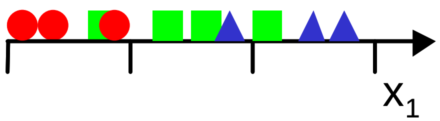
Lecture 10: Dimensionality reduction

- **The curse of dimensionality**
- **Feature extraction vs. feature selection**
- **Principal Components Analysis**
- **Linear Discriminant Analysis**

Dimensionality reduction

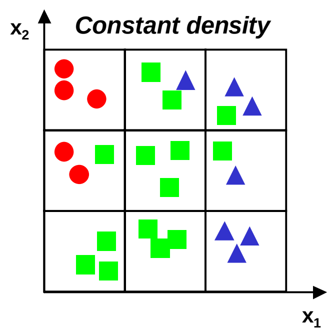
- **The "curse of dimensionality"**
  - Refers to the problems associated with multivariate data analysis as the dimensionality increases
- **Consider a 3-class pattern recognition problem**
  - Three types of objects have to be classified based on the value of a single feature:

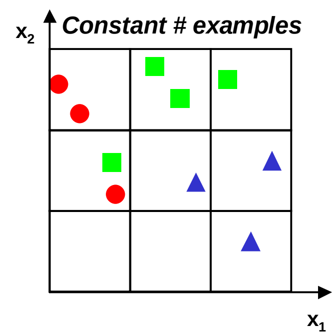


  - A simple procedure would be to
    - Divide the feature space into uniform bins
    - Compute the ratio of examples for each class at each bin and,
    - For a new example, find its bin and choose the predominant class in that bin
  - We decide to start with one feature and divide the real line into 3 bins
    - Notice that there exists a lot of overlap between classes ⇒ to improve discrimination, we decide to incorporate a second feature

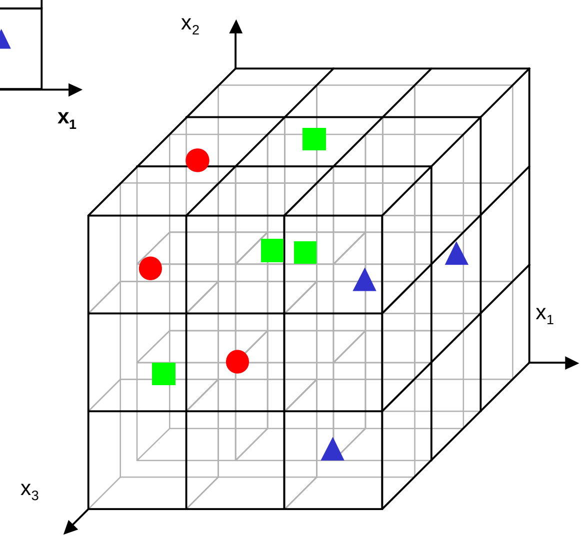
Dimensionality reduction

- **Moving to two dimensions increases the number of bins from 3 to 3^2=9**
  - QUESTION: Which should we maintain constant?
    - The density of examples per bin? This increases the number of examples from 9 to 27
    - The total number of examples? This results in a 2D scatter plot that is very sparse



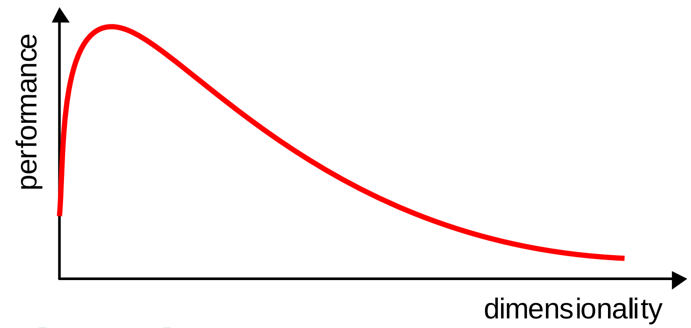


- **Moving to three features …**
  - The number of bins grows to 3^3=27
  - To maintain the initial density of examples, the number of required examples grows to 81
  - For the same number of examples the 3D scatter plot is almost empty



Dimensionality reduction

- **Implications of the curse of dimensionality**
  - Exponential growth with dimensionality in the number of examples required to accurately estimate a function
- **In practice, the curse of dimensionality means that**
  - For a given sample size, there is a maximum number of features above which the performance of our classifier will degrade rather than improve
    - In most cases, the information that was lost by discarding some features is compensated by a more accurate mapping in lower-dimensional space



- **How do we beat the curse of dimensionality?**
  - By incorporating prior knowledge
  - By providing increasing smoothness of the target function
  - By reducing the dimensionality

Dimensionality reduction

- **Two approaches to perform dim. reduction $\Re^N \rightarrow \Re^M$ (M<N)**
  - **Feature selection**: choosing a subset of all the features

$$[x_1\ x_2 \ldots x_N] \xrightarrow{\text{feature selection}} [x_{i_1}\ x_{i_2} \ldots x_{i_M}]$$

  - **Feature extraction**: creating new features by combining existing ones

$$[x_1\ x_2 \ldots x_N] \xrightarrow{\text{feature extraction}} [y_1\ y_2 \ldots y_M] = f\left([x_{i_1}\ x_{i_2} \ldots x_{i_M}]\right)$$

    - In either case, the goal is to find a low-dimensional representation of the data that preserves (most of) the information or structure in the data
    - Feature extraction is covered in more detail in CS790

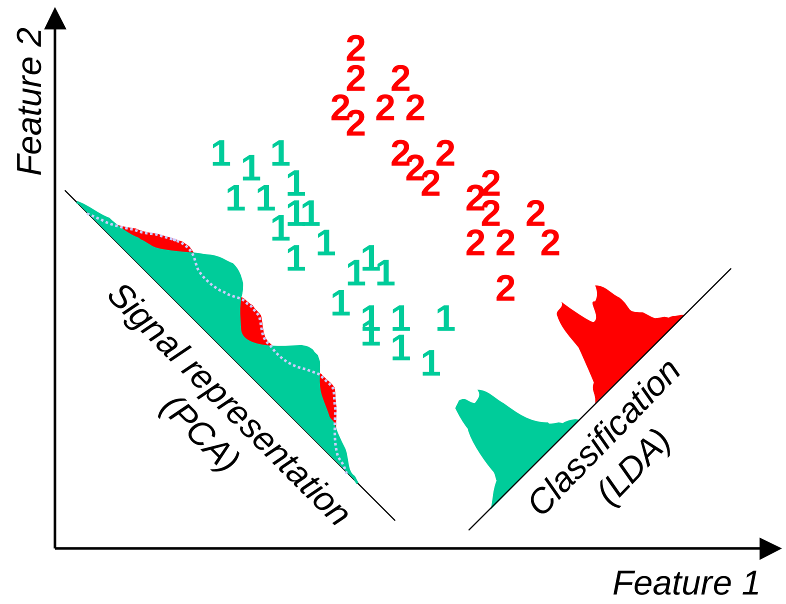
- **Linear feature extraction**
  - The "optimal" mapping y=f(x) is, in general, a non-linear function whose form is problem-dependent
    - Hence, feature extraction is commonly limited to linear projections **y=Wx**

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ \\ x_N \end{bmatrix} \xrightarrow{\text{linear feature extraction}} \begin{bmatrix} y_1 \\ y_2 \\ \\ y_M \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1N} \\ w_{21} & w_{22} & \cdots & w_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ w_{M1} & w_{M2} & & w_{MN} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ \\ x_N \end{bmatrix}$$

Signal representation versus classification

- **Two criteria can be used to find the "optimal" feature extraction mapping y=f(x)**
  - **Signal representation**: The goal of feature extraction is to represent the samples accurately in a lower-dimensional space
  - **Classification**: The goal of feature extraction is to enhance the class-discriminatory information in the lower-dimensional space

- **Within the realm of linear feature extraction, two techniques are commonly used**
  - Principal Components (PCA)
    - Based on signal representation
  - Fisher's Linear Discriminant (LDA)
    - Based on classification



Principal Components Analysis

- **Let us illustrate PCA with a two dimensional problem**
  - Data $\underline{x}$ follows a Gaussian density as depicted in the figure
  - Vectors can be represented by their 2D coordinates:

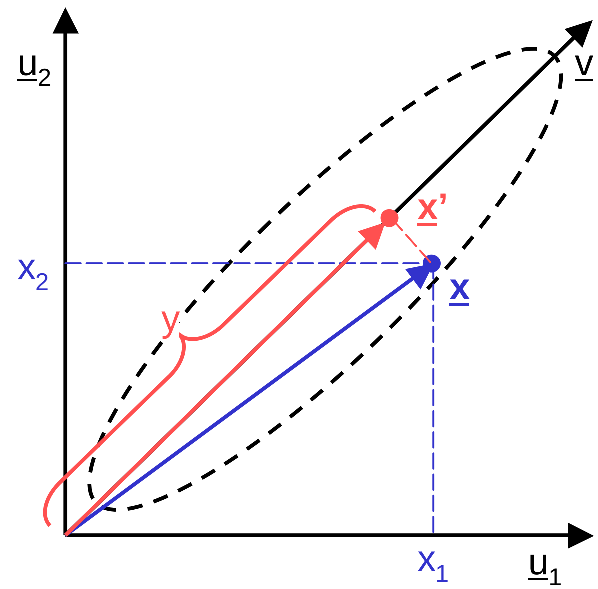
$$\underline{x} = x_1\underline{u}_1 + x_2\underline{u}_2 = (x_1, x_2)_{\underline{u}_1,\underline{u}_2}$$

  - We seek to find a 1D representation $\underline{x}$' "close" to $\underline{x}$

$$\underline{x}' = y\underline{v} = (y)_{\underline{v}}$$

  - Where "closeness" is measured by the mean squared error over all points in the distribution

$$(y)_{\underline{v}} = \operatorname{argmin} E\left[\left[\underline{x}' - \underline{x}\right]^2\right]$$



Principal Components Analysis

- **RESULT (for proof check CS790 notes)**
  - It can be shown that the "optimal"1D representation consists of projecting the vector $\underline{x}$ over the direction of maximum variance in the data (e.g., the longest axis in the ellipse)
- **This result can be generalized for more than two dimensions**

> The optimal* approximation of a random vector $\underline{x} \in \Re^N$ by a linear combination of M (M<N) independent vectors is obtained by projecting the random vector $\underline{x}$ onto the eigenvectors $\underline{v}_i$ corresponding to the largest eigenvalues λ_i of the covariance matrix of x (Σ_x)

Principal Components Analysis

- **Summary**

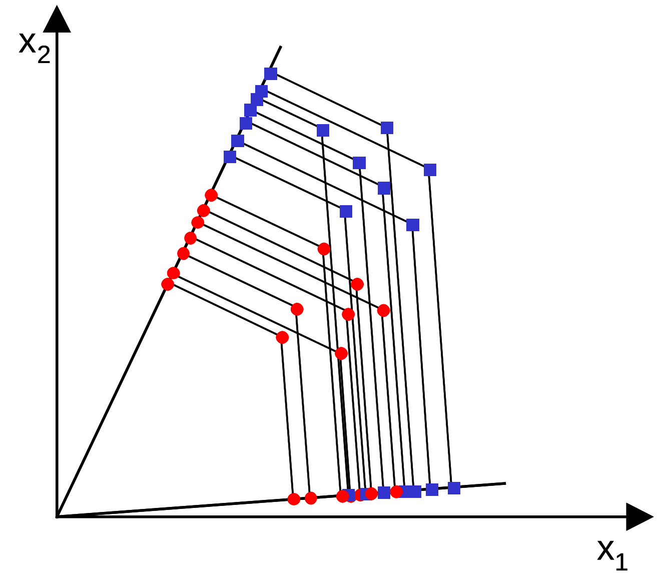
$$\underline{x}' = y_1\underline{v}_1 + y_2\underline{v}_2 \cdots + y_M\underline{v}_M$$

$$\underline{x}' = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{bmatrix} = \begin{bmatrix} \underline{v}_1^T \\ \underline{v}_2^T \\ \\ \underline{v}_M^T \end{bmatrix} \underline{x} = \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1N} \\ v_{21} & v_{22} & & \\ \vdots & & \ddots & \\ v_{M1} & v_{M2} & \cdots & v_{MN} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_N \end{bmatrix}$$

- where $\underline{v}_k$ is the eigenvector corresponding to the k^{th} largest eigenvalue of the covariance matrix

Linear Discriminant Analysis, two-classes

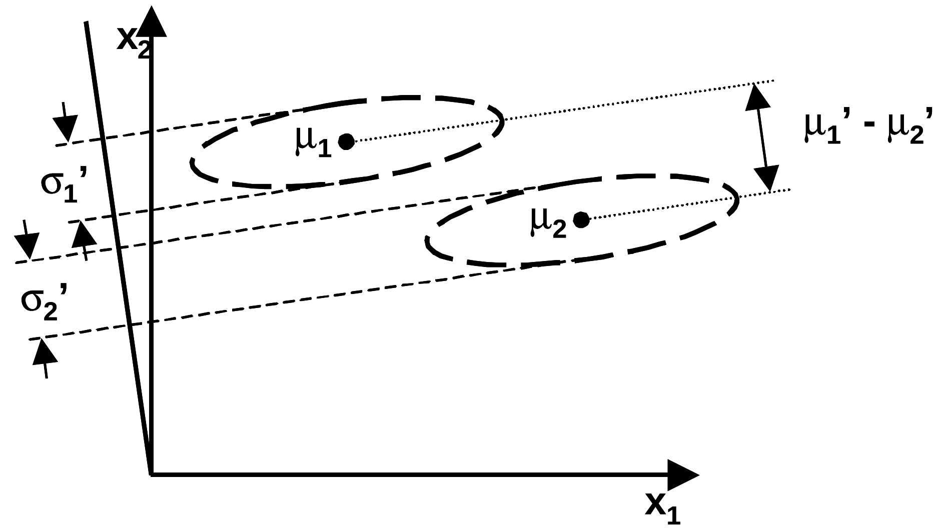
- **The objective of LDA is to perform dimensionality reduction while preserving as much of the class discriminatory information as possible**
  - Assume we have a set of N-dimensional samples $(x_1, x_2, \ldots, x_N)$, P_1 of which belong to class ω_1, and P_2 to class ω_2. We seek to obtain a scalar y by projecting the samples x onto a line
  - Of all the possible lines we would like to select the one that maximizes the separability of the classes



Linear Discriminant Analysis

- **In a nutshell, we want**
  - Maximum separation between the means of the projection
  - Minimum variance within each projected class

$$\text{maximize} \quad \frac{(\mu_1{}'-\mu_2{}')^2}{\sigma_1{}'^2-\sigma_2{}'^2}$$



Linear Discriminant Analysis

- **RESULT (for proof check CS790 notes)**
  - It can be shown that the optimal projection matrix W* is the one whose columns are the eigenvectors corresponding to the largest eigenvalues of the following generalized eigenvalue problem

$$\boxed{V = \left[v_1 \,|\, v_2 \,|\, \cdots \,|\, v_{C-1}\right] = \operatorname{argmin}\left\{\frac{V^T S_B V}{V^T S_W V}\right\} \Rightarrow \left(S_B - \lambda_i S_W\right)v_i = 0}$$

  - Where S_B and S_W are the BETWEEN-CLASS and WITHIN-CLASS covariance matrices

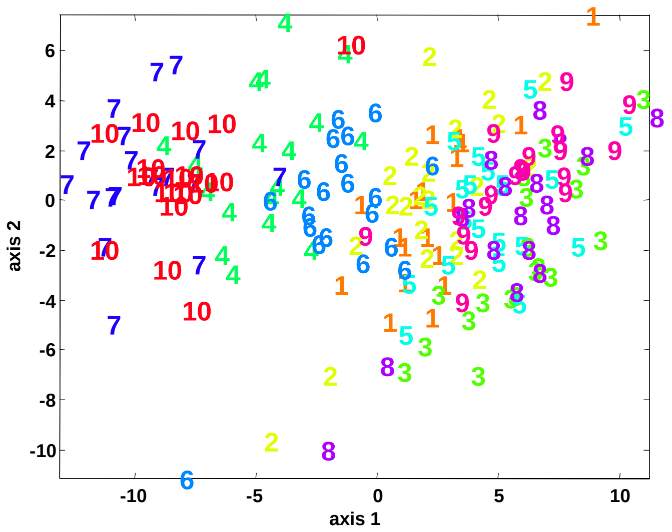
$$S_W = \sum_{i=1}^{C} S_i = \sum_{i=1}^{C} \sum_{x \in \omega_i} (x - \mu_i)(x - \mu_i)^T$$

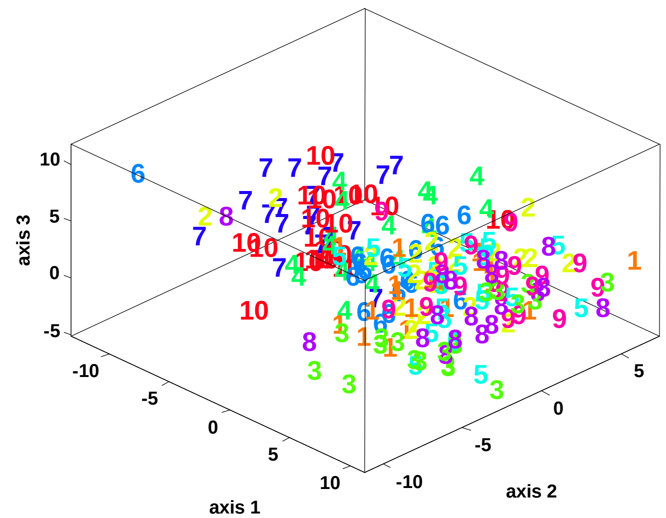
$$S_B = \sum_{i=1}^{C} P_i (\mu_i - \mu)(\mu_i - \mu)^T$$

$$\textit{where} \quad \mu_i = \frac{1}{N_i} \sum_{x \in \omega_i} x \;\; \textit{and} \;\; \mu = \frac{1}{N} \sum_{\forall x} x$$

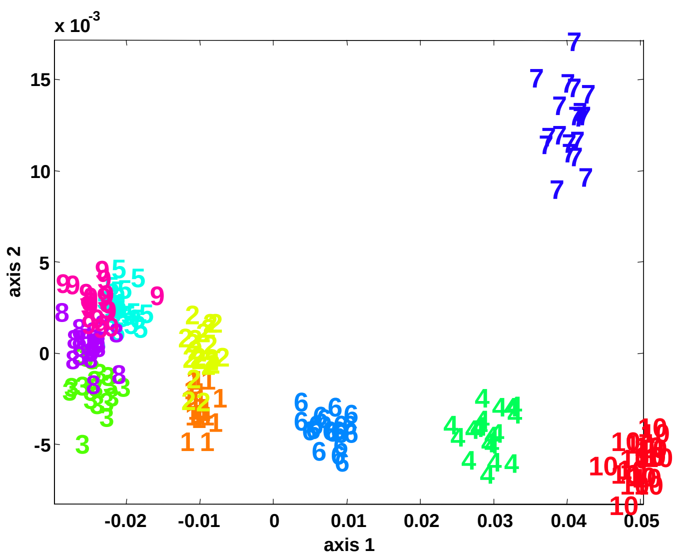
PCA Versus LDA

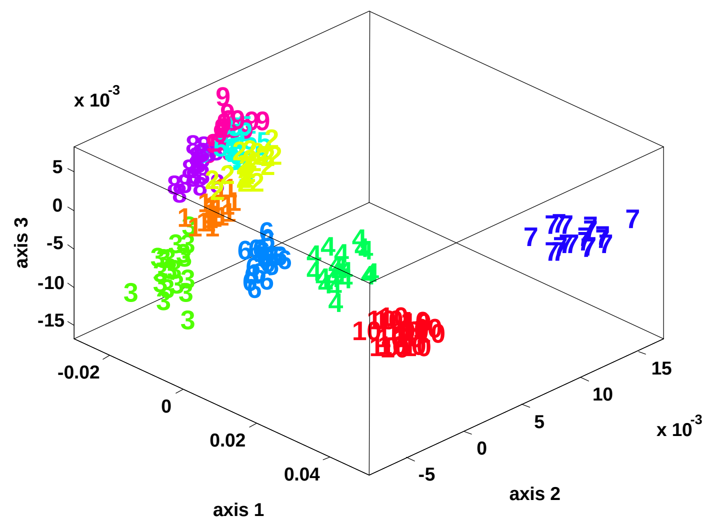
PCA





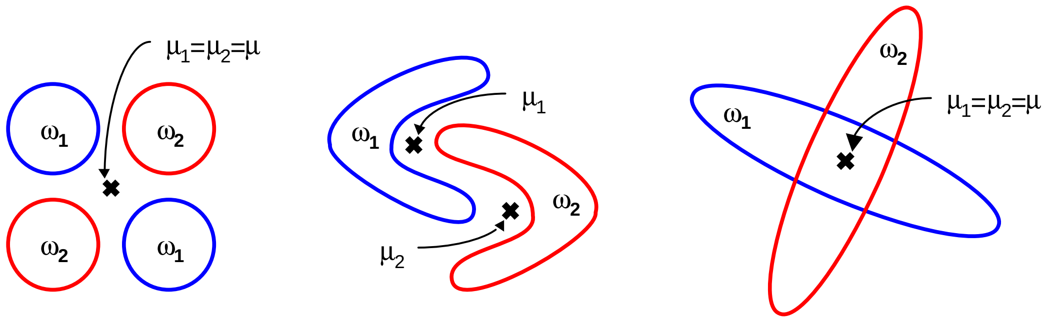
LDA



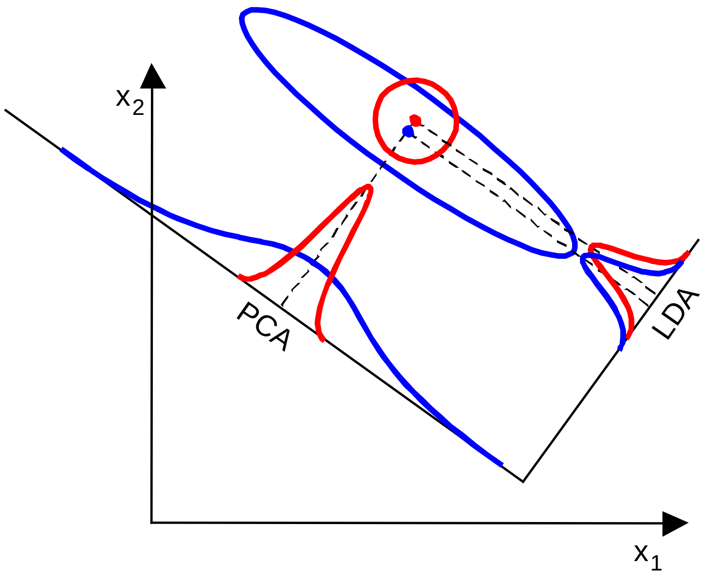


Limitations of LDA

- **LDA assumes unimodal Gaussian likelihoods**
  - If the densities are significantly non-Gaussian, LDA may not preserve any complex structure of the data needed for classification



- **LDA will fail when the discriminatory information is not in the mean but rather in the variance of the data**



Limitations of LDA

- **LDA has a tendency to overfit training data**
  - To illustrate this problem, we generate an artificial dataset
    - Three classes, 50 examples per class, with the <u>exact</u> same likelihood: a multivariate Gaussian with zero mean and identity covariance
    - As we arbitrarily increase the number of dimensions, classes appear to separate better, even though they come from the same distribution

10 dimensions **50 dimensions**

100 dimensions **150 dimensions**