# Foundations of Machine Learning

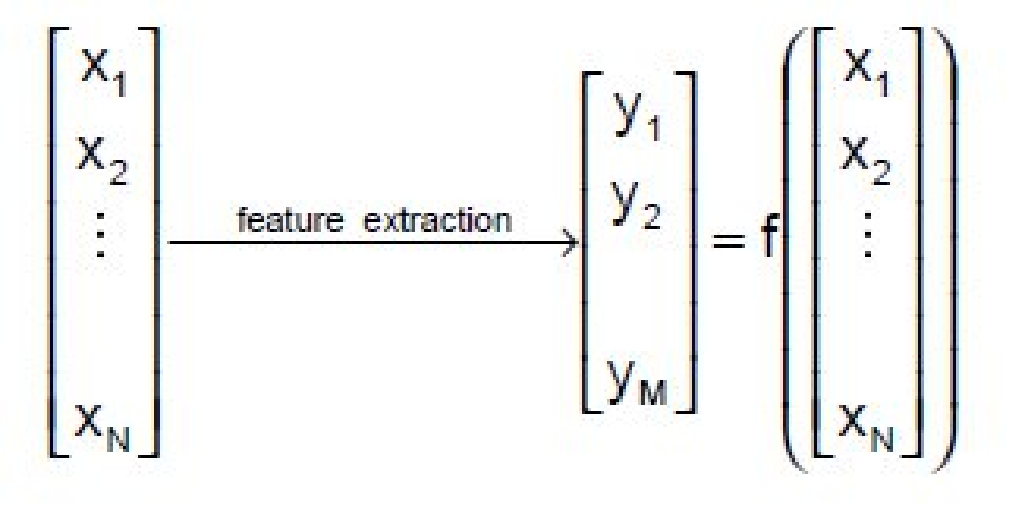
## Module 3: Instance Based Learning and Feature Reduction

## Part C: Feature Extraction

Sudeshna Sarkar
IIT Kharagpur

# Feature extraction - definition

- Given a set of features $F = \{x_1, \ldots, x_N\}$
  the Feature Extraction("Construction") problem is
  is to map $F$ to some feature set $F''$ that maximizes
  the learner's ability to classify patterns



# Feature Extraction

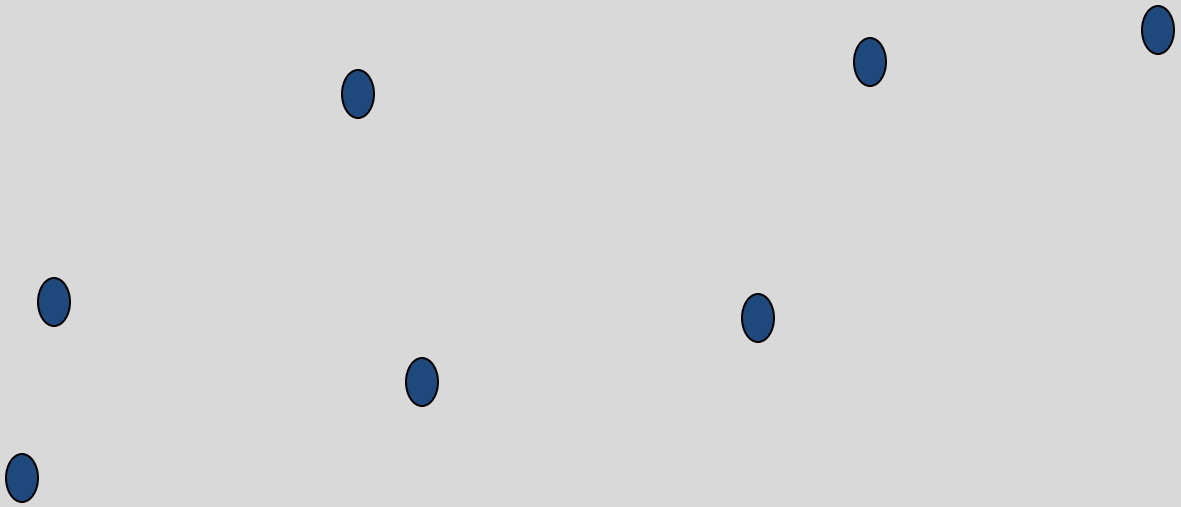
- Find a projection matrix w from N-dimensional to M-dimensional vectors that keeps error low
- Assume that N features are linear combination of $\mathrm{M} < N$ vectors

$$z_i = w_{i1}x_{i1} + \ldots + w_{id}x_{iN}$$

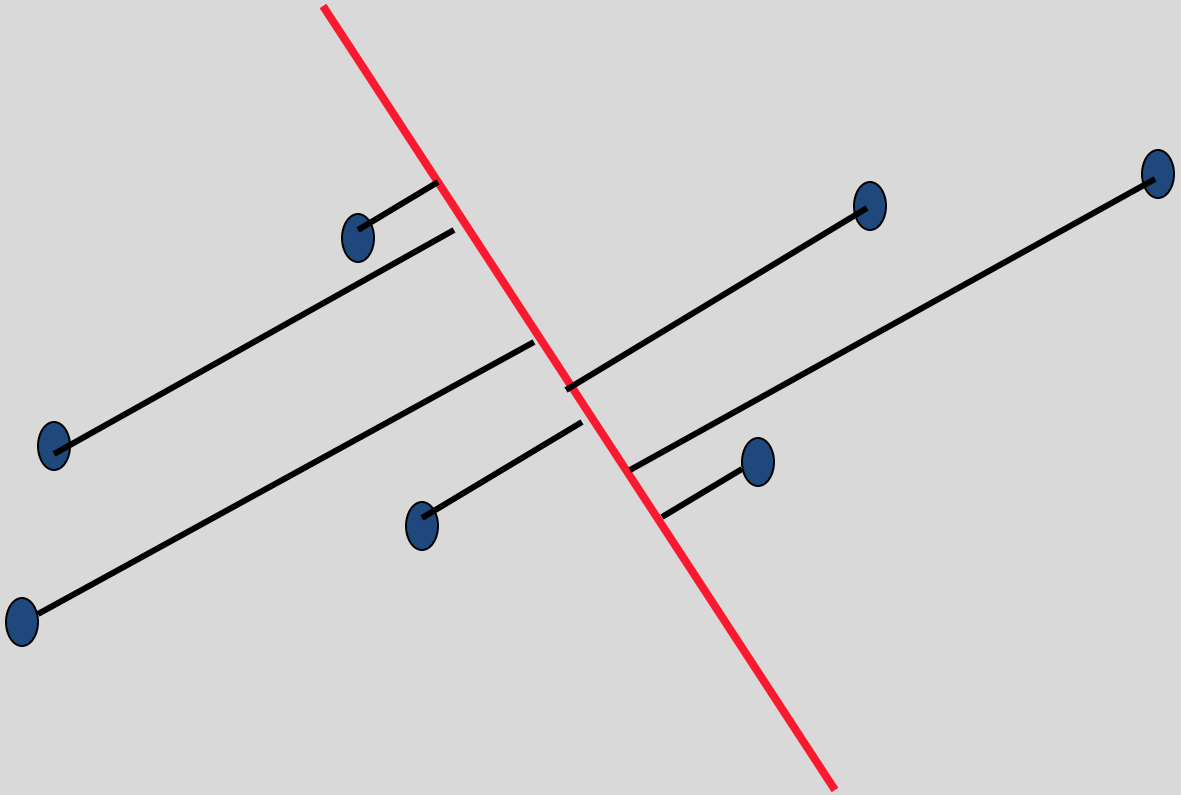
$$\boldsymbol{z} = w^T \boldsymbol{x}$$

- What we expect from such basis
  - Uncorrelated, cannot be reduced further
  - Have large variance or otherwise bear no information

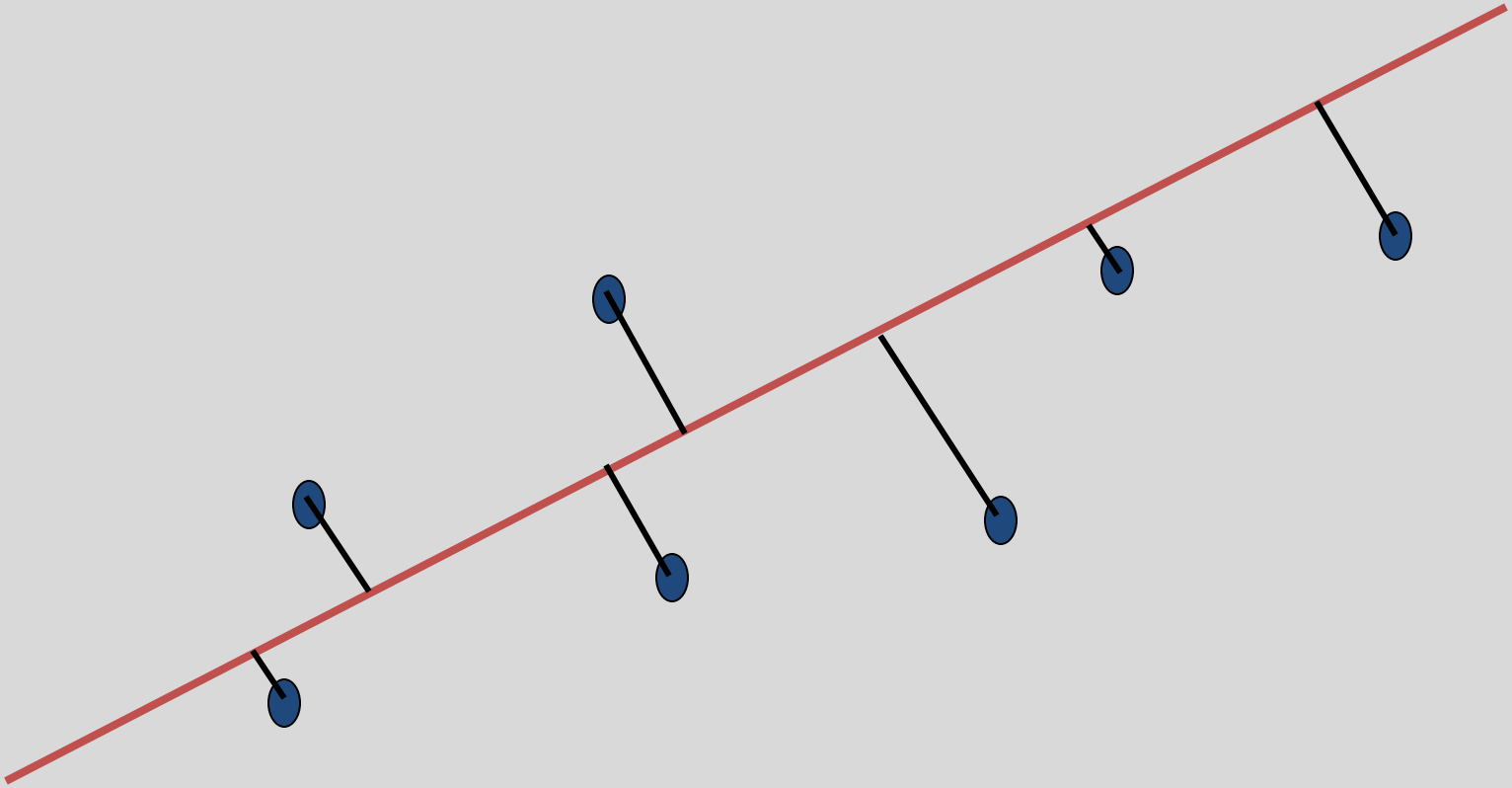
# Geometric picture of principal components (PCs)



# Geometric picture of principal components (PCs)



# Geometric picture of principal components (PCs)



# Algebraic definition of PCs

Given a sample of *p* observations on a vector of *N* variables

$$\{x_1, x_2, \cdots, x_n\} \in \Re^N$$

define the first principal component of the sample
by the linear transformation

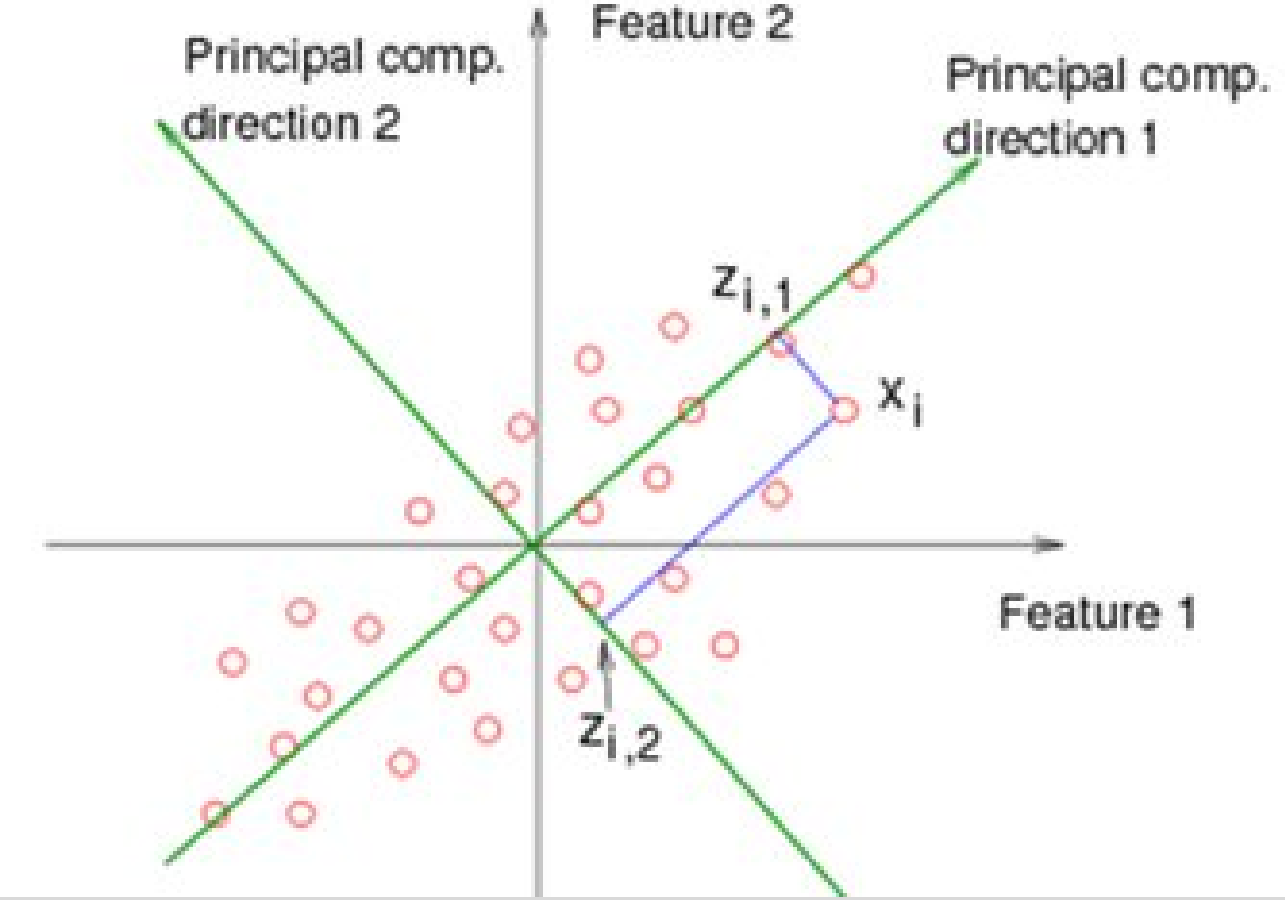
$$z_1 = w_1^T x_j = \sum_{i=1}^{N} w_{i1} x_{ij}, \quad j = 1,2,\cdots, p.$$

where the vector $w_1 = (w_{11}, w_{21}, \cdots, w_{N1})$

$x_j = (x_{1j}, x_{2j}, \cdots, x_{Nj})$

is chosen such that $\mathrm{var}[z_1]$ is maximum.

# PCA


Feature 2
Principal comp.
direction 2
Principal comp.
direction 1
$z_{i,1}$
$x_i$
Feature 1
$z_{i,2}$

# PCA

- Choose directions such that a total variance of data will be maximum
  1. Maximize Total Variance
- Choose directions that are orthogonal
  2. Minimize correlation

- Choose $M < N$ orthogonal directions which maximize total variance

# PCA

- $N$-dimensional feature space
- $\mathrm{N} \times N$ symmetric covariance matrix estimated from samples $Cov(\boldsymbol{x}) = \Sigma$
- Select $M$ largest eigenvalue of the covariance matrix and associated $M$ eigenvectors
- The first eigenvector will be a direction with largest variance

# PCA for image compression

p=1

p=2

p=4

p=8

p=16

p=32

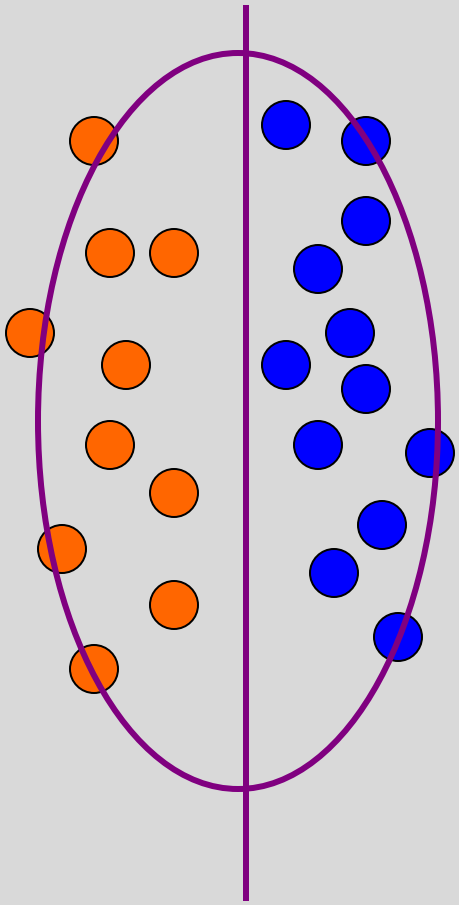
p=64

p=100

Original Image

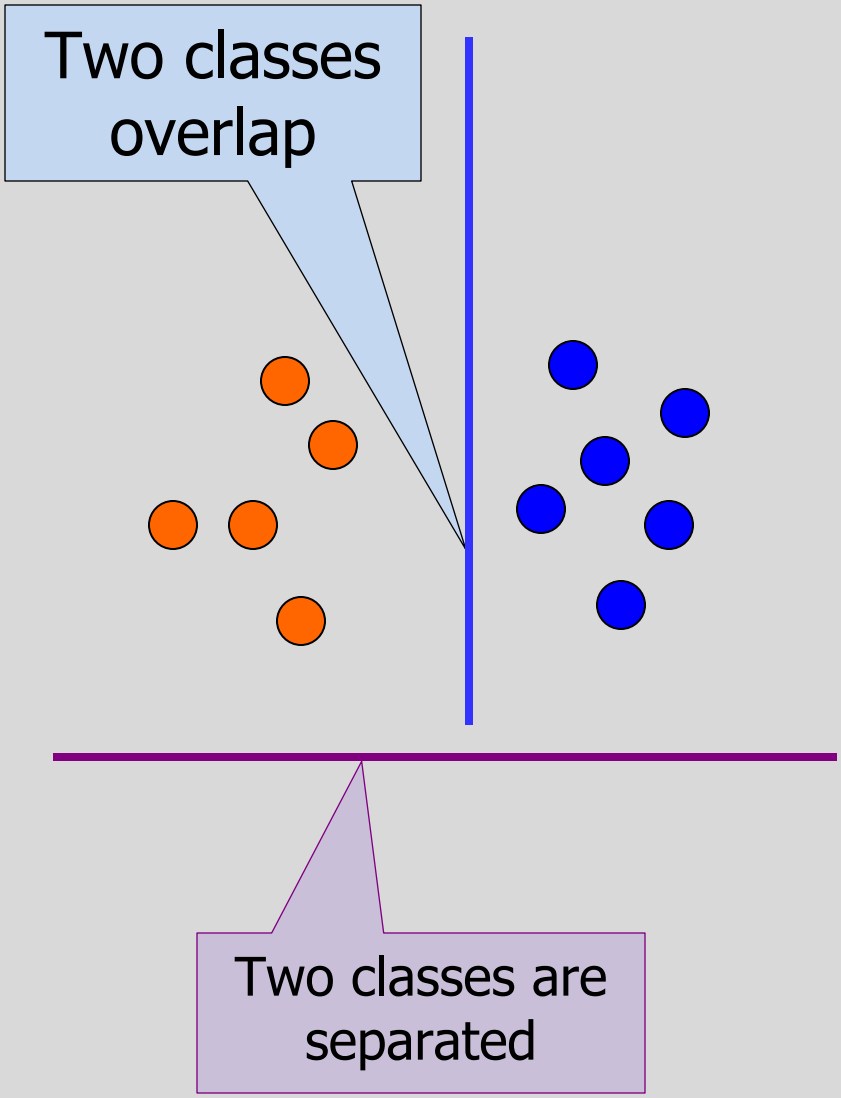
# Is PCA a good criterion for classification?

- Data variation determines the projection direction
- What's missing?
  - Class information



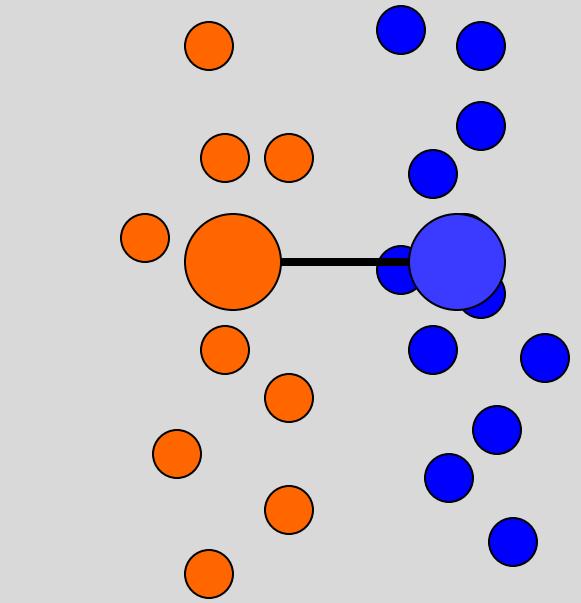
# What is a good projection?

- Similarly, what is a good criterion?
  - Separating different classes



# What class information may be useful?

- Between-class distance
  - Distance between the centroids of different classes

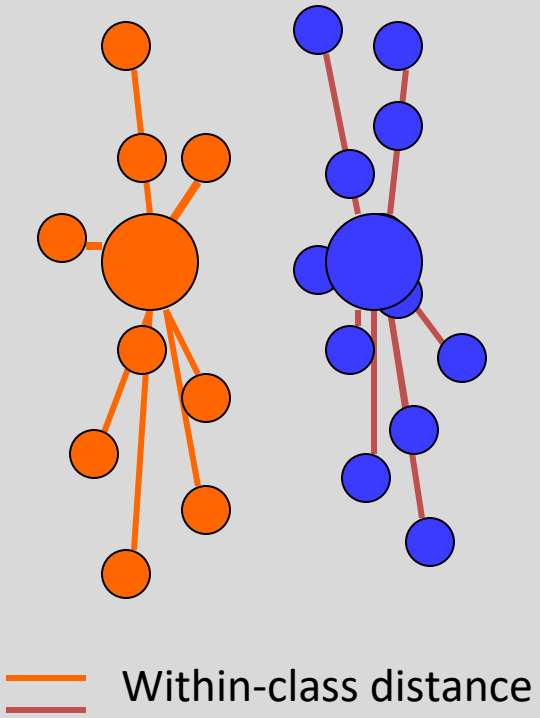


Between-class distance

# What class information may be useful?

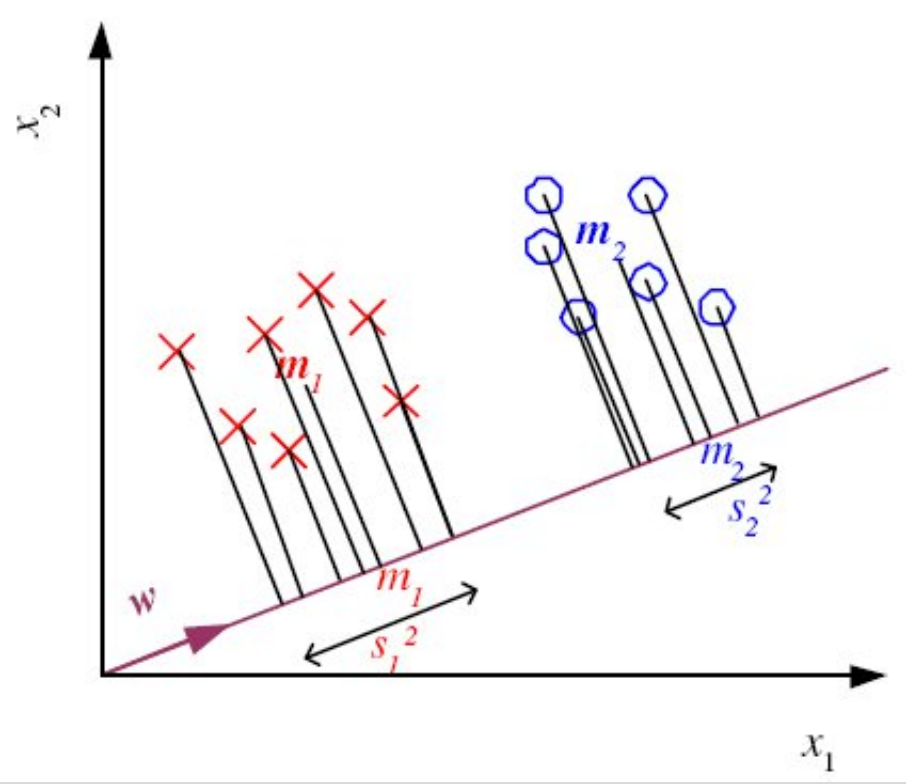
- Between-class distance
  - Distance between the centroids of different classes
- Within-class distance
  - Accumulated distance of an instance to the centroid of its class

- Linear discriminant analysis (LDA) finds most discriminant projection by
  - maximizing between-class distance
  - and minimizing within-class distance



# Linear Discriminant Analysis

- Find a low-dimensional space such that when $\boldsymbol{x}$ is projected, classes are well-separated



# Means and Scatter after projection

$$m_1 = \frac{\sum_t \mathbf{w}^T \mathbf{x}^t r^t}{\sum_t r^t} = \mathbf{w}^T \mathbf{m}_1$$

$$m_2 = \frac{\sum_t \mathbf{w}^T \mathbf{x}^t (1 - r^t)}{\sum_t (1 - r^t)} = \mathbf{w}^T \mathbf{m}_2$$

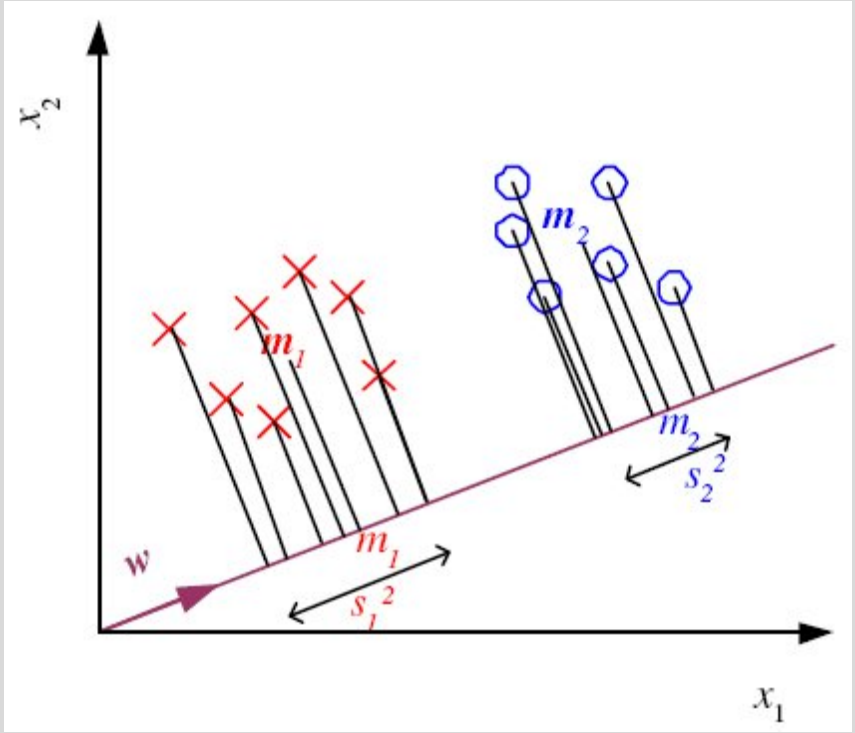
$$s_1^2 = \sum_t (\mathbf{w}^T \mathbf{x}^t - m_1)^2 r^t$$

$$s_2^2 = \sum_t (\mathbf{w}^T \mathbf{x}^t - m_2)^2 (1 - r^t)$$

# Good Projection

- Means are as far away as possible
- Scatter is small as possible
- Fisher Linear Discriminant

$$J(\mathbf{w}) = \frac{(m_1 - m_2)^2}{s_1^2 + s_2^2}$$



# Multiple Classes

- For $c$ classes, compute $c-1$ discriminants, project N-dimensional features into $c-1$ space.

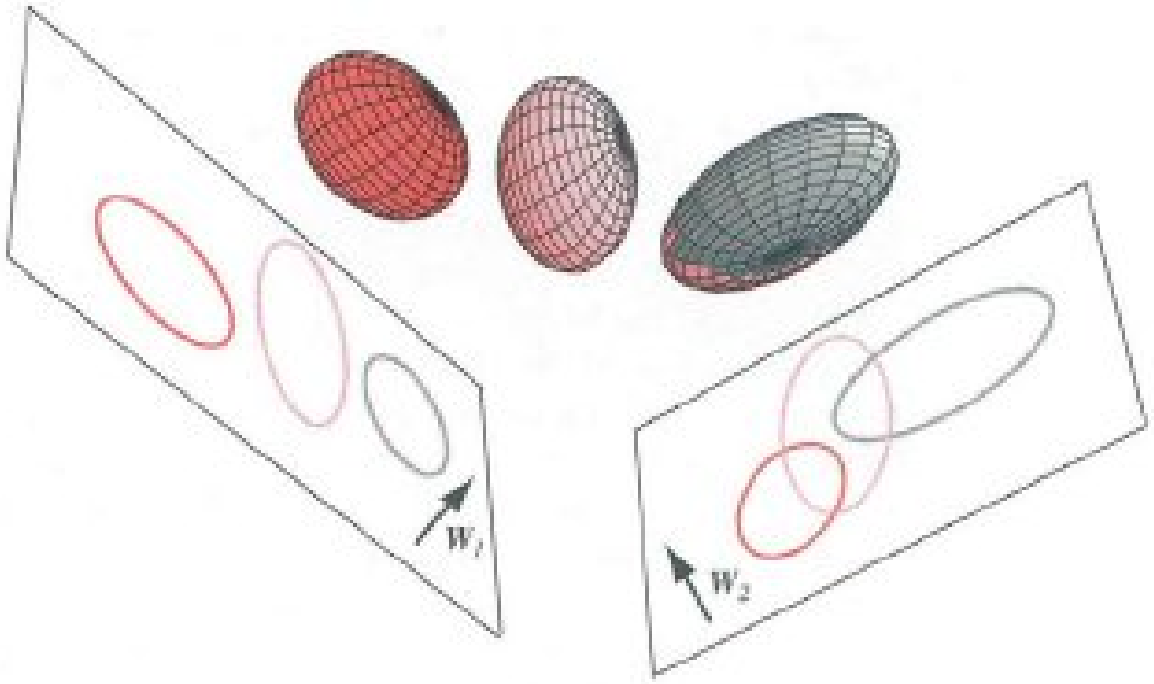



FIGURE 3.6. Three three-dimensional distributions are projected onto two-dimensional subspaces, described by a normal vectors $\mathbf{W}_1$ and $\mathbf{W}_2$. Informally, multiple discriminant methods seek the optimum such subspace, that is, the one with the greatest separation of the projected distributions for a given total within-scatter matrix, here as associated with $\mathbf{W}_1$.

Thank You