DIMENSIONALITY REDUCTION

TRU CAO

HO CHI MINH CITY UNIVERSITY OF TECHNOLOGY
AND JOHN VON NEUMANN INSTITUTE

OUTLINE

- Review of vector and matrix calculus
- Linear Discriminant Analysis (LDA)
- Principal Component Analysis (PCA)
- Feature selection for text classification

REVIEW OF VECTOR AND MATRIX CALCULUS

- A vector as a column matrix
- Dot product in matrix notation:

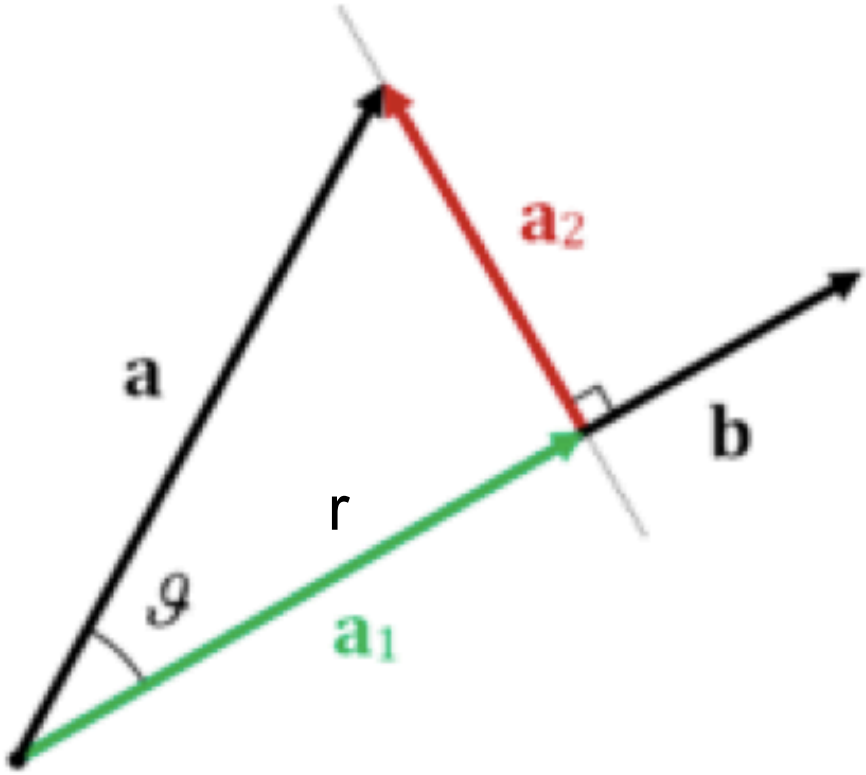
$$\mathbf{a}^T.\mathbf{b}$$

REVIEW OF VECTOR AND MATRIX CALCULUS

- Vector projection:

$$\mathbf{a}_1 = r.\frac{\mathbf{b}}{\|\mathbf{b}\|}$$

$$r = \frac{\mathbf{a}.\mathbf{b}}{\|\mathbf{b}\|}$$

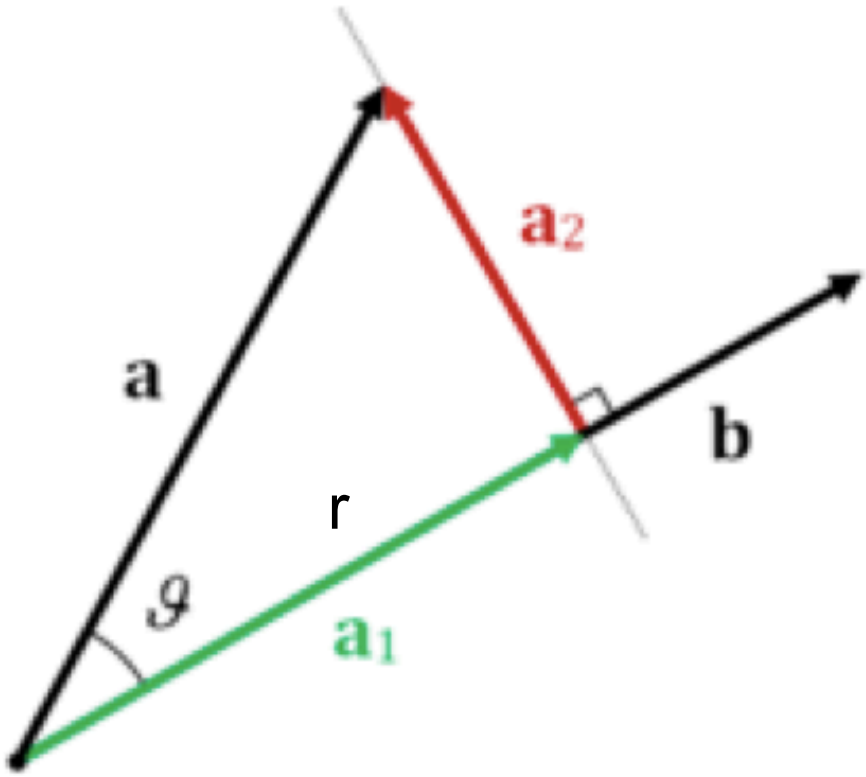


REVIEW OF VECTOR AND MATRIX CALCULUS

- Vector projection:

$$\mathbf{a}_1 = r.\frac{\mathbf{b}}{||\mathbf{b}||}$$

$$r = \frac{\mathbf{a}.\mathbf{b}}{||\mathbf{b}||}$$



a.b is a linear combination of **a**'s dimensions.

REVIEW OF VECTOR AND MATRIX CALCULUS

- Matrix differentiation:

$$\mathbf{y} = \Psi(\mathbf{x})$$

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}$$

y is an m×1 matrix, **x** is an 1×n matrix

REVIEW OF VECTOR AND MATRIX CALCULUS

- **Proposition 1:**

$$\alpha = \mathbf{x}^T\mathbf{A}\mathbf{x}$$

$$\frac{\partial \alpha}{\partial \mathbf{x}} = \mathbf{x}^T\left(\mathbf{A} + \mathbf{A}^T\right)$$

x is n×1, **A** is n×n, **A** does not depend on **x**

REVIEW OF VECTOR AND MATRIX CALCULUS

- **Proposition 1:**

$$\alpha = \mathbf{x}^T\mathbf{A}\mathbf{x}$$

$$\frac{\partial \alpha}{\partial \mathbf{x}} = \mathbf{x}^T\left(\mathbf{A} + \mathbf{A}^T\right)$$

x is n×1, **A** is n×n, **A** does not depend on **x**

Proof:

$$\alpha = \sum_{j=1}^{n}\sum_{i=1}^{n} a_{ij}x_i x_j$$

$$\frac{\partial \alpha}{\partial x_k} = \sum_{j=1}^{n} a_{kj}x_j + \sum_{i=1}^{n} a_{ik}x_i$$

$$\frac{\partial \alpha}{\partial \mathbf{x}} = \mathbf{x}^T\mathbf{A}^T + \mathbf{x}^T\mathbf{A} = \mathbf{x}^T\left(\mathbf{A}^T + \mathbf{A}\right)$$

REVIEW OF VECTOR AND MATRIX CALCULUS

- **Proposition 2: A** is symmetric

$$\alpha = \mathbf{x}^T\mathbf{A}\mathbf{x}$$

$$\frac{\partial \alpha}{\partial \mathbf{x}} = 2\mathbf{x}^T\mathbf{A}$$

x is n×1, **A** is n×n, **A** does not depend on **x**

REVIEW OF VECTOR AND MATRIX CALCULUS

- **Proposition 3: A** is symmetric

$$\alpha = \mathbf{x}^T\mathbf{A}\mathbf{x}$$

$$\left(\frac{\partial \alpha}{\partial \mathbf{x}}\right)^T = 2\mathbf{A}\mathbf{x}$$

x is n×1, **A** is n×n, **A** does not depend on **x**

REVIEW OF VECTOR AND MATRIX CALCULUS

- Eigenvalues and eigenvectors:

$$\mathbf{Av} = \lambda\mathbf{v}$$

A is n×n (linear transformation)

v is n×1

λ is an eigenvalue of **A**'s

v is an eigenvector of **A**'s

REVIEW OF VECTOR AND MATRIX CALCULUS

- Eigenvalues and eigenvectors:

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$$

A is n×n (linear transformation)

v is n×1

λ is an eigenvalue of **A**'s

v is an eigenvector of **A**'s

A stretches **v** by an amount of λ

REVIEW OF VECTOR AND MATRIX CALCULUS

- Example:

$$\mathbf{Av} = \lambda\mathbf{v}$$

$$\mathbf{A} = \begin{bmatrix} 19 & 20 & -16 \\ 20 & 13 & 4 \\ -16 & 4 & 31 \end{bmatrix} \quad \mathbf{v} = \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix} \quad \lambda = 27$$

REVIEW OF VECTOR AND MATRIX CALCULUS

- To find eigenvalues:

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{v} = \mathbf{0}$$

$$\det(\mathbf{A} - \lambda\mathbf{I}) = 0$$

REVIEW OF VECTOR AND MATRIX CALCULUS

- To find eigenvalues:

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{v} = \mathbf{0}$$

$$\det(\mathbf{A} - \lambda\mathbf{I}) = 0$$

- Example:

$$\mathbf{A} = \begin{bmatrix} .8 & .3 \\ .2 & .7 \end{bmatrix} \quad \det(\mathbf{A} - \lambda\mathbf{I}) = \det \begin{bmatrix} .8 - \lambda & .3 \\ .2 & .7 - \lambda \end{bmatrix}$$

$$\lambda^2 - \frac{3}{2}\lambda + \frac{1}{2} = 0 \Rightarrow \lambda_1 = 1 \text{ and } \lambda_2 = 1/2$$

REVIEW OF VECTOR AND MATRIX CALCULUS

- To find eigenvectors:

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{v} = \mathbf{0}$$

REVIEW OF VECTOR AND MATRIX CALCULUS

- To find eigenvectors:

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{v} = \mathbf{0}$$

- Example:

$$\mathbf{A} = \begin{bmatrix} .8 & .3 \\ .2 & .7 \end{bmatrix} \qquad \lambda_1 = 1 \qquad \lambda_2 = 1/2$$

$$(\mathbf{A} - \mathbf{I})\mathbf{v} = \mathbf{0} \Rightarrow \mathbf{v}_1 = \begin{bmatrix} .6 \\ .4 \end{bmatrix} \quad (\mathbf{A} - \tfrac{1}{2}\mathbf{I})\mathbf{v} = \mathbf{0} \Rightarrow \mathbf{v}_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

REVIEW OF VECTOR AND MATRIX CALCULUS

- **Proposition 4: A** is a $n \times n$ symmetric matrix
  - All of its eigenvalues are real, and
  - There are n linearly independent eigenvectors for A.

REVIEW OF VECTOR AND MATRIX CALCULUS

- **Proposition 5:** $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ are linearly independent eigenvectors of $\mathbf{A}$, and $\lambda_1, \lambda_2, \ldots, \lambda_n$ are their corresponding eigenvalues

$$\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^{-1}$$

where

$$\mathbf{P} = [\mathbf{v}_1 \ \mathbf{v}_2 \ \cdots \ \mathbf{v}_n]$$

$$\mathbf{D} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}$$

REVIEW OF VECTOR AND MATRIX CALCULUS

- **Proposition 5:** $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ are linearly independent eigenvectors of $\mathbf{A}$, and $\lambda_1, \lambda_2, \ldots, \lambda_n$ are their corresponding eigenvalues

$$\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^{-1}$$

Proof:

$\mathbf{P} = [\mathbf{v}_1 \ \mathbf{v}_2 \ \ldots \ \mathbf{v}_n]$ $\quad$ $\mathbf{P}^{-1}\mathbf{P} = \mathbf{I}$ $\quad$ $\mathbf{P}^{-1}\mathbf{v}_i = \mathbf{e}_i$

$\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{P}^{-1}\mathbf{A}[\mathbf{v}_1 \ \mathbf{v}_2 \ \ldots \ \mathbf{v}_n] = \mathbf{P}^{-1}[\mathbf{A}\mathbf{v}_1 \ \mathbf{A}\mathbf{v}_2 \ \ldots \ \mathbf{A}\mathbf{v}_n]$

$= \mathbf{P}^{-1}[\lambda_1\mathbf{v}_1 \ \lambda_2\mathbf{v}_2 \ \ldots \ \lambda_n\mathbf{v}_n] = [\lambda_1\mathbf{P}^{-1}\mathbf{v}_1 \ \lambda_2\mathbf{P}^{-1}\mathbf{v}_2 \ \ldots \ \lambda_n\mathbf{P}^{-1}\mathbf{v}_n]$

$= [\lambda_1\mathbf{e}_1 \ \lambda_2\mathbf{e}_2 \ \ldots \ \lambda_n\mathbf{e}_n] = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}$

REVIEW OF VECTOR AND MATRIX CALCULUS

- Example:

$$\mathbf{A} = \begin{bmatrix} 19 & 20 & -16 \\ 20 & 13 & 4 \\ -16 & 4 & 31 \end{bmatrix}$$

REVIEW OF VECTOR AND MATRIX CALCULUS

- Example:

$$\mathbf{A} = \begin{bmatrix} 19 & 20 & -16 \\ 20 & 13 & 4 \\ -16 & 4 & 31 \end{bmatrix}$$

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix} \quad \mathbf{v}_2 = \begin{bmatrix} -2 \\ -1 \\ 2 \end{bmatrix} \quad \mathbf{v}_3 = \begin{bmatrix} 2 \\ -2 \\ 1 \end{bmatrix}$$

$$\lambda_1 = 27 \qquad \lambda_2 = 45 \qquad \lambda_3 = -9$$

REVIEW OF VECTOR AND MATRIX CALCULUS

- Example:

$$\mathbf{A} = \begin{bmatrix} 19 & 20 & -16 \\ 20 & 13 & 4 \\ -16 & 4 & 31 \end{bmatrix}$$

$$\mathbf{P} = \begin{bmatrix} 1 & -2 & 2 \\ 2 & -1 & -2 \\ 2 & 2 & 1 \end{bmatrix} \qquad \mathbf{P}^{-1} = \frac{1}{9}\begin{bmatrix} 1 & 2 & 2 \\ -2 & -1 & 2 \\ 2 & -2 & 1 \end{bmatrix}$$

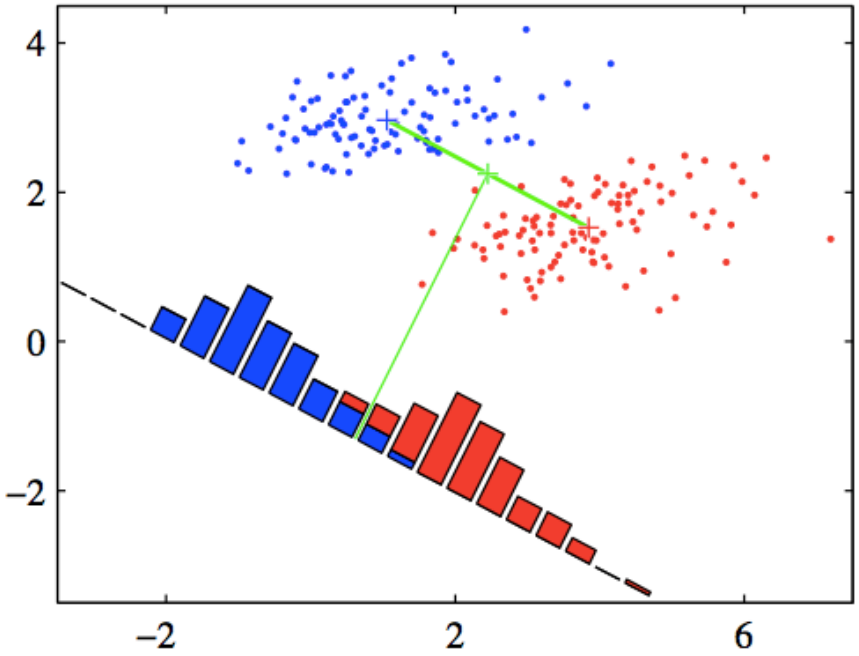
$$\lambda_1 = 27 \qquad \lambda_2 = 45 \qquad \lambda_3 = -9$$

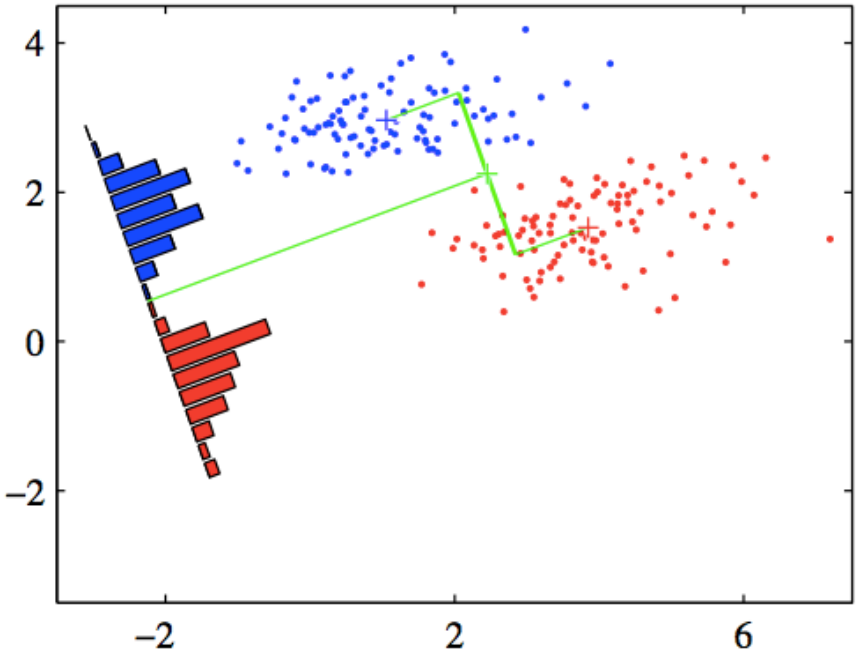
DIMENSIONALITY REDUCTION

- Case 1: feature combinations that are sufficient to classify samples.

DIMENSIONALITY REDUCTION

- Example 1: linear combination of features.



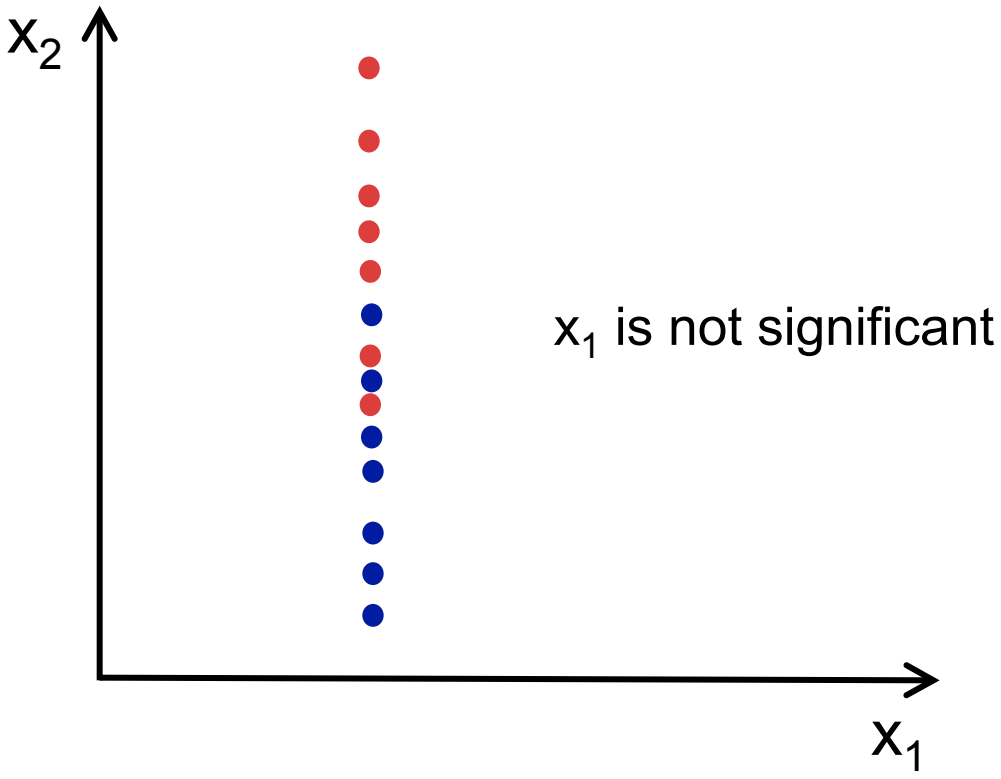


DIMENSIONALITY REDUCTION

- Case 2: a feature with invariant values over the samples is not useful for classification.

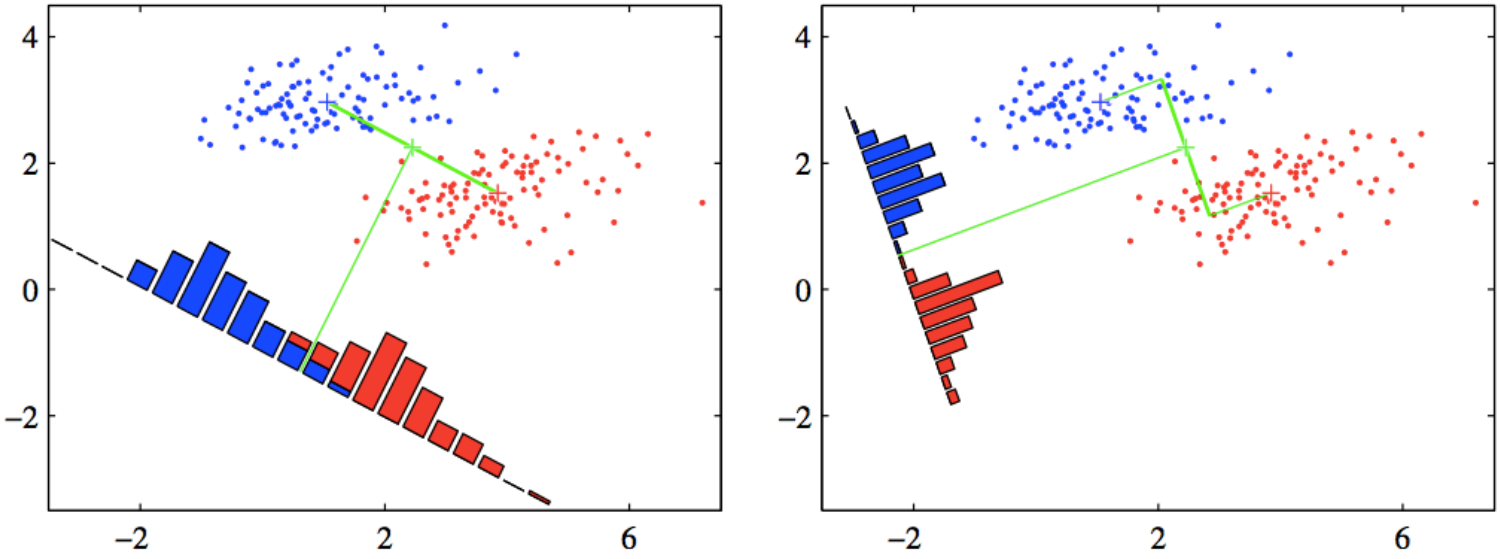
DIMENSIONALITY REDUCTION

- Example 2:



LINEAR DISCRIMINANT ANALYSIS (LDA) Phân tích phân biệt tuyến tính

- To project a high dimensional vector to one dimension (linear combination of features) so that:
  - The between-class distance is maximized.
  - The within-class variance is minimized.



LINEAR DISCRIMINANT ANALYSIS (LDA)

- To project a high dimensional vector to one dimension (linear combination of features) so that:
  - The between-class distance is maximized.
  - The within-class variance is minimized.

- To optimize **w** in:

$$y = \mathbf{w}^T\mathbf{x}$$

LINEAR DISCRIMINANT ANALYSIS (LDA)

- Two-class problem: N_1 points of class C_1 and N_2 points of class C_2.

LINEAR DISCRIMINANT ANALYSIS (LDA)

- Two-class problem: N_1 points of class C_1 and N_2 points of class C_2.
- Mean vectors:

$$\mathbf{m}_1 = \frac{1}{N_1} \sum_{n \in \mathcal{C}_1} \mathbf{x}_n, \qquad \mathbf{m}_2 = \frac{1}{N_2} \sum_{n \in \mathcal{C}_2} \mathbf{x}_n$$

LINEAR DISCRIMINANT ANALYSIS (LDA)

- Two-class problem: N_1 points of class C_1 and N_2 points of class C_2.
- Mean vectors:

$$\mathbf{m}_1 = \frac{1}{N_1} \sum_{n \in \mathcal{C}_1} \mathbf{x}_n, \qquad \mathbf{m}_2 = \frac{1}{N_2} \sum_{n \in \mathcal{C}_2} \mathbf{x}_n$$

- Mean of projected data:

$$m_k = \mathbf{w}^{\mathrm{T}} \mathbf{m}_k$$

LINEAR DISCRIMINANT ANALYSIS (LDA)

- Two-class problem: N_1 points of class C_1 and N_2 points of class C_2.
- Mean vectors:

$$\mathbf{m}_1 = \frac{1}{N_1} \sum_{n \in \mathcal{C}_1} \mathbf{x}_n, \qquad \mathbf{m}_2 = \frac{1}{N_2} \sum_{n \in \mathcal{C}_2} \mathbf{x}_n$$

- Mean of projected data:

$$m_k = \mathbf{w}^{\mathrm{T}} \mathbf{m}_k$$

- To be maximized:

$$(m_2 - m_1)^2$$

LINEAR DISCRIMINANT ANALYSIS (LDA)

- Two-class problem: N_1 points of class C_1 and N_2 points of class C_2.
- Within-class variance:

$$s_k^2 = \sum_{n \in \mathcal{C}_k} (y_n - m_k)^2$$

- To be minimized:

$$s_1^2 + s_2^2$$

LINEAR DISCRIMINANT ANALYSIS (LDA)

- To be maximized:

$$J(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}$$

LINEAR DISCRIMINANT ANALYSIS (LDA)

- To be maximized:

$$J(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}$$

- Between-class covariance matrix:

$$\mathbf{S}_\mathrm{B} = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^\mathrm{T}$$

- Within-class covariance matrix:

$$\mathbf{S}_\mathrm{W} = \sum_{n \in \mathcal{C}_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^\mathrm{T} + \sum_{n \in \mathcal{C}_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^\mathrm{T}$$

LINEAR DISCRIMINANT ANALYSIS (LDA)

- To be maximized:

$$J(\mathbf{w}) = \frac{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{B}}\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}\mathbf{w}}$$

LINEAR DISCRIMINANT ANALYSIS (LDA)

- To be maximized:

$$J(\mathbf{w}) = \frac{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{B}}\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}\mathbf{w}}$$

$$\partial J(\mathbf{w})/\partial \mathbf{w} = 0$$

$$\Rightarrow \; 2\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}(\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{B}}\mathbf{w}) = 2\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{B}}(\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}\mathbf{w})$$

$$\Rightarrow \; (\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{B}}\mathbf{w})\mathbf{S}_{\mathrm{W}}\mathbf{w} = (\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}\mathbf{w})\mathbf{S}_{\mathrm{B}}\mathbf{w}$$

$$\Rightarrow \; (\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{B}}\mathbf{w})\mathbf{w} = \mathbf{S}_{\mathrm{W}}^{-1}(\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}\mathbf{w})\mathbf{S}_{\mathrm{B}}\mathbf{w}$$

LINEAR DISCRIMINANT ANALYSIS (LDA)

- Solution:

$$(\mathbf{w}^T\mathbf{S}_B\mathbf{w})\mathbf{w} = \mathbf{S}_W^{-1}(\mathbf{w}^T\mathbf{S}_W\mathbf{w})\mathbf{S}_B\mathbf{w}$$

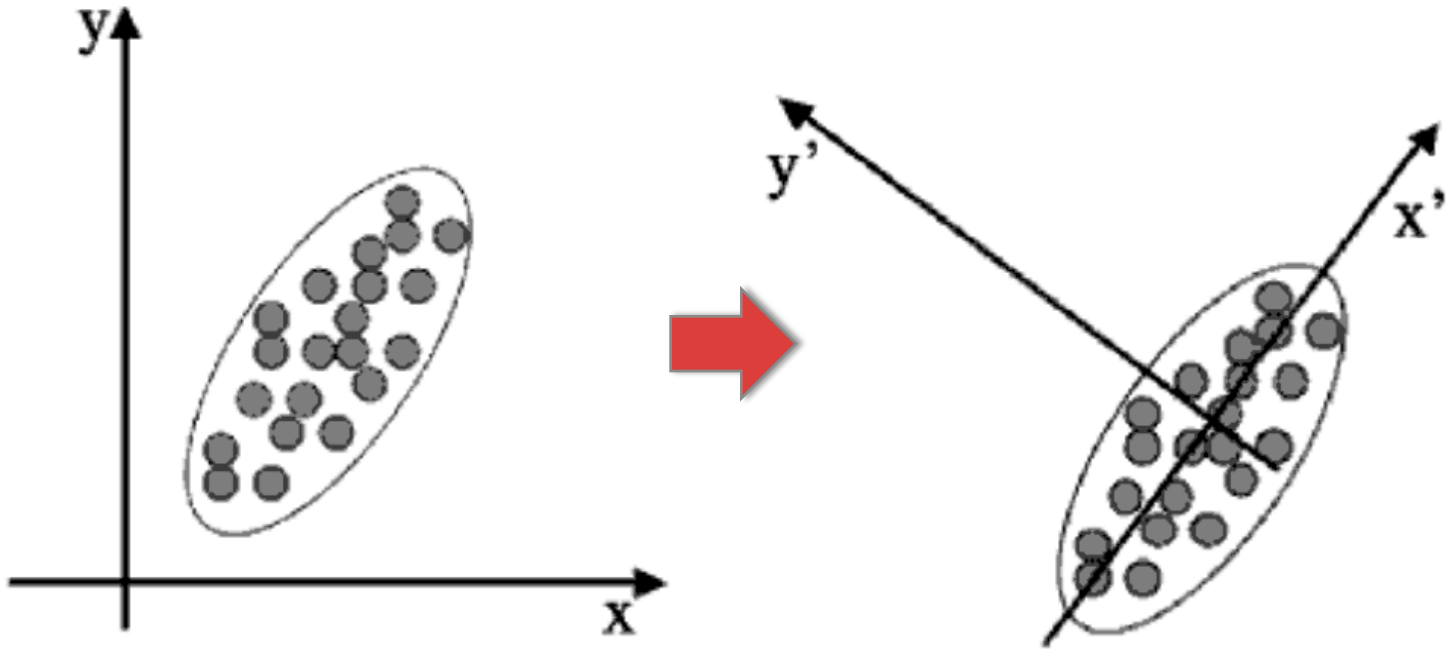
$(\mathbf{w}^T\mathbf{S}_B\mathbf{w})$ and $(\mathbf{w}^T\mathbf{S}_W\mathbf{w})$: scalar factors

$\mathbf{S}_B\mathbf{w} = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^T\mathbf{w}$: in the direction of $(\mathbf{m}_2 - \mathbf{m}_1)$

$\Rightarrow$ $\mathbf{w}$ is in the direction of $\mathbf{S}_W^{-1}(\mathbf{m}_2 - \mathbf{m}_1)$

PRINCIPAL COMPONENT ANALYSIS (PCA) phân tích nguyên tắc cấu thành

- To find the dimensions for which the projected data have the largest variance.



PRINCIPAL COMPONENT ANALYSIS (PCA)

- Consider N points of unlabeled data $\{\mathbf{x}_n\}$.

PRINCIPAL COMPONENT ANALYSIS (PCA)

- Consider N points of unlabeled data $\{\mathbf{x}_n\}$.
- Mean vector:

$$\mathbf{m} = \frac{1}{N} \sum_{n=1..N} \mathbf{x}_n$$

PRINCIPAL COMPONENT ANALYSIS (PCA)

- Consider N points of unlabeled data $\{\mathbf{x}_n\}$.
- Mean vector:

$$\mathbf{m} = \frac{1}{N} \sum_{n=1..N} \mathbf{x}_n$$

- Variance of projected data on dimension $\mathbf{u}_1$:

$$\frac{1}{N} \sum_{n=1..N} (\mathbf{u}_1^T.\mathbf{x}_n - \mathbf{u}_1^T.\mathbf{m})^2 = \mathbf{u}_1^T.\mathbf{S}.\mathbf{u}_1$$

where **S** is the data covariance matrix:

$$\mathbf{S} = \frac{1}{N} \sum_{n=1..N} (\mathbf{x}_n - \mathbf{m}).(\mathbf{x}_n - \mathbf{m})^T$$

PRINCIPAL COMPONENT ANALYSIS (PCA)

- To be maximized:

$$\frac{1}{N}\sum_{n=1..N}(\mathbf{u}_1^T.\mathbf{x}_n - \mathbf{u}_1^T.\mathbf{m})^2 = \mathbf{u}_1^T.\mathbf{S}.\mathbf{u}_1$$

with the constraint on the unit vector $\mathbf{u}_1$:

$$\mathbf{u}_1^T.\mathbf{u}_1 = 1$$

PRINCIPAL COMPONENT ANALYSIS (PCA)

- To be maximized:

$$\frac{1}{N}\sum_{n=1..N}(\mathbf{u}_1^T.\mathbf{x}_n - \mathbf{u}_1^T.\mathbf{m})^2 = \mathbf{u}_1^T.\mathbf{S}.\mathbf{u}_1$$

with the constraint on the unit vector $\mathbf{u}_1$:

$$\mathbf{u}_1^T.\mathbf{u}_1 = 1$$

- Lagrange function to be maximized:

$$\mathbf{u}_1^T.\mathbf{S}.\mathbf{u}_1 + \lambda_1.(1 - \mathbf{u}_1^T.\mathbf{u}_1)$$

PRINCIPAL COMPONENT ANALYSIS (PCA)

- To be maximized:

$$\frac{1}{N}\sum_{n=1..N}(\mathbf{u}_1^T.\mathbf{x}_n - \mathbf{u}_1^T.\mathbf{m})^2 = \mathbf{u}_1^T.\mathbf{S}.\mathbf{u}_1$$

with the constraint on the unit vector $\mathbf{u}_1$:

$$\mathbf{u}_1^T.\mathbf{u}_1 = 1$$

- Lagrange function to be maximized:

$$\mathbf{u}_1^T.\mathbf{S}.\mathbf{u}_1 + \lambda_1.(1 - \mathbf{u}_1^T.\mathbf{u}_1)$$

- Solution:

$$\mathbf{S}.\mathbf{u}_1 = \lambda_1.\mathbf{u}_1$$

PRINCIPAL COMPONENT ANALYSIS (PCA)

- Solution is an eigenvalue and eigenvector:

$$\mathbf{S}.\mathbf{u}_1 = \lambda_1.\mathbf{u}_1$$

- PCA:
  - Compute the data covariance matrix (square and symmetric) in the original space of D dimensions.
  - Find D eigenvalues and eigenvectors of the covariance matrix.
  - Select the largest M < D eigenvalues and the corresponding eigenvectors to be the new space.

PRINCIPAL COMPONENT ANALYSIS (PCA)

- Solution is an eigenvalue and eigenvector:

$$\mathbf{S}.\mathbf{u}_1 = \lambda_1.\mathbf{u}_1$$

- PCA:

$$\mathbf{S} = [\mathbf{u}_1 \ \mathbf{u}_2 \ \dots \ \mathbf{u}_D] \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}_D [\mathbf{u}_1 \ \mathbf{u}_2 \ \dots \ \mathbf{u}_D]^{-1}$$

FEATURE SELECTION FOR TEXT CLASSIFICATION

- This is important because of:
  - High dimensionality of text features.
  - Existence of irrelevant/noisy features (not only redundant, bt also with negative effects).

FEATURE SELECTION FOR TEXT CLASSIFICATION

- Representation of a document:
  - Bag of words.
  - Sequence of words (strings).
  - Plus grammatical and semantic elements.
  - Probabilistic distribution on topics (topic modeling).

FEATURE SELECTION FOR TEXT CLASSIFICATION

- The most basic and common feature filtering:
  - Removal of stop-words (the common words such as articles, conjunctions, prepositions, …).
  - Stemming: different forms (e.g. singular, plural, difference tenses, …) of the same word are consolidated into a single word.

FEATURE SELECTION FOR TEXT CLASSIFICATION

- The basic idea of word feature selection: retain only those words that discriminate document classes.
- How to measure the discriminative power of a word?

GINI INDEX

- Consider a word w and assume k class-labels.
- Let $p_n(w)$ be the fraction of those documents containing w that belong to class n:

$$p_n(w) = \text{Prob}(\text{a document} \in \text{class } n \mid \text{it contains } w)$$

$$\sum_{n=1..k} p_n(w) = 1$$

GINI INDEX

- Consider a word w and assume k class-labels.
- Let $p_n(w)$ be the fraction of those documents containing w that belong to class n:

$$p_n(w) = \text{Prob}(\text{a document} \in \text{class } n \mid \text{it contains } w)$$

$$\sum_{n=1..k} p_n(w) = 1$$

- Gini-index of w:

$$G(w) = \sum_{n=1..k} p_n(w)^2 \in [1/k, 1]$$

the higher, the greater discriminative power of w.

GINI INDEX

- Criticism: $p_n(w)$ may be biased when the global class distribution of documents is usually not uniform.

GINI INDEX

- Let P_n be the fraction of documents that belong to class n.
- Define the normalized probability $q_n(w)$:

$$q_n(w) = \frac{p_n(w)/P_n}{\sum_{m=1..k} p_m(w)/P_m}$$

GINI INDEX

- Let P_n be the fraction of documents that belong to class n.
- Define the normalized probability $q_n(w)$:

$$q_n(w) = \frac{p_n(w)/P_n}{\sum_{m=1..k} p_m(w)/P_m}$$

- New gini-index of w:

$$G(w) = \sum_{n=1..k} q_n(w)^2 \in [1/k, 1]$$

INFORMATION GAIN

- Prior inhomogeneity of a set of documents:

$$E = -\sum_{n=1..k} P_n.\log(P_n)$$

INFORMATION GAIN

- Prior inhomogeneity of a set of documents:

$$E = -\sum_{n=1..k} P_n.\log(P_n)$$

- Let F(w) be the fraction of documents that contains w.
- Inhomogeneity after using w:

$$E(w) = -F(w).\sum_{n=1..k} p_n(w).\log(p_n(w))$$

$$-(1-F(w)).\sum_{n=1..k} (1-p_n(w)).\log(1-p_n(w))$$

INFORMATION GAIN

- Prior inhomogeneity of a set of documents:

$$E = -\sum_{n=1..k} P_n.\log(P_n)$$

- Let F(w) be the fraction of documents that contains w.
- Inhomogeneity after using w:

$$E(w) = -F(w).\sum_{n=1..k} p_n(w).\log(p_n(w))$$

$$-(1-F(w)).\sum_{n=1..k} (1-p_n(w)).\log(1-p_n(w))$$

- Information gain:

$$I(w) = E - E(w)$$

MUTUAL INFORMATION

- How a word w and a class n are correlated?

MUTUAL INFORMATION

- How a word w and a class n are correlated?
- Expected co-occurrence of w and class n:

$$P_n.F(w)$$

- True co-occurrence of w and class n:

$$p_n(w).F(w)$$

MUTUAL INFORMATION

- How a word w and a class n are correlated?
- Expected co-occurrence of w and class n:

$$P_n.F(w)$$

- True co-occurrence of w and class n:

$$p_n(w).F(w)$$

- Mutual information of w and class n:

$$M_n(w) = \log\left(\frac{p_n(w).F(w)}{P_n.F(w)}\right) = \log\left(\frac{p_n(w)}{P_n}\right)$$

MUTUAL INFORMATION

- Mutual information of w and class n:

$$M_n(w) = \log\left(\frac{p_n(w).F(w)}{P_n.F(w)}\right) = \log\left(\frac{p_n(w)}{P_n}\right)$$

- $M_n(w) = 0$: w is not relevant to class n.

  $M_n(w) > 0$: w is positively correlated to class n.

  $M_n(w) < 0$: w is negatively correlated to class n.

MUTUAL INFORMATION

- Mutual information of w and class n:

$$M_n(w) = \log\left(\frac{p_n(w).F(w)}{P_n.F(w)}\right) = \log\left(\frac{p_n(w)}{P_n}\right)$$

- Average and maximum values of $M_n(w)$:

$$M_{avg}(w) = \sum_{n=1..k} P_n.M_n(w)$$

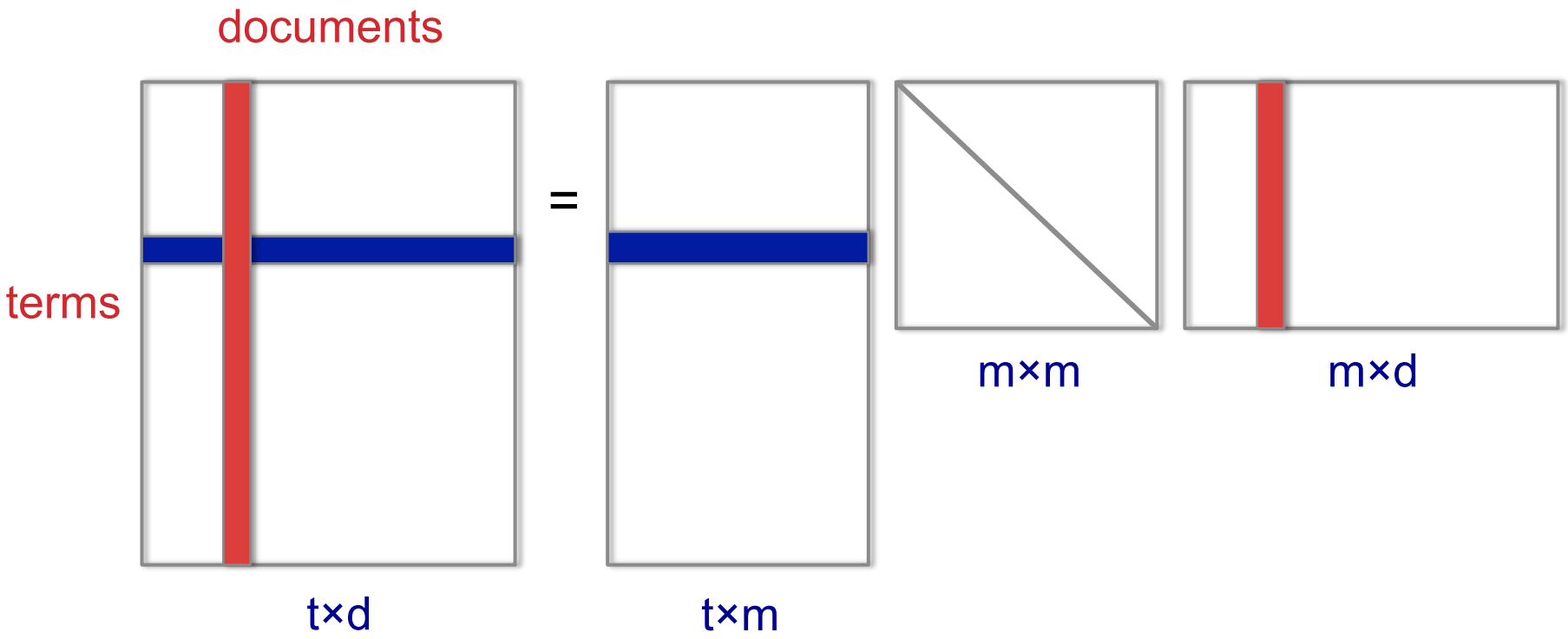
$$M_{max}(w) = \max_n\{M_n(w)\}$$

LATENT SEMANTIC INDEXING (LSI)

- Deerwester, S. et al. (1990), Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science, 41(6), 391-407.
- To find the latent/hidden semantic components that can represent a document.
- The number of those components is much less than the number of words in a document.

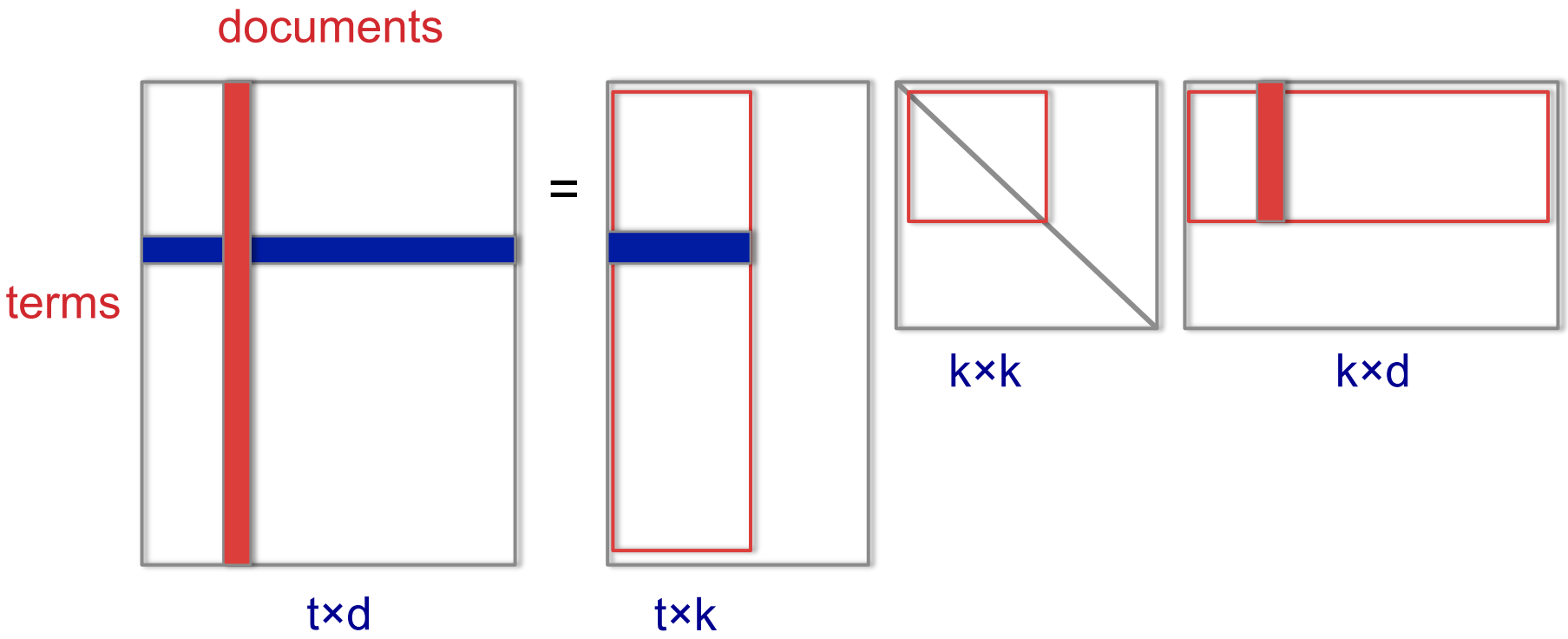
LATENT SEMANTIC INDEXING (LSI)

- The idea is similar to PCA:



LATENT SEMANTIC INDEXING (LSI)

- The idea is similar to PCA:



READING HOMEWORK

- Deerwester, S. et al. (1990), Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science, 41(6), 391-407.