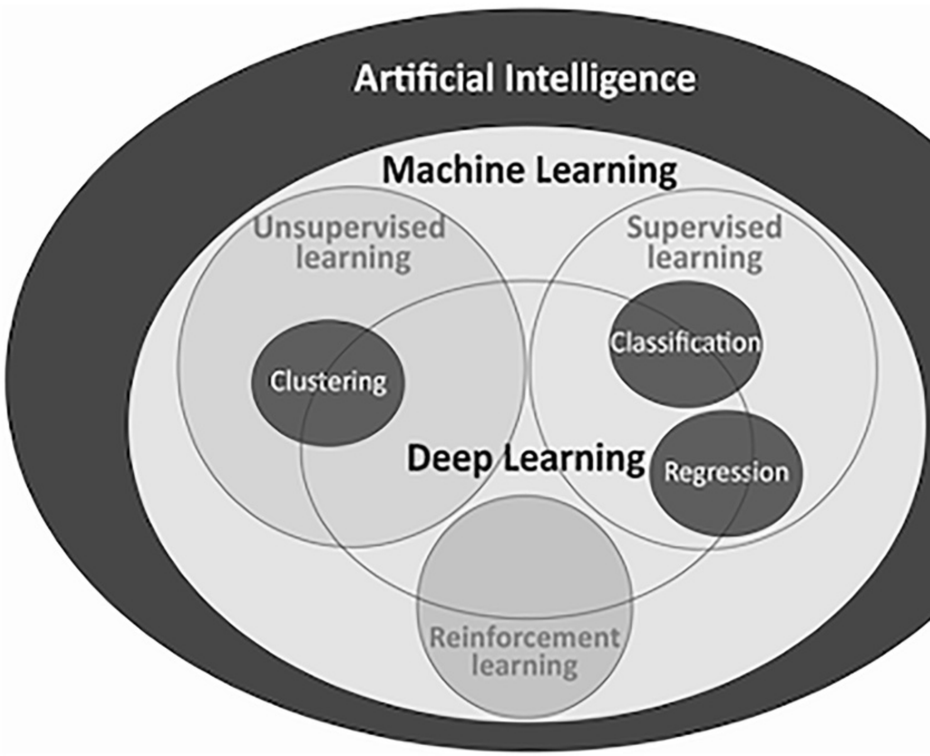
COSC 522 – Machine Learning

Dimensionality Reduction

Hairong Qi, Gonzalez Family Professor
Electrical Engineering and Computer Science
University of Tennessee, Knoxville
https://www.eecs.utk.edu/people/hairong-qi/
Email: hqi@utk.edu



	Part 1: Statistical Methods	
	Baysian Learning	
08/20 (T)		Introduction
08/22 (R)		Baysian Decision Theory and Parametric Learning
08/27 (T)		Baysian Decision Theory and Non-Parametric Learning
08/29 (R)		Case Study: Representation for Natural Language (taught by Andr Cozma)
09/03 (T)		Parametric vs. Non-Parametric Learning: Some In-Depth Discussi
09/05 (R)		Homework and Project Discussion (taught by Fanqi Wang)
	Neural Networks	
09/10 (T)		Biological Neuron and Perceptron
09/12 (R)		Perceptron
09/17 (T)		Back Propagation and Gradient Descent
09/19 (R)		Back Propagation
09/20 (F)		TRUST-AI Seminar
09/24 (T)		Kernel Methods and Review
09/26 (R)	Test 1	
10/01 (T)		Kernel Methods and Support Vector Machine
	Regression	
10/03 (R)		Regression
~~**10/08 (T)**~~	~~Fall Break (No Class)~~	
	Unsupervised Learning	
10/10 (R)		Logistic Regression; k-means
10/15 (T)		Hierarchical Clustering
	Dimensionality Reduction	
10/17 (R)		Supervised methods
10/22 (T)		Unsupervised methods

Questions

AICIP RESEARCH

- What is the curse of dimensionality?

- What are the different objectives of the two dimensionality reduction approaches?
- What is the cost function for FLD? Can you verbally describe it in one sentence? What is the optimization approach taken?
- What is scatter matrix? What are between-class scatter and within-class scatter?
- Is FLD supervised or unsupervised?

- What is the cost function for PCA? Can you verbally describe it in one sentence? What is the optimization approach taken?
- What is major principal axis?
- Is PCA supervised or unsupervised?

The Curse of Dimensionality – 1st Aspect

- The number of training samples
- What would the probability density function look like if the dimensionality is very high?
  - For a 7-dimensional space, where each variable could have 20 possible values, then the 7-d histogram contains 20^7 cells. To distribute a training set of some reasonable size (1000) among this many cells is to leave virtually all the cells empty

Curse of Dimensionality – 2nd Aspect

- Accuracy and overfitting
- In theory, the higher the dimensionality, the less the error, the better the performance. However, in realistic ML problems, the opposite is often true. Why?
  - The assumption that pdf behaves like Gaussian is only approximately true
  - When increasing the dimensionality, we may be overfitting the training set.
  - Problem: excellent performance on the training set, poor performance on new data points which are in fact very close to the data within the training set

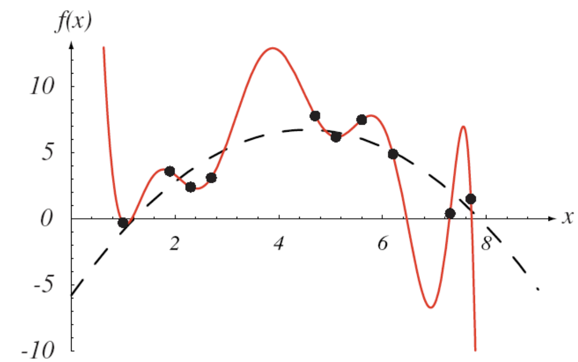


FIGURE 3.4. The "training data" (black dots) were selected from a quadratic function plus Gaussian noise, i.e., $f(x) = ax^2 + bx + c + \epsilon$ where $p(\epsilon) \sim N(0, \sigma^2)$. The 10th-degree polynomial shown fits the data perfectly, but we desire instead the second-order function $f(x)$, because it would lead to better predictions for new samples. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Curse of Dimensionality - 3rd Aspect

- Computational complexity

PART I: SUPERVISED DR - FLD

Dimensionality Reduction

AICIP RESEARCH

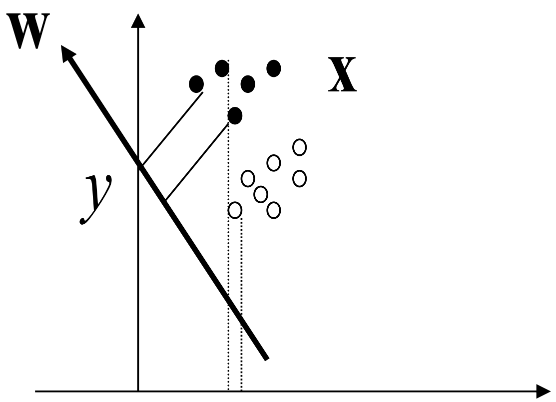
- Linear
  - Fisher's linear discriminant (Linear Discriminant Analysis – LDA)
    - Best discriminating the data
    - Supervised
  - Principal component analysis (PCA)
    - Best representing the data
    - Unsupervised

Fisher's Linear Discriminant

AICIP RESEARCH

- For two-class cases, projection of data from d-dimension onto a line
- Principle: We'd like to find vector **w** (direction of the line) such that the projected data set can be best separated

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



$$J(\mathbf{w}) = \left|\tilde{m}_1 - \tilde{m}_2\right|^2 = \left|\mathbf{w}^T(\mathbf{m}_1 - \mathbf{m}_2)\right|^2$$

$$\tilde{m}_i = \frac{1}{n_i}\sum_{y \in Y_i} y = \frac{1}{n_i}\sum_{\mathbf{x} \in D_i} \mathbf{w}^T \mathbf{x} = \mathbf{w}^T \mathbf{m}_i \qquad \mathbf{m}_i = \frac{1}{n_i}\sum_{\mathbf{x} \in D_i} \mathbf{x}$$

Projected mean Sample mean

Other Approaches?

AICIP RESEARCH

- Solution 1: make the projected mean as apart as possible
- Solution 2?

$$J(\mathbf{w}) = \frac{|\tilde{m}_1 - \tilde{m}_2|^2}{\tilde{s}_1^2 + \tilde{s}_2^2} = \frac{|\mathbf{w}^T(\mathbf{m}_1 - \mathbf{m}_2)|^2}{\mathbf{w}^T\mathbf{S}_1\mathbf{w} + \mathbf{w}^T\mathbf{S}_2\mathbf{w}} = \boxed{\frac{\mathbf{w}^T\mathbf{S}_B\mathbf{w}}{\mathbf{w}^T\mathbf{S}_W\mathbf{w}}}$$

$$\tilde{s}_i^2 = \sum_{y \in Y_i}(y - \tilde{m}_i)^2 = \sum_{\mathbf{x} \in D_i}(\mathbf{w}^T\mathbf{x} - \mathbf{w}^T\mathbf{m}_i)^2 = \sum_{\mathbf{x} \in D_i}\mathbf{w}^T(\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T\mathbf{w} = \mathbf{w}^T\mathbf{S}_i\mathbf{w}$$

<u>Scatter matrix</u> $\mathbf{S}_i = \sum_{\mathbf{x} \in D_i}(\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T$

Between-class scatter matrix $\mathbf{S}_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T$

Within-class scatter matrix $\mathbf{S}_W = \mathbf{S}_1 + \mathbf{S}_2 = \sum_{i=1}^{2}(\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T$

*The Generalized Rayleigh Quotient

AICIP RESEARCH

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}}$$

$$\frac{dJ(w)}{dw} = \frac{2\mathbf{S}_B \mathbf{w}\left(\mathbf{w}^T \mathbf{S}_W \mathbf{w}\right) - 2\mathbf{S}_W \mathbf{w}\left(\mathbf{w}^T \mathbf{S}_B \mathbf{w}\right)}{\left(\mathbf{w}^T \mathbf{S}_W \mathbf{w}\right)^2} = 0$$

$$\mathbf{S}_B \mathbf{w} = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}} \mathbf{S}_W \mathbf{w} \Rightarrow \mathbf{S}_W^{-1} \mathbf{S}_B \mathbf{w} = \lambda \mathbf{w}$$

$\mathbf{S}_B \mathbf{w}$ is always in the direction of $\mathbf{m}_1 - \mathbf{m}_2$

$$\mathbf{w} = \mathbf{S}_W^{-1}\left(\mathbf{m}_1 - \mathbf{m}_2\right)$$ Canonical variate

Some Math Preliminaries

AICIP RESEARCH

- Positive definite
  - A matrix **S** is positive definite if $y=\mathbf{x}^T\mathbf{S}\mathbf{x}>0$ for all R^d except 0
  - $\mathbf{x}^T\mathbf{S}\mathbf{x}$ is called the quadratic form
  - The derivative of a quadratic form is particularly useful

$$\frac{d}{d\mathbf{x}}\left(\mathbf{x}^T\mathbf{S}\mathbf{x}\right)=\left(\mathbf{S}+\mathbf{S}^T\right)\mathbf{x}$$

- Eigenvalue and eigenvector
  - **x** is called the eigenvector of **A** iff **x** is not zero, and $\mathbf{A}\mathbf{x}=\lambda\mathbf{x}$
  - λ is the eigenvalue of **x**

Multiple Discriminant Analysis

- For c-class problem, the projection is from d-dimensional space to a (c-1)-dimensional space (assume d >= c)
- Between-class scatter matrix: $S_B = \sum_{k=1}^{c} n_k(\mathbf{m}_k - \mathbf{m})(\mathbf{m}_k - \mathbf{m})^{\mathrm{T}}$ where **m** is the global mean, $\mathbf{m}_k$ is the class mean, and n_k is the number of samples in class *k*, *c* is the total number of classes.
- Within-class scatter matrix: $S_W = \sum_{k=1}^{c} S_k$, $S_k = \sum_{i \in D_k}(\mathbf{x}_i - \mathbf{m}_k)(\mathbf{x}_i - \mathbf{m}_k)^{\mathrm{T}}$
- $J(W) = Tr(\frac{W^T S_B W}{W^T S_W W})$: trace is the sum of elements along the main diagonal direction. Can only calculate trace for a square matrix.
- $W = eig(S_w^{-1} S_B)$: At most c-1 non-zero eigenvalues as S_B has a rank of c-1.

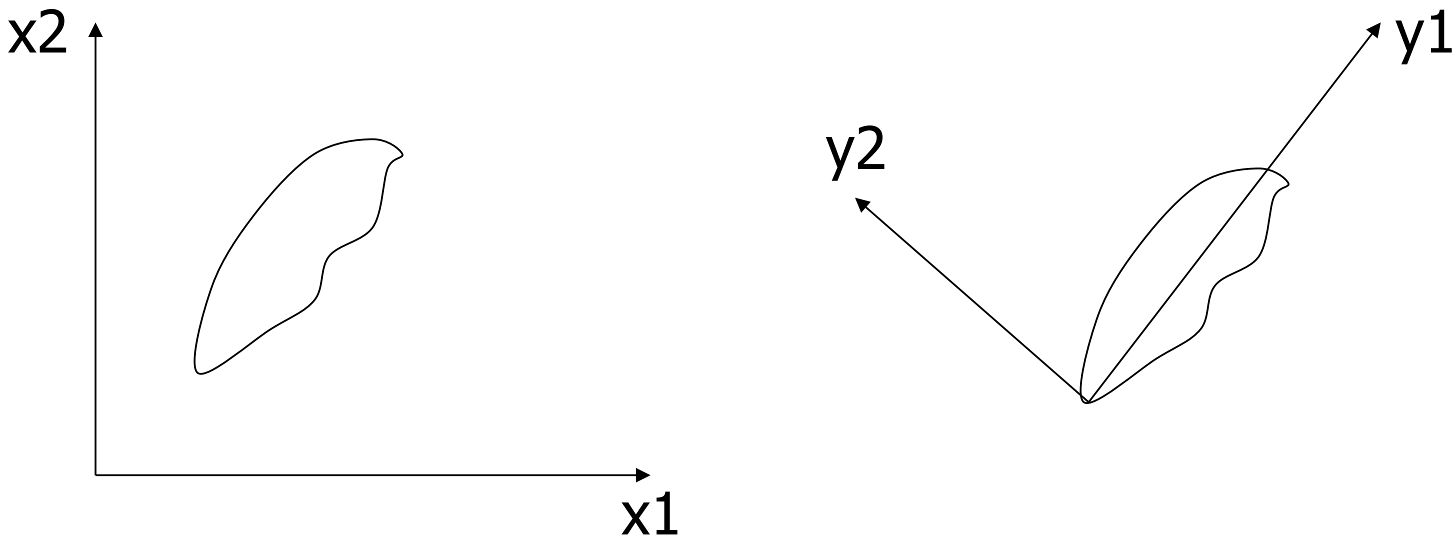
PART II: UNSUPERVISED DR - PCA

PCA Procedure

- Raw data → covariance matrix → eigenvalue → eigenvector → principal component
- How to use error rate?

Principal Component Analysis or K-L Transform

- How to find a new feature space (m-dimensional) that is adequate to describe the original feature space (d-dimensional). Suppose m<d



K-L Transform (1)

AICIP RESEARCH

- Describe vector **x** in terms of a set of basis vectors $\mathbf{b}_i$.

$$\mathbf{x} = \sum_{i=1}^{d} y_i \mathbf{b}_i \qquad\qquad y_i = \mathbf{b}_i^T \mathbf{x}$$

- The basis vectors ($\mathbf{b}_i$) should be linearly independent and orthonormal, that is,

$$\mathbf{b}_i^T \mathbf{b}_j = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

K-L Transform (2)

AICIP RESEARCH

- Suppose we wish to ignore all but *m* (*m*<*d*) components of **y** and still represent **x**, although with some error. We will thus calculate the first *m* elements of **y** and replace the others with constants

$$\mathbf{x} = \sum_{i=1}^{m} y_i \mathbf{b}_i + \sum_{i=m+1}^{d} \boxed{y_i} \mathbf{b}_i \approx \sum_{i=1}^{m} y_i \mathbf{b}_i + \sum_{i=m+1}^{d} \boxed{\alpha_i} \mathbf{b}_i$$

Error: $$\Delta \mathbf{x} = \sum_{i=m+1}^{d} \left(y_i - \alpha_i \right) \mathbf{b}_i$$

K-L Transform (3)

AICIP RESEARCH

- Use mean-square error to quantify the error

$$\varepsilon^2(m) = E\left\{ \sum_{i=m+1}^{d} \sum_{j=m+1}^{d} (y_i - \alpha_i)\mathbf{b}_i^T (y_j - \alpha_j)\mathbf{b}_j \right\}$$

$$= E\left\{ \sum_{i=m+1}^{d} \sum_{j=m+1}^{d} (y_i - \alpha_i)(y_j - \alpha_j)\mathbf{b}_i^T \mathbf{b}_j \right\}$$

$$= \sum_{i=m+1}^{d} E\left\{ (y_i - \alpha_i)^2 \right\}$$

K-L Transform (4)

AICIP RESEARCH

- Find the optimal α_i to minimize ε^2

$$\frac{\partial \varepsilon^2}{\partial \alpha_i} = -2\left(E\{y_i\} - \alpha_i\right) = 0$$

$$\alpha_i = E\{y_i\}$$

- Therefore, the error is now equal to

$$\varepsilon^2(m) = \sum_{i=m+1}^{d} E\left\{\left(y_i - E\{y_i\}\right)^2\right\}$$

$$= \sum_{i=m+1}^{d} E\left\{\left(\mathbf{b}_i^T \mathbf{x} - E\{\mathbf{b}_i^T \mathbf{x}\}\right)^2\right\} = \sum_{i=m+1}^{d} E\left\{\left(\mathbf{b}_i^T \mathbf{x} - E\{\mathbf{b}_i^T \mathbf{x}\}\right)\left(\mathbf{x}^T \mathbf{b}_i - E\{\mathbf{x}^T \mathbf{b}_i\}\right)\right\}$$

$$= \sum_{i=m+1}^{d} \mathbf{b}_i^T E\left\{\left(\mathbf{x} - E\{\mathbf{x}\}\right)\left(\mathbf{x} - E\{\mathbf{x}\}\right)^T\right\} \mathbf{b}_i = \sum_{i=m+1}^{d} \mathbf{b}_i^T \Sigma_{\mathbf{x}} \mathbf{b}_i = \boxed{\sum_{i=m+1}^{d} \lambda_i}$$

K-L Transform (5)

AICIP RESEARCH

- The optimal choice of basis vectors is the eigenvectors of $\Sigma_{\mathbf{x}}$
- The expansion of a random vector in terms of the eigenvectors of the covariance matrix is referred to as the Karhunen-Loeve expansion, or the "K-L expansion"
- Without loss of generality, we will sort the eigenvectors $\mathbf{b}_i$ in terms of their eigenvalues. That is $\lambda_1 >= \lambda_2 >= \ldots >= \lambda_d$. Then we refer to $\mathbf{b}_1$, corresponding to λ_1, as the "major eigenvector", or "principal component"

THE UNIVERSITY OF TENNESSEE KNOXVILLE

Dimensionality Reduction

AICIP RESEARCH

- Linear
  - Fisher's linear discriminant
    - Best discriminating the data
    - Supervised
  - Principal component analysis (PCA)
    - Best representing the data
    - Unsupervised