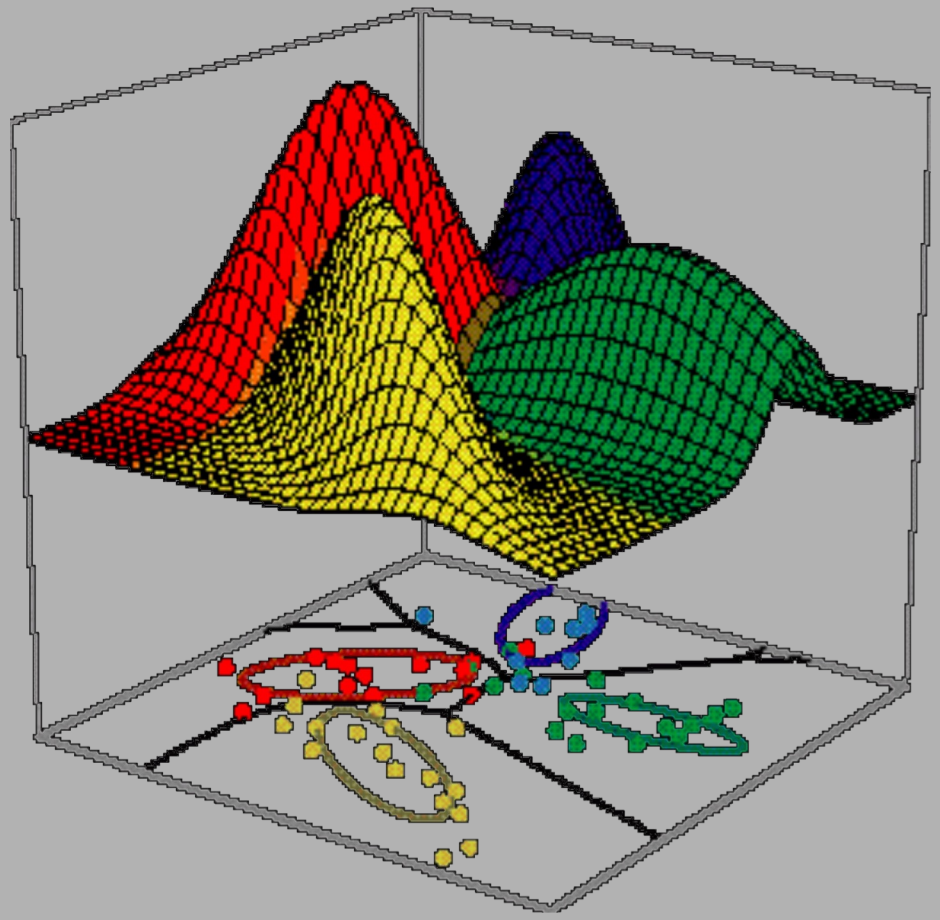


# Pattern Recognition (Αναγνώριση Προτύπων)

## Dimensionality Issues

**(Θέματα Σχετικά με τη Διάσταση των Χαρακτηριστικών)**

Panos Trahanias

UNIVERSITY OF CRETE
DEPARTMENT of COMPUTER SCIENCE

# Dimensionality Reduction

- Large number of features may give rise to overfitting if the features are not appropriate and if the employed distributions are not correctly estimated.
- Moreover, large number of features affects the design of the classifier, due to space / time complexity issues.
- Combination of features can be used to confront the "dimensionality" problem
- Appropriate approaches:
  - Principal component analysis
  - Fisher linear discriminant, etc.

# Overfitting

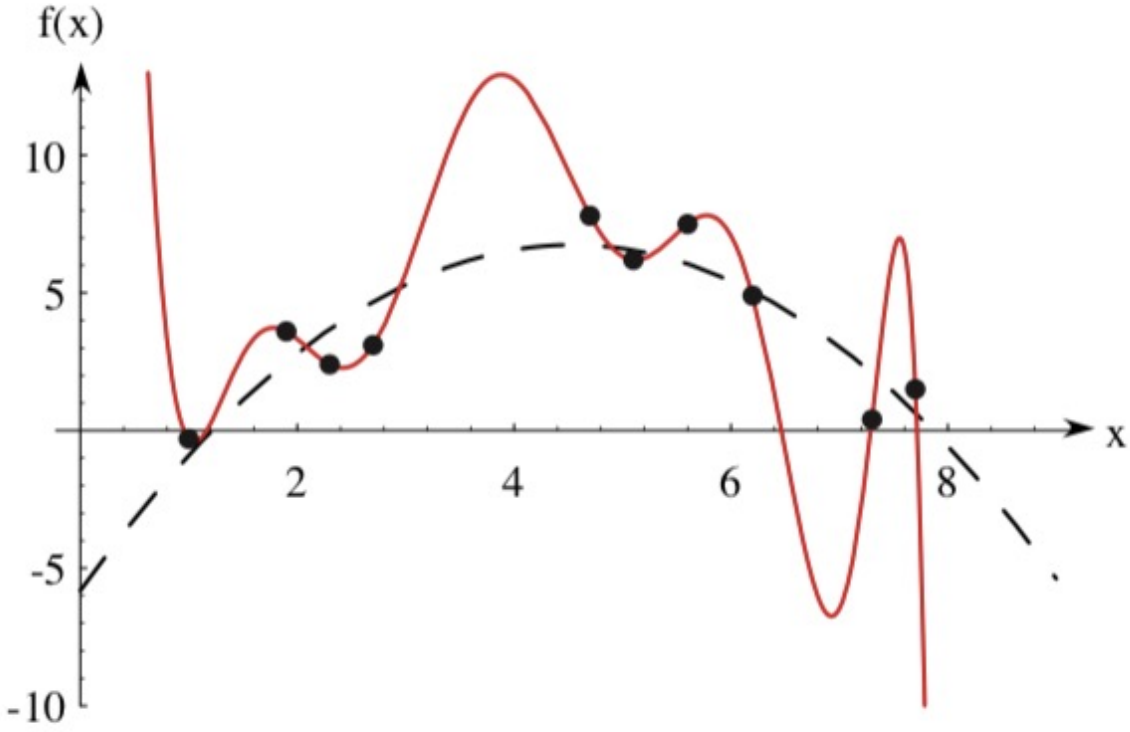



Figure 3.4: The "training data" (black dots) were selected from a quadradic function plus Gaussian noise, i.e., $f(x) = ax^2 + bx + c + \epsilon$ where $p(\epsilon) \sim N(0, \sigma^2)$. The 10th degree polynomial shown fits the data perfectly, but we desire instead the second-order function $f(x)$, since it would lead to better predictions for new samples.

# Principal Component Analysis (PCA)

- Employed to represent a set of *n* *d-dimensional* vectors via the use of a unique, *l-dimensional vector*, where $l<d$.

- The optimal 1-D representation of the samples, in the least squares sense, is their projection onto a line that passes from the mean of the samples and is in the direction of the eigenvector of the ***scatter matrix*** that corresponds to the largest eigenvalue!

- In the case of l-dimensions, the optimal representation of the samples is the *l*-eigenvectors of the scatter matrix, that correspond to the *l* largest eigenvalues.

# Principal Component Analysis (PCA)

- In the case of l-dimensions, the optimal representation of the samples is the *l*-eigenvectors of the scatter matrix, that correspond to the *l* largest eigenvalues.

data scatter matrix

$$S_D = \sum_{x_i} (x_i - m)(x_i - m)^T$$

$$\mathbf{Mx} = \lambda\mathbf{x}$$

$$(\mathbf{M} - \lambda\mathbf{I})\mathbf{x} = 0,$$

$$\mathbf{Me}_j = \lambda_j\mathbf{e}_j.$$

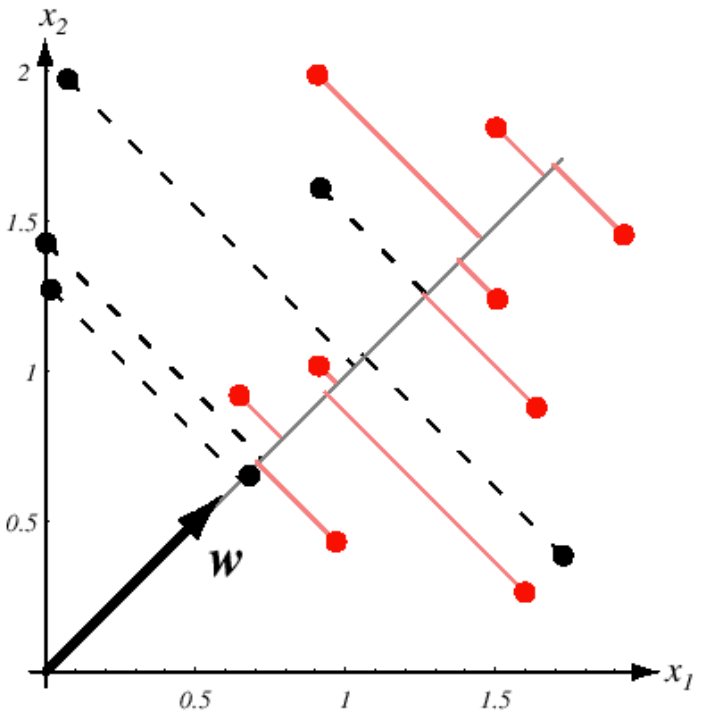
# Fisher Linear Discriminant (FLD)

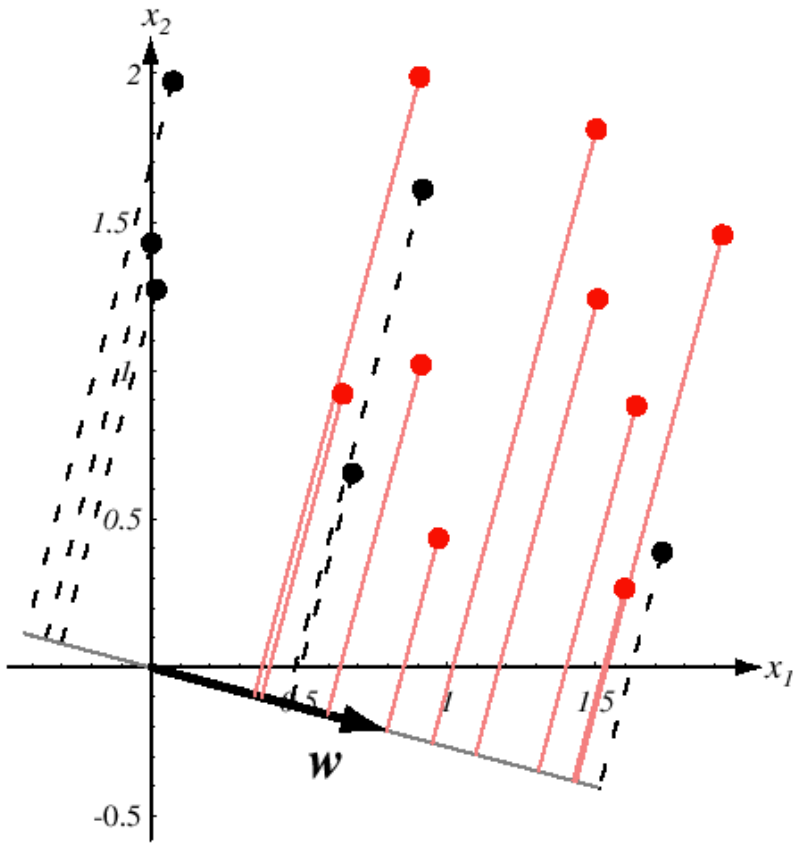
- PCA defines the minimum number of components that represent the samples.
- This representation is in the least squares sense. It does not however guarantee its appropriateness with respect to classification.

- Ideally we'd like to reduce dimensionality under the constraint of maximizing separability of patterns.
- Pattern separability maximization can be achieved by increasing *intercluster* distances and decreasing *intracluster* distances. These distances are calculated via the use of intercluster scatter matrices and intracluster scatter matrices, respectively.
- FLD is based on such a transformation.

$$y = \boldsymbol{\omega}^T \mathbf{x}$$

# Fisher Linear Discriminant (FLD)





- 2-D FLD maximizes the criterion function

Intercluster scatter matrix

$$J(\omega) = \frac{\boldsymbol{\omega}^T S_B \boldsymbol{\omega}}{\boldsymbol{\omega}^T S_W \boldsymbol{\omega}}$$

Intracluster scatter matrix

$$S_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T$$

$$S_W = \sum_{i=1}^{2} \sum_{x \in D_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T$$

$$\boldsymbol{\omega} = S_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$$

# FLD in Larger Dimensions

- In a *c*-class problem, FLD can reduce dimension from *d* to *c-1*. Obviously, $c<d$ . As before, we need to estimate matrix **W** for the following transformation:

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

- Note that dimensions will be as:
  - **x:** [d x m] where *m* is the number of samples and *d* the dimension of each sample (number of features)
  - **W:** [d x c-1], where *c* is the number of classes
  - **y:** [c-1 x m]

# FLD in Larger Dimensions

- We start by defining the «inter» and «intra» scatter matrices as follows, for the initial feature space **x**:

$$S_i = \sum_{x \in D_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T \quad \Rightarrow \quad S_W = \sum_{i=1}^{c} S_i$$

$D_i$ is the sample set of class $i$

$S_W$ is the interclass scatter matrix

$$S_B = \sum_{i=1}^{c} n_i (\mathbf{m}_1 - \mathbf{m})(\mathbf{m}_1 - \mathbf{m})^T$$

$n_i$ is the number of samples in class $i$

$S_B$ is the intraclass scatter matrix

- In the transformed (y) space, the above matrices become

$$\widetilde{S}_W = \mathbf{W}^T S_W \mathbf{W}$$

$$\widetilde{S}_B = \mathbf{W}^T S_B \mathbf{W}$$

# FLD in Larger Dimensions

- FLD derives the transformation matrix **W** that maximizes the following criterion function. Note that by maximizing this function the interclass distance decreases (by minimizing interclass matrix) and the intraclass distance increases (by maximizing intraclass matrix).

$$J(\mathbf{W}) = \frac{\left|\widetilde{S}_B\right|}{\left|\widetilde{S}_W\right|} = \frac{\left|\mathbf{W}^T S_B \mathbf{W}\right|}{\left|\mathbf{W}^T S_W \mathbf{W}\right|}$$

- Turns out that, the columns of optimal **W** are the eigenvectors that correspond to the largest eigenvalues of

$$S_B \mathbf{w}_i = \lambda_i S_W \mathbf{w}_i$$

# Wrap-Up

- Bayesian Decision Theory
- Maximum Likelihood and Bayesian Estimation
- Nonparametric Techniques
  - Parzen Windows
  - K-n Nearest Neighbor
- Linear Discriminant Functions
  - Perceptron
  - Relaxation Procedures
  - MSE
  - Multicategory Generalization
- Dimensionality Reduction
  - Principal component analysis
  - Fisher linear discriminant