F \;\; \text{subject to } \operatorname{rank}(A_r) \leq r$$

if) rank=2 , A와 A_2의 차가 가장 작은 matrix

- The optimal solution is given as

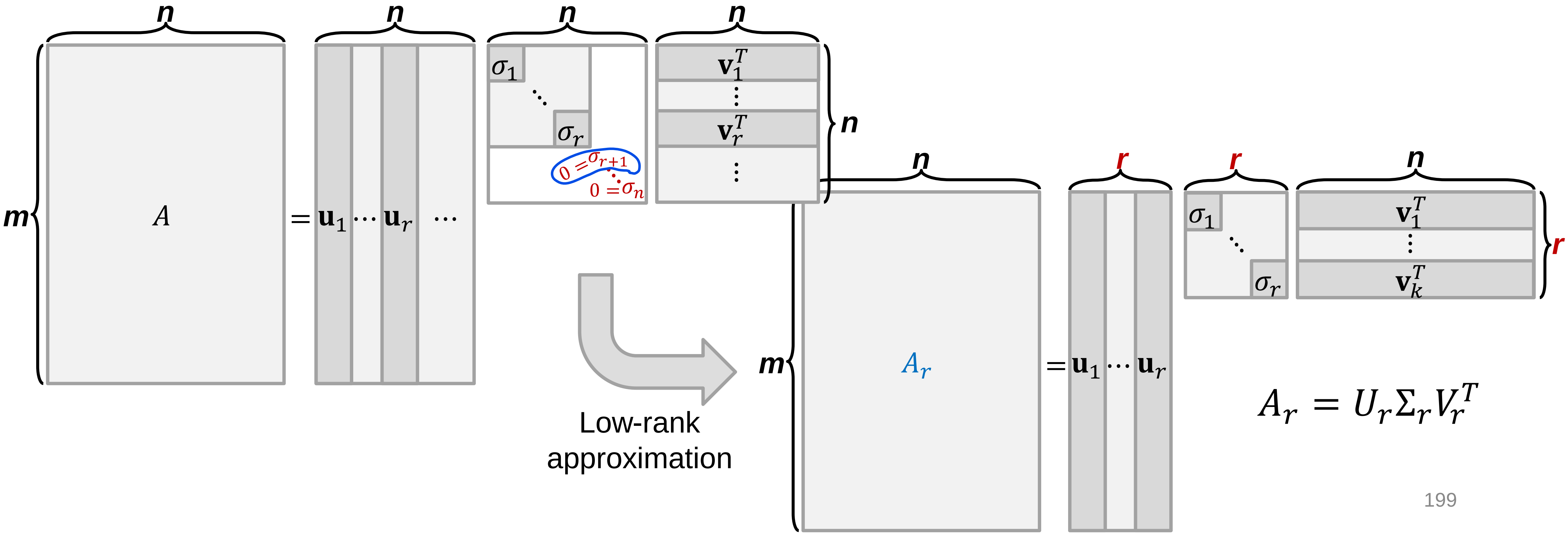
$$\hat{A}_r = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T , \qquad \text{where } \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$$

* 머신러닝에서는 뒤에 잘라서 근사값으로 하는 것이 성능 더 좋음

Low-Rank Approximation of a Matrix

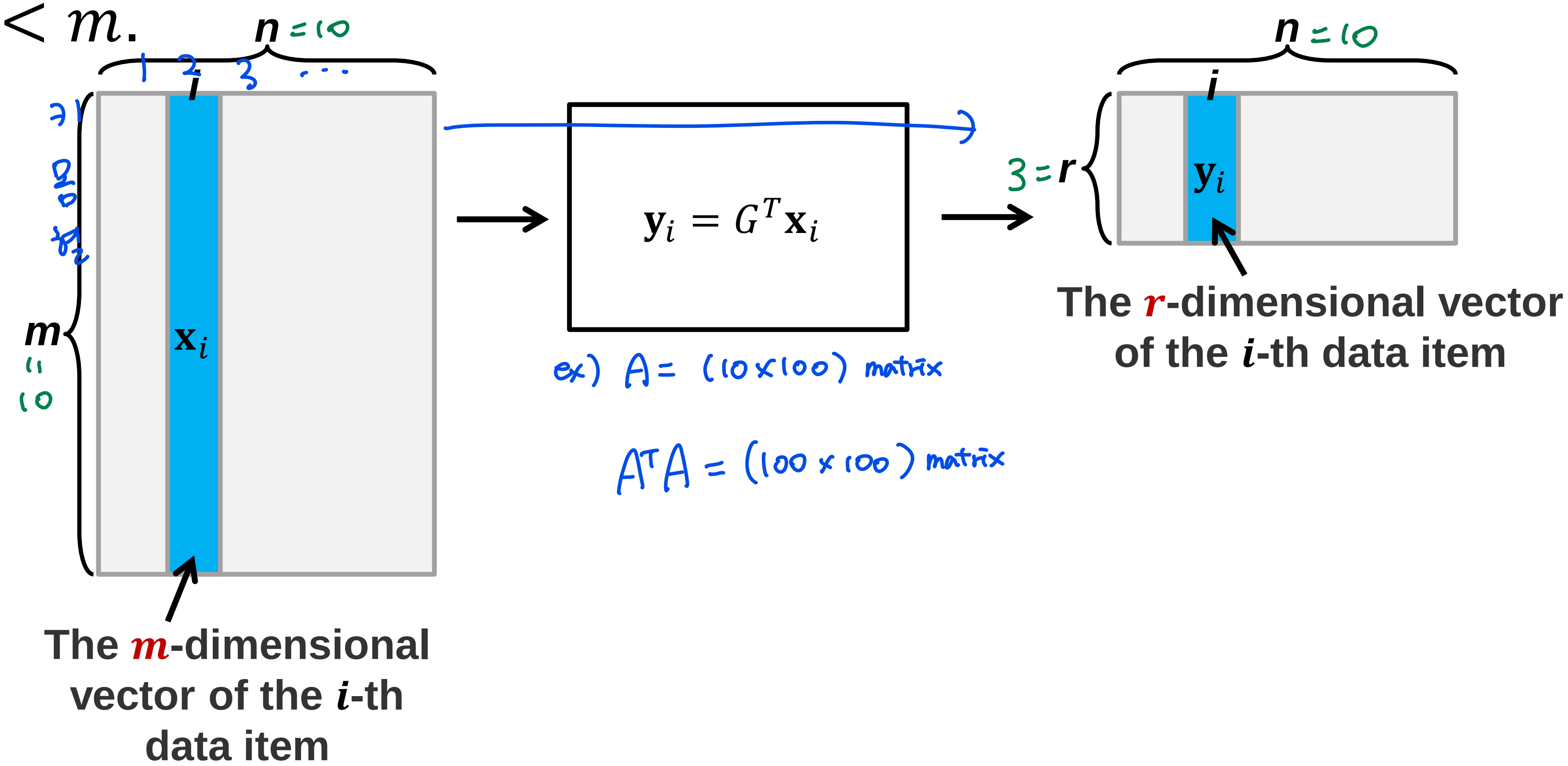
- We approximate A as A_r by setting $\sigma_i = 0$ for $\forall i \geq (r+1)$

$$A = \sum_{i=1}^{n} \sigma_i \mathbf{u}_i \mathbf{v}_i^T \simeq A_r = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T = U_r \Sigma_r V_r^T$$



Dimension-Reducing Transformation

- Given a (feature-by-data item) matrix $X \in \mathbb{R}^{m \times n}$, consider the linear transformation, $G^T : \mathbf{x} \in \mathbb{R}^m \mapsto \mathbf{y} \in \mathbb{R}^r$, where $G \in \mathbb{R}^{m \times r}$ and $r < m$.



ex) $A = (10 \times 100)$ matrix

$A^T A = (100 \times 100)$ matrix

Dimension-Reducing Transformation

$A = U\Sigma V^T$

10 3 10

길이 1인 수직

- Can we find the linear transformation, $\mathbf{y}_i = G^T\mathbf{x}_i$, where the columns of $G \in \mathbb{R}^{m\times r}$ are orthonormal, that best preserves the pairwise similarity between data items, $S = X^T X$?
- $Y = G^T X$, and their pairwise similarity is written as

100 3

$$Y^T Y = (G^T X)^T G^T X = X^T G G^T X$$

- Then, the above problem is written as

100 100 − 100 100

$$\hat{G} = \arg\min_{G} \| S - \underline{X^T G G^T X} \|_F \ \text{ subject to } G^T G = I_k$$

선형 변환으로 만든 것의 pairwise similarity matrix

- Given $X = U\Sigma V^T = \sum_{i=1}^{n} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$, where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$, the optimal solution is given as

$$\hat{G} = U_r = [\mathbf{u}_1 \quad \mathbf{u}_2 \quad \cdots \quad \mathbf{u}_r]$$

Dimension-Reducing Transformation

- In this case, $Y = \hat{G}^T X = U_r^T U \Sigma V^T = \Sigma_r V_r^T$.
- We can show that this generates the best solution for the best rank-r approximation of S.

Further Study

- Principal component analysis
  - http://www.math.union.edu/~jaureguj/PCA.pdf
  - http://pages.cs.wisc.edu/~jerryzhu/cs540/handouts/PCA.pdf

- Lectures on low-rank matrix factorization for topic modeling and word2vec
  - https://www.youtube.com/playlist?list=PLep-kTP3NkcNqn2MtzkscRlTDYTiqKjzD

- Lecture on gram matrix in style transfer
  - https://youtu.be/VC-YFRSp7lM