CS5489
Lecture 7.2: Linear Dimensionality Reduction

Kede Ma

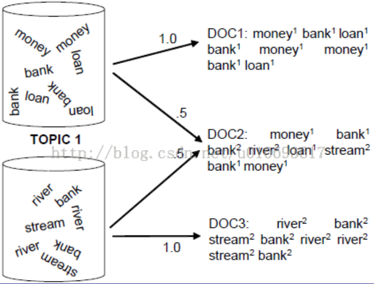
City University of Hong Kong (Dongguan)

Slide template by courtesy of Benjamin M. Marlin

Outline

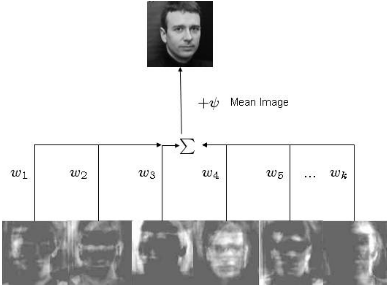
Dimensionality Reduction

- Transform high-dim vectors into low-dim vectors
  - Dimensions in the low-dim data represent co-occuring features in high-dim data
  - Dimensions in the low-dim data may have semantic meaning
- For example: document analysis
  - High-dim: bag-of-word vectors of documents
  - Low-dim: each dimension represents similarity to a topic



Example: Image Analysis

- Approximate an image as a weighted combination of several basis images
- Represent the image as the weights



Reasons for Dimensionality Reduction

- Preprocessing - make the dataset easier to use
- Reduce computational cost of running machine learning algorithms
- Can be used to "de-noise" data by projecting to lower-dim space and then projecting back to the original high-dim space
- Make the results easier to understand (visualization)

Dimensionality Reduction vs Feature Selection

- The goal of feature selection is to remove features that are not informative with respect to the class label. This obviously reduces the dimensionality of the feature space
- Dimensionality reduction can be used to find a meaningful lower-dim feature space even when there is information in each feature dimension so that none can be discarded
- Another important property of dimensionality reduction is that it is **unsupervised**

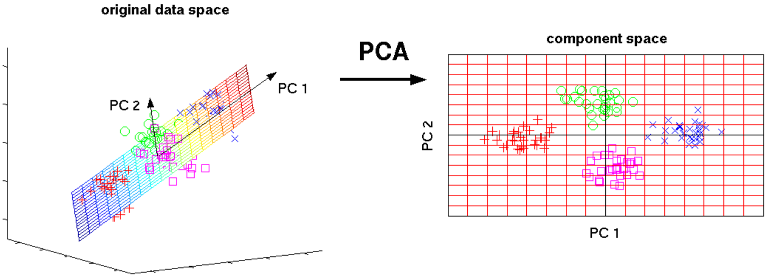
Dimensionality Reduction vs Data Compression

- While dimensionality reduction can be seen as a simplistic form of data compression, it is not equivalent to it, as the goal of data compression is to reduce the expected code length (which is lower bounded by **entropy**) of the representation not only the dimensionality
- For example, in lossless data compression, **arithmetic coding** encodes the entire data into a single number, an arbitrary-precision fraction q where $0.0 \leq q < 1.0$. In some science fiction books, this is paraphrased as a pinpoint representing all information of the whole universe

Outline

Linear Dimensionality Reduction

- Project the original data onto a lower-dimensional hyperplane (e.g., line, plane)
  - I.e., move and rotate the coordinate axis of the data
  - Represent the data with coordinates in the new component space



Linear Dimensionality Reduction

- Mathematically, this can be written as follows:

$$\mathbf{x}^{(i)} = \sum_{k=1}^{K} z_k^{(i)} \mathbf{b}_k$$

  - $\mathbf{b}_k = [b_{1k}, \ldots, b_{Nk}]^T$ is a basis vector
  - $z_k^{(i)} \in \mathbb{R}$ is the corresponding weight

Connection to Linear Regression

- Focus on the j-th entry of $\mathbf{x}^{(i)}$:

$$x_j^{(i)} = \sum_{k=1}^{K} z_k^{(i)} b_{jk}$$

  - This expression can be seen linear regression
    - $x_j^{(i)}$ is the target
    - $z_k^{(i)}$ for each k are the weights
    - b_{jk} for each k are the features
  - Alternatively, we may view $z_k^{(i)}$ as feature and b_{jk} as weight
  - Unlike linear regression, we only know "targets." We must learn both features and weights

Matrix Form

- Data matrix: $\mathbf{X} \in \mathbb{R}^{M\times N}$ with one data case $\mathbf{x}^{(i)} \in \mathbb{R}^N$ per row

$$\mathbf{X} = \begin{bmatrix} — & (\mathbf{x}^{(1)})^T & — \\ — & (\mathbf{x}^{(2)})^T & — \\ & \vdots & \\ — & (\mathbf{x}^{(M)})^T & — \end{bmatrix}$$

- Loading matrix $\mathbf{Z} \in \mathbb{R}^{M\times K}$ and factor matrix $\mathbf{B} \in \mathbb{R}^{K\times N}$

$$\mathbf{Z} = \begin{bmatrix} z_1^{(1)} & z_2^{(1)} & \cdots & z_K^{(1)} \\ z_1^{(2)} & z_2^{(2)} & \cdots & z_K^{(2)} \\ \vdots & \ddots & \ddots & \vdots \\ z_1^{(M)} & z_2^{(M)} & \cdots & z_K^{(M)} \end{bmatrix}, \mathbf{B} = \begin{bmatrix} — & (\mathbf{b}^{(1)})^T & — \\ — & (\mathbf{b}^{(2)})^T & — \\ & \vdots & \\ — & (\mathbf{b}^{(K)})^T & — \end{bmatrix}$$

Observation Noise

- We can express $\mathbf{X}$ as follows:

$$\mathbf{X} = \mathbf{Z} \times \mathbf{B}$$

- Most real world data will be subject to noise. If we assume that $\boldsymbol{\epsilon} \in \mathbb{R}^{M \times N}$ is a matrix of noise values from some probability distribution, we have

$$\mathbf{X} = \mathbf{Z} \times \mathbf{B} + \boldsymbol{\epsilon}$$

Learning Criterion

- The learning problem for linear dimensionality reduction is to estimate values for both $\mathbf{Z}$ and $\mathbf{B}$ given only the noisy observations $\mathbf{X}$
- One possible learning criterion is to minimize the sum of squared errors when **reconstructing** $\mathbf{X}$ from $\mathbf{Z}$ and $\mathbf{B}$. This leads to:

$$\arg\min_{\mathbf{Z},\mathbf{B}} \|\mathbf{X} - \mathbf{Z}\mathbf{B}\|_F^2$$

  - $\|\mathbf{A}\|_F$ is the Frobenius norm of matrix $\mathbf{A}$
    - $\|\mathbf{A}\|_F = \sqrt{\sum_{ij} A_{ij}^2}$

Alternating Least Squares

- By leveraging the OLS solution to linear regression, we can estimate $\mathbf{Z}$ and $\mathbf{B}$ using Alternating Least Squares (ALS)
- Starting from some random initialization, ALS iterates between two steps until convergence:
  - Assume $\mathbf{Z}$ is given and optimize $\mathbf{B}$:

$$\mathbf{B} \leftarrow (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{X}$$

  - Assume $\mathbf{B}$ is given and optimize $\mathbf{Z}$:

$$\mathbf{Z}^T \leftarrow (\mathbf{B}\mathbf{B}^T)^{-1}\mathbf{B}\mathbf{X}^T$$

Lack of Uniqueness for Optimal Parameters

- Suppose we run the ALS algorithm to convergence and obtain optimal parameters $\mathbf{Z}^\star$ and $\mathbf{B}^\star$ such that:

$$\ell^\star = \|\mathbf{X} - \mathbf{Z}^\star\mathbf{B}^\star\|_F^2$$

- Assume an invertible matrix $\mathbf{R} \in \mathbb{R}^{K\times K}$
  - A $K \times K$ matrix $\mathbf{R}$ is invertible, if there exists a $K \times K$ square matrix $\mathbf{S}$ such that $\mathbf{RS} = \mathbf{SR} = \mathbf{I}$. $\mathbf{S}$ is often denoted by $\mathbf{R}^{-1}$
- We obtain a different set of parameters $\tilde{\mathbf{Z}} = \mathbf{Z}^\star\mathbf{R}$ and $\tilde{\mathbf{B}} = \mathbf{R}^{-1}\mathbf{B}^\star$ with the same optimal value:

$$\ell^\star = \|\mathbf{X} - \mathbf{Z}^\star(\mathbf{I})\mathbf{B}^\star\|_F^2 = \|\mathbf{X} - \mathbf{Z}^\star(\mathbf{R}\mathbf{R}^{-1})\mathbf{B}^\star\|_F^2 = \|\mathbf{X} - \tilde{\mathbf{Z}}\tilde{\mathbf{B}}\|_F^2$$

- We can obtain the **global** optimal solutions and make them **unique** by specifying additional criteria

Outline

Singular Value Decomposition (SVD)

- Let $\mathbf{X}$ be an $M \times N$ matrix, with $M \geq N$. It can be factorized

$$\mathbf{X} = \mathbf{U} \begin{pmatrix} \mathbf{\Sigma} \\ 0 \end{pmatrix} \mathbf{V}^T$$

  - $\mathbf{U} \in \mathbb{R}^{M \times M}$ and $\mathbf{V} \in \mathbb{R}^{N \times N}$ are orthogonal, *i.e.*,

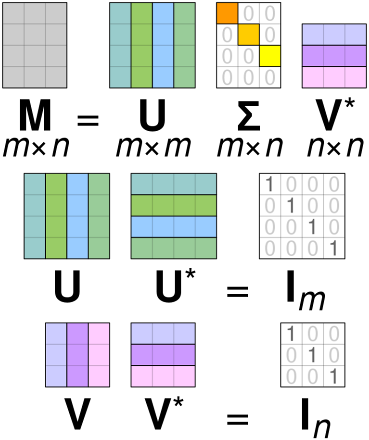
$$\mathbf{U}^T\mathbf{U} = \mathbf{U}\mathbf{U}^T = \mathbf{I}_M, \quad \mathbf{V}^T\mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbf{I}_N$$

    - Columns of $\mathbf{U}$ and $\mathbf{V}$ are called the left and right **singular vectors** of $\mathbf{X}$, respectively
  - $\mathbf{\Sigma} \in \mathbb{R}^{N \times N}$ is diagonal

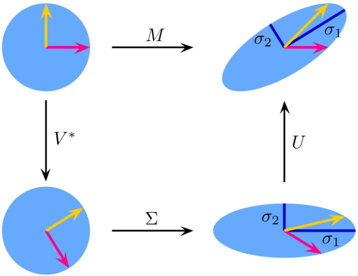
$$\mathbf{\Sigma} = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_N), \sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_N \geq 0$$

    - σ_i's are called **singular values** of $\mathbf{X}$

Singular Value Decomposition (SVD)



Singular Value Decomposition (SVD)



$$M = U \cdot \Sigma \cdot V^*$$

- Upper left: the unit disc with the two canonical unit vectors
- Upper right: transformed with $\mathbf{M}$
- Lower left: the action of $\mathbf{V}^T$. This is just a rotation
- Lower right: the action of $\mathbf{\Sigma V}^T$. $\mathbf{\Sigma}$ scales vertically and horizontally

Reduced-Form SVD

- If only $K < \min\{M, N\}$ singular values are non-zeros, the SVD of $\mathbf{X} \in \mathbb{R}^{M \times N}$ can be represented in reduced form as follows

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}_K\mathbf{V}^T = \sum_{k=1}^{K} \sigma_k \mathbf{u}_k \mathbf{v}_k^T$$

  - $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_K] \in \mathbb{R}^{M \times K}$
    - $\mathbf{U}^T\mathbf{U} = \mathbf{I}_K$
  - $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_K] \in \mathbb{R}^{N \times K}$
    - $\mathbf{V}^T\mathbf{V} = \mathbf{I}_K$
  - $\mathbf{\Sigma}_K = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_K)$
    - $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_K > 0$
  - $\mathbf{u}_k\mathbf{v}_k^T \in \mathbb{R}^{M \times N}$ is the product of a column vector $\mathbf{u}_k$ and a row vector $\mathbf{v}_k^T$
    - It has **rank** 1
    - $\mathbf{X}$ is a weighted summation of K rank-1 matrices

Eckart-Young-Mirsky Theorem

- Given an $M \times N$ matrix $\mathbf{X}$ of rank $R \leq \min\{M, N\}$ and its singular value decomposition $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}_R\mathbf{V}^T$, with singular values $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_R > 0$ and $\sigma_{R+1} = \sigma_{R+2} = \ldots = \sigma_{\min\{M,N\}} = 0$, then among all $M \times N$ matrices of lower rank $K \leq R$, the best approximation is $\mathbf{Y}^\star = \mathbf{U}\mathbf{\Sigma}_K\mathbf{V}^T$, where $\mathbf{\Sigma}_K$ is the diagonal matrix with singular values $\sigma_1, \sigma_2, \ldots, \sigma_K$ in the sense that

$$\|\mathbf{X} - \mathbf{Y}^\star\|_F^2 = \min\{\|\mathbf{X} - \mathbf{Y}\|_F^2; \mathbf{Y} \in \mathbb{R}^{M \times N}, \mathrm{rank}\mathbf{Y} \leq K\}$$

- SVD provides a unique solution to minimum Frobenius norm linear dimensionality reduction