COSC 522 – Machine Learning
Lecture 7 – Dimensionality Reduction – FLD & PCA

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

Course Website: http://web.eecs.utk.edu/~hqi/cosc522/
Recap from Previous Lecture

- Understand the essential difference between supervised and unsupervised learning
- Understand agglomerative clustering
- Cluster characteristics using different distance metrics
- Understand kmeans and wta
Racap - Bayes Decision Rule

- Supervised learning
  - Baysian based - Maximum Posterior Probability (MPP): For a given \( x \), if \( P(w_1|x) > P(w_2|x) \), then \( x \) belongs to class 1, otherwise 2.
    - Parametric Learning
    - Case 1: Minimum Euclidean Distance (Linear Machine), \( \Sigma_i = \sigma^2 I \)
    - Case 2: Minimum Mahalanobis Distance (Linear Machine), \( \Sigma_i = \Sigma \)
    - Case 3: Quadratic classifier, \( \Sigma_i = \) arbitrary
    - Estimate Gaussian parameters using MLE
  - Nonparametric Learning
    - K-Nearest Neighbor
    - Least-square based
- Unsupervised learning
  - Kmeans
  - Winner-takes-all
- Supporting preprocessing techniques
  - Normalization
  - Dimensionality Reduction (FLD, PCA)
### COSC 522 - Machine Learning (Fall 2020) Syllabus

<table>
<thead>
<tr>
<th>Lecture</th>
<th>Date</th>
<th>Content</th>
<th>Tests</th>
<th>Assignment</th>
<th>Due Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8/20</td>
<td>Introduction</td>
<td></td>
<td>Project 1 - Supervised Learning</td>
<td>9/10</td>
</tr>
<tr>
<td>2</td>
<td>8/25</td>
<td>Baysian Decision Theory - MPP</td>
<td>HW0</td>
<td></td>
<td>8/25</td>
</tr>
<tr>
<td>3</td>
<td>8/27</td>
<td>Discriminant Function - MD</td>
<td>HW1</td>
<td></td>
<td>9/1</td>
</tr>
<tr>
<td>4</td>
<td>9/1</td>
<td>Parametric Learning - MLE</td>
<td>Proj1</td>
<td></td>
<td>9/10</td>
</tr>
<tr>
<td>5</td>
<td>9/3</td>
<td>Non-parametric Learning - kNN</td>
<td>Proj2</td>
<td></td>
<td>9/24</td>
</tr>
<tr>
<td>6</td>
<td>9/8</td>
<td>Supervised Learning - kmeans</td>
<td>Proj3</td>
<td></td>
<td>10/8</td>
</tr>
<tr>
<td>7</td>
<td>9/10</td>
<td>Dimensionality Reduction - FLD</td>
<td>Proj4</td>
<td>Final Project - Milestone 2: Choosing Topic</td>
<td>10/13</td>
</tr>
<tr>
<td>8</td>
<td>9/15</td>
<td>Dimensionality Reduction - PCA</td>
<td>Proj5</td>
<td>Final Project - Milestone 3: Literature Survey</td>
<td>10/27</td>
</tr>
<tr>
<td>9</td>
<td>9/17</td>
<td>Linear Regression</td>
<td></td>
<td></td>
<td>11/5</td>
</tr>
<tr>
<td>10</td>
<td>9/22</td>
<td>Performance Evaluation</td>
<td></td>
<td></td>
<td>9/29</td>
</tr>
<tr>
<td>11</td>
<td>9/24</td>
<td>Fusion</td>
<td>HW3</td>
<td></td>
<td>10/8</td>
</tr>
<tr>
<td>12</td>
<td>9/29</td>
<td>Midterm Exam</td>
<td></td>
<td>Final Project - Milestone 1: Forming Team</td>
<td>9/29</td>
</tr>
<tr>
<td>13</td>
<td>10/1</td>
<td>Gradient Descent</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>10/6</td>
<td>Neural Network - Perceptron</td>
<td>HW4</td>
<td></td>
<td>10/13</td>
</tr>
<tr>
<td>15</td>
<td>10/8</td>
<td>Neural Network - BPNN</td>
<td>Proj4 - BPNN</td>
<td></td>
<td>10/22</td>
</tr>
<tr>
<td>16</td>
<td>10/13</td>
<td>Neural Network - Practices</td>
<td>Final Project 4</td>
<td>Final Project - Milestone 4: Prototype</td>
<td>11/17</td>
</tr>
<tr>
<td>17</td>
<td>10/15</td>
<td>Kernel Methods - SVM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>10/20</td>
<td>Kernel Methods - SVM</td>
<td>HW5</td>
<td></td>
<td>10/27</td>
</tr>
<tr>
<td>19</td>
<td>10/22</td>
<td>Kernel Methods - SVM</td>
<td>Proj5 - SVM &amp; DT</td>
<td></td>
<td>11/5</td>
</tr>
<tr>
<td>20</td>
<td>10/27</td>
<td>Decision Tree</td>
<td></td>
<td>Final Project - Milestone 3: Literature Survey</td>
<td>10/27</td>
</tr>
<tr>
<td>21</td>
<td>10/29</td>
<td>Random Forest</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>11/3</td>
<td>From PCA to t-SNE</td>
<td>HW6</td>
<td></td>
<td>11/10</td>
</tr>
<tr>
<td>23</td>
<td>11/5</td>
<td>From Gaussian to Mixture and EM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>11/10</td>
<td>From Supervised/Unsupervised to RL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>11/17</td>
<td>From Classification/Regression to Generation</td>
<td></td>
<td>Final Project - Milestone 4: Prototype</td>
<td>11/17</td>
</tr>
<tr>
<td>26</td>
<td>11/19</td>
<td>From NN to CNN</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>11/24</td>
<td>Final Exam</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>12/3</td>
<td>Final Presentation</td>
<td>HW6</td>
<td>Final Project - Report</td>
<td>12/4</td>
</tr>
</tbody>
</table>

**Final Exam:** 8:00-10:15

**Final Presentation:** Final Project - Report
Questions

• 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 – 1<sup>st</sup> 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.

![Graph](image)

**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 \( \epsilon \sim N(0, \sigma^2) \). 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
Dimensionality Reduction

• Fisher’s linear discriminant
  – Best discriminating the data
• Principal component analysis (PCA)
  – Best representing the data
Questions

- 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?
Fisher’s Linear Discriminant

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

\[
y = \mathbf{w}^T \mathbf{x}
\]

\[
J(\mathbf{w}) = |\tilde{m}_1 - \tilde{m}_2|^2 = |\mathbf{w}^T (\mathbf{m}_1 - \mathbf{m}_2)|^2
\]

- Projected mean
- Sample mean

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

\[
\mathbf{m}_i = \frac{1}{n_i} \sum_{x \in D_i} \mathbf{x}
\]
Other Approaches?

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

\[
J(w) = \frac{\left| \mathbf{\bar{m}}_1 - \mathbf{\bar{m}}_2 \right|^2}{\hat{S}_1^2 + \hat{S}_2^2} = \frac{\left| w^T (\mathbf{m}_1 - \mathbf{m}_2) \right|^2}{w^T \mathbf{S}_1 w + w^T \mathbf{S}_2 w} = \frac{w^T \mathbf{S}_B w}{w^T \mathbf{S}_W w}
\]

\[
\hat{S}_i^2 = \sum_{y \in \mathcal{Y}_i} \left( y - \mathbf{\bar{m}}_i \right)^2 = \sum_{x \in \mathcal{D}_i} \left( w^T \mathbf{x} - w^T \mathbf{m}_i \right)^2 = \sum_{x \in \mathcal{D}_i} w^T (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T w = w^T S_i w
\]

Scatter matrix: \[S_i = \sum_{x \in \mathcal{D}_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T\]

Between-class scatter matrix: \[S_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T\]

Within-class scatter matrix: \[S_W = S_1 + S_2 = \sum_{i=1}^{2} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T\]
The Generalized Rayleigh Quotient

\[ J(w) = \frac{w^T S_B w}{w^T S_W w} \]

\[ \frac{dJ(w)}{dw} = \frac{2S_B w (w^T S_W w) - 2S_W w (w^T S_B w)}{(w^T S_W w)^2} = 0 \]

\[ S_B w = \frac{w^T S_B w}{w^T S_W w} S_W w \Rightarrow S_W^{-1} S_B w = \lambda w \]

\[ S_B w \text{ is always in the direction of } m_1 - m_2 \]

\[ w = S_W^{-1} (m_1 - m_2) \quad \text{Canonical variate} \]
Some Math Preliminaries

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

$$\frac{d}{dx}(x^T S x) = (S + S^T) x$$

Eigenvalue and eigenvector
- $x$ is called the eigenvector of $A$ iff $x$ is not zero, and $Ax = \lambda x$
- $\lambda$ 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)
Questions

• 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?
PCA Procedure

- Raw data $\rightarrow$ covariance matrix $\rightarrow$ eigenvalue $\rightarrow$ eigenvector $\rightarrow$ 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)

Describe vector \( \mathbf{x} \) in terms of a set of basis vectors \( \mathbf{b}_i \).

\[
\mathbf{x} = \sum_{i=1}^{d} y_i \mathbf{b}_i \quad \quad y_i = \mathbf{b}_i^T \mathbf{x}
\]

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)

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.

$$x = \sum_{i=1}^{m} y_i b_i + \sum_{i=m+1}^{d} y_i b_i \approx \sum_{i=1}^{m} y_i b_i + \sum_{i=m+1}^{d} \alpha_i b_i$$

Error: $\Delta x = \sum_{i=m+1}^{d} (y_i - \alpha_i) b_i$
K-L Transform (3)

- Use mean-square error to quantify the error

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

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

$$= \sum_{i=m+1}^{d} E \left\{ (y_i - \alpha_i)^2 \right\}$$
K-L Transform (4)

Find the optimal $\alpha_i$ to minimize $\varepsilon^2$

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

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

Therefore, the error is now equal to

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

$$= \sum_{i=m+1}^{d} E \left\{ b_i^T x - E\{b_i^T x\} \right\} = \sum_{i=m+1}^{d} E \left\{ b_i^T x - E\{b_i^T x\} \right\} \left\{ x^T b_i - E\{x^T b_i\} \right\}$$

$$= \sum_{i=m+1}^{d} b_i^T E \left\{ x - E\{x\} \right\} \left\{ x - E\{x\} \right\}^T b_i = \sum_{i=m+1}^{d} b_i^T \Sigma_x b_i = \sum_{i=m+1}^{d} \lambda_i$$
K-L Transform (5)

- The optimal choice of basis vectors is the eigenvectors of $\Sigma_x$
- 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 $b_i$ in terms of their eigenvalues. That is $\lambda_1 >= \lambda_2 >= \ldots >= \lambda_d$. Then we refer to $b_1$, corresponding to $\lambda_1$, as the “major eigenvector”, or “principal component”
Dimensionality Reduction

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