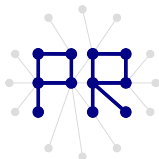# Pattern Recognition Lecture
"Feature Generation"

Prof. Dr. Marcin Grzegorzek

Research Group for Pattern Recognition
www.pr.informatik.uni-siegen.de

Institute for Vision and Graphics
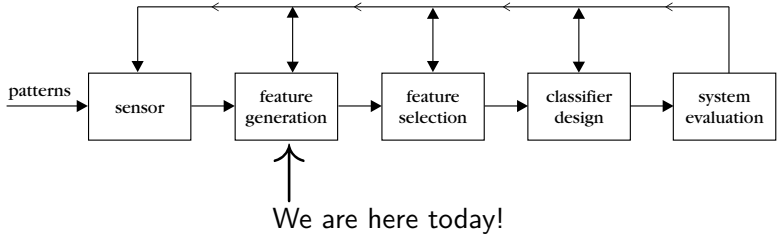University of Siegen, Germany

Pattern Recognition Chain

We are here today!

Overview

1 Introduction

2 Basis Vectors and Images

3 The Karhunen-Loeve Transform

4 The Singular Value Decomposition

Overview

Introduction (1)

- Given a set of measurements, the goal is to discover compact and informative representations of the obtained data - features.

- The representations are generated after processing a large amount of sensory data.

- Measurements are transformed to a new set of features.

- In good features, the classification-related information is "squeezed" in a relatively small number of features.

- Appropriately chosen feature transform can exploit and remove information redundancies.

- E. g., using image pixels as features would be highly inefficient as pixels have a large degree of correlation.

- However, the Fourier transform turns out to be much more efficient for feature extraction. Why?

- Fourier transform is just one of the tools from a palette of possible transforms.

Overview

Basis Vectors and Images (1)

- Let $x(0), x(1), \ldots, x(N-1)$ be a set of input samples and $\mathbf{x}$ be the $N \times 1$ corresponding vector

$$\mathbf{x}^{\mathrm{T}} = [x(0), x(1), \ldots, x(N-1)]$$

- Given a unitary[1] $N \times N$ matrix $\mathbf{A}$, we define the transformed vector $\mathbf{y}$ of $\mathbf{x}$ as

$$\mathbf{y} = \mathbf{A}^H \mathbf{x} \equiv \begin{bmatrix} \mathbf{a}_0^H \\ \vdots \\ \mathbf{a}_{N-1}^H \end{bmatrix} \mathbf{x}$$

- H denotes the Hermitian operation (complex conjugation and transposition).

[1] A complex matrix is called unitary if $A^{-1} = A^H$.

Basis Vectors and Images (2)

- From the definition of unitary matrices we have

$$\mathbf{x} = \mathbf{A}\mathbf{y} = \sum_{i=0}^{N-1} y(i)\mathbf{a}_i$$

- The columns of $\mathbf{A}$, namely $\mathbf{a}_{i=0,1,\ldots,N-1}$ are called the basis vectors of the transformation.

- The elements $y(i)$ of $\mathbf{y}$ are projections of $\mathbf{x}$ onto these basis vectors.

Basis Vectors and Images (3)

- In many problems, such as image analysis, the input set is an $N \times N$ matrix $\mathbf{X}$ with elements $X(i,j)$ instead of a vector.

- In such cases we define an equivalent N^2 vector $\mathbf{x}$ by ordering the rows in the matrix one after another

  $$\mathbf{x}^{\mathrm{T}} = [X(0,0), \ldots, X(0, N-1), \ldots X(N-1, 0), \ldots, X(N-1, N-1)]$$

  and now we can transform this vector.

- However, this is not the most efficient way as the number of operations to multiply an $N^2 \times N^2$ matrix with an $N^2 \times 1$ vector $\mathbf{x}$ is of the order $O(N^4)$.

Basis Vectors and Images (4)

- An alternative possibility is to transform matrix $\mathbf{X}$ via a set of basis matrices or basis images.

- Let $\mathbf{U}$ and $\mathbf{V}$ be unitary $N \times N$ matrices. Define the transformed matrix $\mathbf{Y}$ of $\mathbf{X}$ as

$$\mathbf{Y} = \mathbf{U}^H \mathbf{X} \mathbf{V} \qquad \Rightarrow \qquad \mathbf{X} = \mathbf{U} \mathbf{Y} \mathbf{V}^H$$

- The number of operations is now reduced to $O(N^3)$.

Basis Vectors and Images (5)

- It can be shown that

$$\mathbf{X} = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} Y(i,j)\mathbf{u}_i\mathbf{v}_j^H$$

  where $\mathbf{u}_i$ are column vectors of $\mathbf{U}$ and $\mathbf{v}_j$ of $\mathbf{V}$.

- Each of the outer products $\mathbf{u}_i\mathbf{v}_j^H$ is an $N \times N$ matrix[2]

$$\mathbf{u}_i\mathbf{v}_j^H = \left[\begin{array}{ccc} u_{i0}v_{j0}^* & \cdots & u_{i0}v_{jN-1}^* \\ \vdots & \vdots & \vdots \\ u_{iN-1}v_{j0}^* & \cdots & u_{iN-1}v_{jN-1}^* \end{array} \right] \equiv \mathbf{A}_{ij}$$

- A_{ij} are called basis images (matrices).

---

[2] $*$ denotes complex conjugation.

- If **Y** turns out to be diagonal, then

$$\mathbf{X} = \sum_{i=0}^{N-1} Y(i,i)\mathbf{u}_i \mathbf{v}_i^H$$

- The number of basis images is reduced to N in this case.

- If we define the inner product between two matrices as

$$\langle \mathbf{A}, \mathbf{B} \rangle \equiv \sum_{m=0}^{N-1} \sum_{n=0}^{N-1} A^*(m,n) B(m,n)$$

- Then it can be shown that

$$Y(i,j) = \langle \mathbf{A}_{ij}, \mathbf{X} \rangle$$

Overview

- The Karhunen-Loeve transform known also as Principal Component Analysis (PCA) is one of the most popular methods for feature generation and dimensionality reduction in pattern recognition.

- The computation of the transformation matrix exploits the statistical information describing the data.

- The labels of the training samples are not used (unsupervised mode).

- Let $\mathbf{x}$ be vectors representing samples. In order to simplify the presentation, the data samples are assumed to have zero mean. If this is not the case, we can always subtract the mean value.

- Let $\mathbf{y} = \mathbf{A}^{\mathrm{T}}\mathbf{x}$

- From the definition of the correlation matrix we have

$$\mathbf{R}_y \equiv E[\mathbf{y}\mathbf{y}^{\mathrm{T}}] = E[\mathbf{A}^{\mathrm{T}}\mathbf{x}\mathbf{x}^{\mathrm{T}}\mathbf{A}] = \mathbf{A}^{\mathrm{T}}\mathbf{R}_x\mathbf{A}$$

- In practice, $\mathbf{R}_x$ is estimated as an average over the given set of training vectors

$$\mathbf{R}_x = \frac{1}{n} \sum_{k=1}^{n} \mathbf{x}_k \mathbf{x}_k^{\mathrm{T}}$$

- Note that $\mathbf{R}_x$ is a symmetric matrix and hence its eigenvectors are mutually orthogonal.

- Thus, if matrix **A** is chosen so that its columns are the orthonormal eigenvectors $\mathbf{a}_{i=0,\dots,N-1}$, then

$$\mathbf{R}_y = \mathbf{A}^{\mathrm{T}}\mathbf{R}_x\mathbf{A} = \mathbf{\Lambda}$$

where $\mathbf{\Lambda}$ is a diagonal matrix having as elements the respective eigenvalues $\lambda_{i=0,\dots,N-1}$ of $\mathbf{R}_x$.

- The resulting transform is known as the Karhunen-Loeve (KL) transform.

- The KL transform generates mutually uncorrelated features.

Example (1)

- In the figure below, we see 100 points in the two-dimensional space spread around the $x_1 = x_2$ line generated by the model $x_2 = x_1 + \epsilon$, where ϵ is a noise source following the uniform distribution in $[-0.5, 0.5]$.
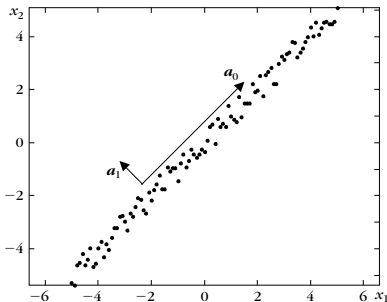
Example (2)

- We first compute the covariance matrix and perform an eigen decomposition. The resulting eigenvectors are

  $$\mathbf{a}_0 = [0.7045, 07097]^{\mathrm{T}} \quad \text{and} \quad \mathbf{a}_1 = [-0.7097, 0.7045]^{\mathrm{T}}$$

- The corresponding eigenvalues are

  $$\lambda_0 = 17.26 \quad \text{and} \quad \lambda_1 = 0.04 \quad \text{thus} \quad \lambda_0 \gg \lambda_1$$

- Projecting along the direction of maximum variability ($\mathbf{a}_0$) retains most of the variance.

- The number of dimensions of the feature space can be reduced to 1.

Overview

SVD Statement (1)

- Given a $l \times n$ matrix $\mathbf{X}$ of rank r ($r \leq \min\{l, n\}$) there exist unitary matrices $\mathbf{U}$ and $\mathbf{V}$ of dimensions $l \times l$ and $n \times n$, respectively, so that

$$\mathbf{X} = \mathbf{U} \left[\begin{array}{cc} \mathbf{\Lambda}^{\frac{1}{2}} & O \\ O & 0 \end{array} \right] \mathbf{V}^H \Rightarrow \mathbf{Y} \equiv \left[\begin{array}{cc} \mathbf{\Lambda}^{\frac{1}{2}} & O \\ O & 0 \end{array} \right] = \mathbf{U}^H \mathbf{X} \mathbf{V}$$

- $\mathbf{\Lambda}^{\frac{1}{2}}$ is the $r \times r$ diagonal matrix with elements $\sqrt{\lambda_i}$ and λ_i are the r nonzero eigenvalues of the associated matrix $\mathbf{X}^H \mathbf{X}$. O denotes a zero element matrix.

- In other words, there exist unitary matrices $\mathbf{U}$ and $\mathbf{V}$ that transform $\mathbf{X}$ into the special diagonal structure of $\mathbf{Y}$.

SVD Statement (2)

- If $\mathbf{u}_i$ and $\mathbf{v}_i$ denote the column vectors of matrices $\mathbf{U}$ and $\mathbf{V}$, respectively, then the SVD can be written as

$$\mathbf{X} = [\mathbf{u}_0, \mathbf{u}_1, \ldots, \mathbf{u}_{r-1}] \begin{bmatrix} \sqrt{\lambda_0} & 0 & \ldots & 0 \\ 0 & \sqrt{\lambda_1} & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & \sqrt{\lambda_{r-1}} \end{bmatrix} \begin{bmatrix} \mathbf{v}_0^H \\ \mathbf{v}_1^H \\ \vdots \\ \mathbf{v}_{r-1}^H \end{bmatrix}$$

$$\Downarrow$$

$$\mathbf{X} = \sum_{i=0}^{r-1} \sqrt{\lambda_i} \mathbf{u}_i \mathbf{v}_i^H$$

SVD Statement (3)

- Sometimes, the above is also written as

$$\mathbf{X} \equiv \mathbf{U}_r \mathbf{\Lambda}^{\frac{1}{2}} \mathbf{V}_r^H$$

- $\mathbf{U}_r$ denotes the $l \times r$ matrix that consists of the first r columns of $\mathbf{U}$ and $\mathbf{V}_r$ the $r \times n$ matrix formed by using the first r columns of $\mathbf{V}$.

- More precisely, $\mathbf{u}_i$ and $\mathbf{v}_i$ are the eigenvectors corresponding to the nonzero eigenvalues of the matrices $\mathbf{X}\mathbf{X}^H$ and $\mathbf{X}^H\mathbf{X}$, respectively.

- The eigenvalues λ_i are known as singular values of $\mathbf{X}$ and the expansion as the singular value decomposition of $\mathbf{X}$ or the spectral representation of $\mathbf{X}$.

Low Rank Approximation via SVD (1)

- The SVD proposes an exact representation of matrix $\mathbf{X}$

$$\mathbf{X} = \sum_{i=0}^{r-1} \sqrt{\lambda_i} \mathbf{u}_i \mathbf{v}_i^H$$

- A very interesting implication occurs, if one uses less than r (the rank of $\mathbf{X}$) terms in the summation.

- Let $\mathbf{X}$ be approximated by

$$\mathbf{X} \approx \widehat{\mathbf{X}} = \sum_{i=0}^{k-1} \sqrt{\lambda_i} \mathbf{u}_i \mathbf{v}_i^H, \qquad k \leq r$$

- Matrix $\widehat{\mathbf{X}}$ being a sum of $k \leq r$ rank-one independent $l \times n$ matrices is of rank k.

- If the k largest eigenvalues are involved, it can be shown that the squared error

$$\epsilon^2 = \sum_{i=0}^{l-1} \sum_{j=0}^{n-1} |X(i,j) - \widehat{X}(i,j)|^2$$

  is the minimum one with respect to all $l \times n$ matrices of rank k.

- The square root of the right-hand side is also known as the Frobenius norm $\|\mathbf{X} - \widehat{\mathbf{X}}\|_{\mathrm{F}}$ of the difference $\mathbf{X} - \widehat{\mathbf{X}}$.

- The error in the approximation turns out to be

$$\epsilon^2 = \sum_{i=k}^{r-1} \lambda_i$$

- Hence, if we order the eigenvalues in descending order $\lambda_0 \geq \lambda_1 \geq \ldots \geq \lambda_{r-1}$, then for a given number of k terms in the expansion, the SVD leads to the minimum square error.

- Thus, $\widehat{\mathbf{X}}$ is the best rank-k approximation of $\mathbf{X}$ with respect to the Frobenius norm.

KL Transform vs. SVD

- The statement of the previous slide reminds us of the Karhunen-Loeve expansion.

- However, while KL is optimal with respect to the mean square error, SVD is optimal in accordance with the Frobenius norm.
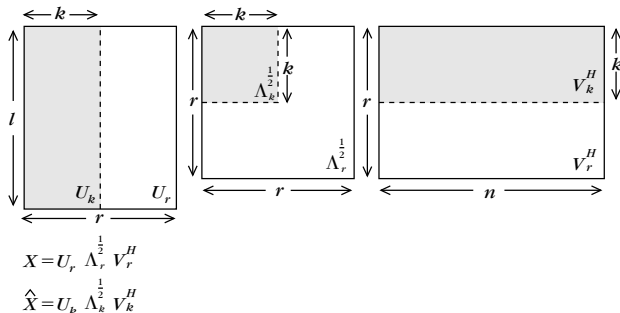
Dimensionality Reduction via SVD (1)

- SVD is used for dimensionality reduction in pattern recognition and information retrieval.

- Using the equations before, we can see that

$$\mathbf{X} \approx \widehat{\mathbf{X}} = [\mathbf{u}_0, \dots, \mathbf{u}_{k-1}] \left[\begin{array}{c} \mathbf{v}_0^H \sqrt{\lambda_o} \\ \vdots \\ \mathbf{v}_{k-1}^H \sqrt{\lambda_{k-1}} \end{array} \right] = \mathbf{U}_k [\mathbf{a}_0, \dots, \mathbf{a}_{n-1}]$$

where $\mathbf{U}_k$ consists of the first k columns of $\mathbf{U}$ and the k-dimensional vectors $\mathbf{a}_{i=0,\dots,n-1}$ are the column vectors of the $k \times n$ product matrix $\mathbf{\Lambda}_k^{\frac{1}{2}} \mathbf{V}_k^H$.

Dimensionality Reduction via SVD (2)

- The figure below gives a diagrammatic interpretation of the matrix products involved in the SVD

$$X = U_r \ \Lambda_r^{\frac{1}{2}} \ V_r^H$$

$$\widehat{X} = U_k \ \Lambda_k^{\frac{1}{2}} \ V_k^H$$

- In the approximation of **X** by $\widehat{\mathbf{X}}$, the first k columns of $\mathbf{U}_r$ and the first k rows of $\mathbf{V}_r^H$ are involved.

- Each *l*-dimensional column vector $\mathbf{x}_i$ of $\mathbf{X}$ is approximated by the *k*-dimensional vector $\mathbf{a}_i$ lying in the subspace spanned by $\mathbf{u}_{i=0,\dots,k-1}$ ($\mathbf{a}_i$ is the projection of $\mathbf{x}_i$ on this subspace)

$$\mathbf{x}_i \approx \mathbf{U}_k \mathbf{a}_i = \sum_{m=0}^{k-1} \mathbf{u}_m a_i(m), \quad i = 0, \dots, n-1$$

- Due to the orthonormality of the columns $\mathbf{u}_{i=0,ldots,k-1}$ can be seen that

$$\|\mathbf{x}_i - \mathbf{x}_j\| \approx \|\mathbf{U}_k(\mathbf{a}_i - \mathbf{a}_j)\| = \|\sum_{m=0}^{k-1} \mathbf{u}_m(a_i(m) - a_j(m))\| =$$

$$= \|\mathbf{a}_i - \mathbf{a}_j\|, \quad i = 0, \dots, n-1$$

SVD Conclusions

- The SVD has excellent information packing properties.
  E, g., an image array can be represented efficiently by a few
  of its singular values. Thus, SVD is a natural candidate as
  a tool for feature generation/selection in classification.

- Since SVD approximately keeps the Euclidean distance
  (see previous slide), it is suitable for information retrieval
  tasks.