THE UNIVERSITY OF TENNESSEE KNOXVILLE

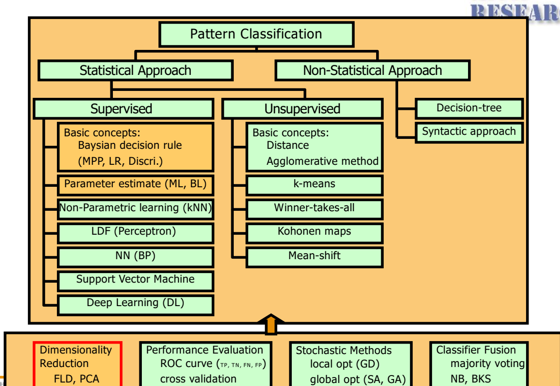
AICIP RESEARCH

ECE471-571 – Pattern Recognition

Lecture 6 – Dimensionality Reduction – Fisher's Linear Discriminant

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

AICIP RESEARCH



Review - Bayes Decision Rule

AICIP RESEARCH

$$P(\omega_j \mid x) = \frac{p(x \mid \omega_j)P(\omega_j)}{p(x)}$$

Maximum Posterior Probability: For a given x, if $P(\omega_1 \mid x) > P(\omega_2 \mid x)$, then x belongs to class 1, otherwise, 2.

Likelihood Ratio: If $\frac{p(x|\omega_1)}{p(x|\omega_2)} > \frac{P(\omega_2)}{P(\omega_1)}$, then x belongs to class 1, otherwise, 2.

Discriminant Function: The classifier will assign a feature vector x to class ω_i if $g_i(x) > g_j(x)$

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 , Σ_i = arbitrary

The Curse of Dimensionality – 1st Aspect

AICIP RESEARCH

- 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

THE UNIVERSITY OF TENNESSEE KNOXVILLE

4

Curse of Dimensionality – 2nd Aspect

AICIP RESEARCH

- Accuracy and overfitting
- In theory, the higher the dimensionality, the less the error, the better the performance. However, in realistic pattern recognition 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

THE UNIVERSITY OF TENNESSEE KNOXVILLE

5

Curse of Dimensionality - 3rd Aspect

AICIP RESEARCH

- Computational complexity

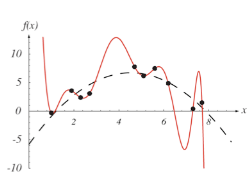


FIGURE 3.4. The "training data" (black dots) were selected from a quadratic function plus Gaussian noise, i.e., $f(x) = ax^2+bx+c+\epsilon$ where $p(\epsilon) \sim N(0, \sigma^2)$. The 10th-degree polynomial shown fits the data perfectly, but we desire instead the second-order function $f(x)$, because 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.

THE UNIVERSITY OF TENNESSEE KNOXVILLE

6

Dimensionality Reduction

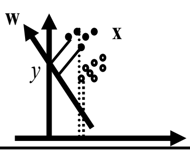
AICIP RESEARCH

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

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



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

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

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

Projected mean　　Sample mean

Other Approaches?

AICIP RESEARCH

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

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

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

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

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 $\mathbf{m}_1 - \mathbf{m}_2$

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

THE UNIVERSITY OF TENNESSEE KNOXVILLE

10

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**

THE UNIVERSITY OF TENNESSEE KNOXVILLE

11

* Multiple Discriminant Analysis

AICIP RESEARCH

- For c-class problem, the projection is from d-dimensional space to a (c-1)-dimensional space (assume d >= c)
- Sec. 3.8.3

THE UNIVERSITY OF TENNESSEE KNOXVILLE

12