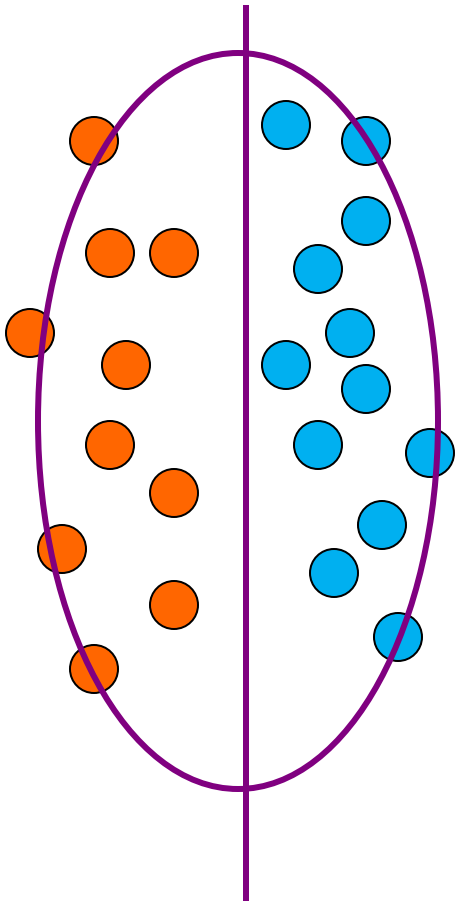
CSC345/M45:
Big Data & Machine Learning
(dimensionality reduction: LDA)

Prof. Xianghua Xie

x.xie@swansea.ac.uk
http://csvision.swan.ac.uk
224 Computational Foundry, Bay Campus

PCA Limitations

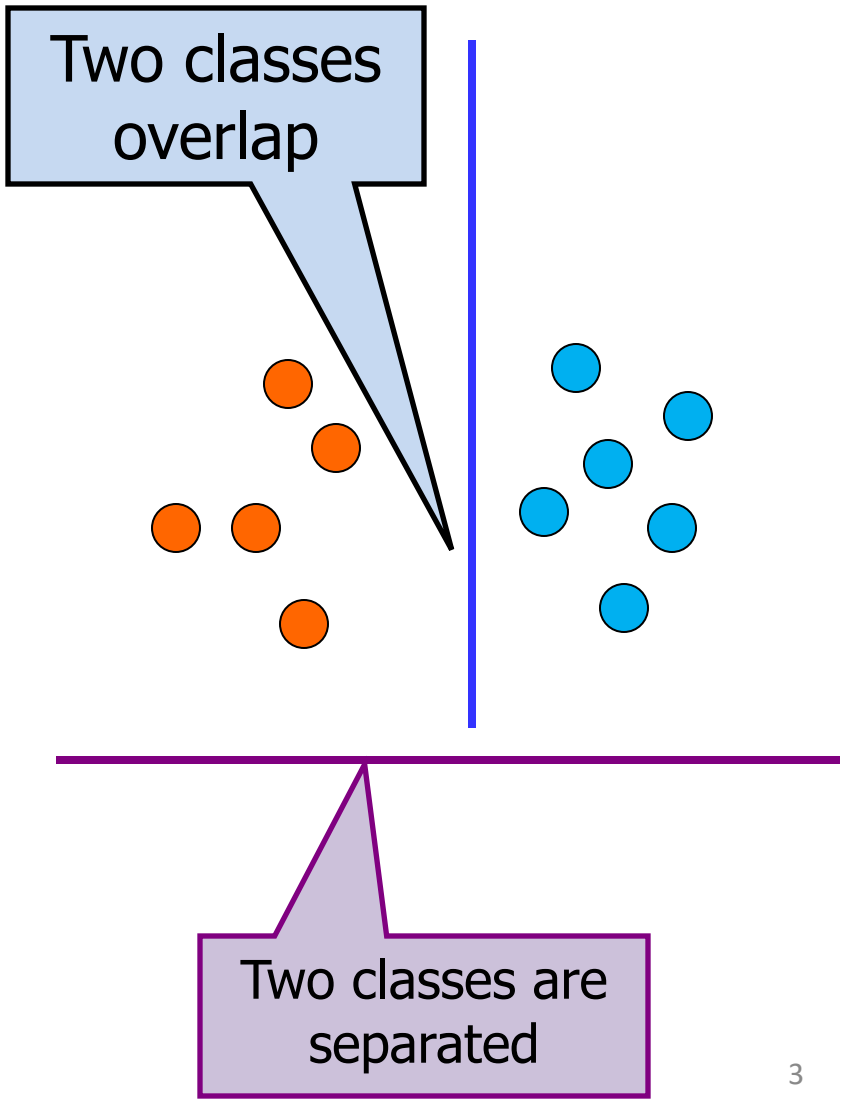
- PCA is unsupervised analysis (no class labels)
- In the case of supervised learning (with class labels)
  - PCA is generally not optimal for classification or dimensionality reduction
  - In PCA: global data variation determines projection orientation



What is a good projection?

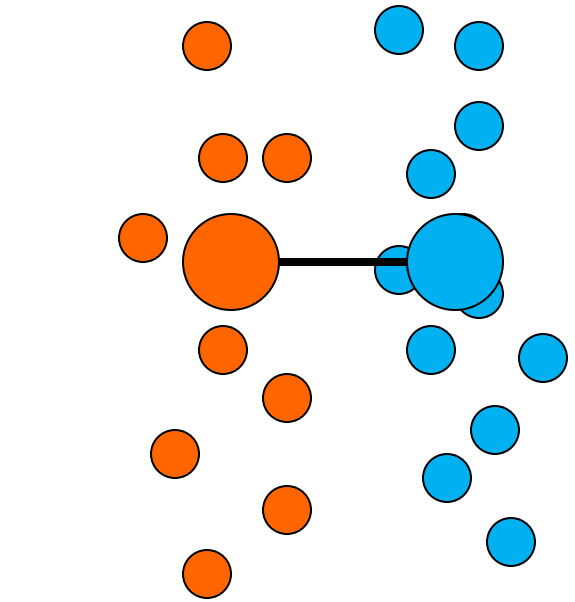
- A good projection that reduces dimensionality but also preserves class separability.

- For the example on the left
  - Both projections reduce dimensionality: 2D -> 1D
  - Projection onto the blue axis (vertical) will result in two classes overlap each other
  - Project onto the purple axis (horizontal) will result in two well separated classes



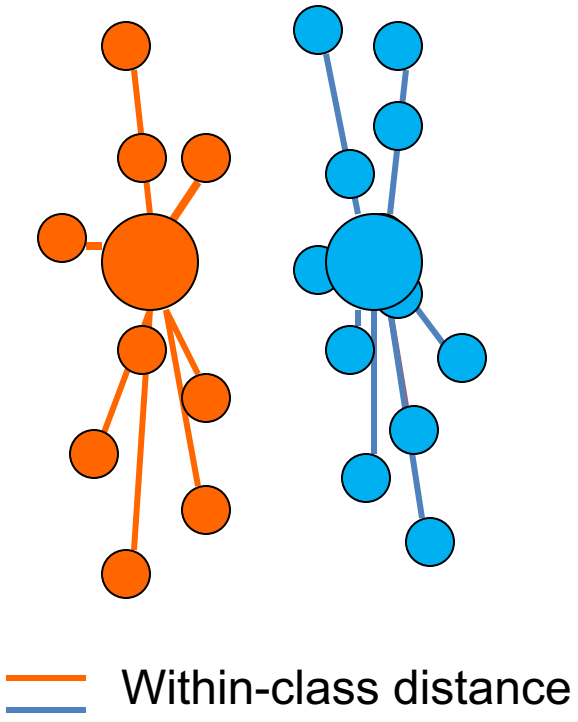
What class information may be useful?

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



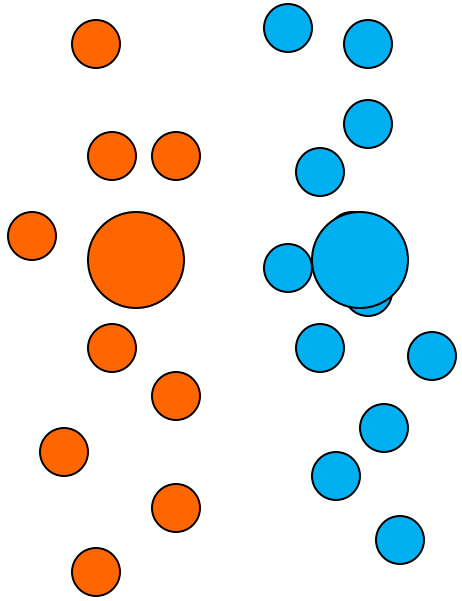
What class information may be useful?

- 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

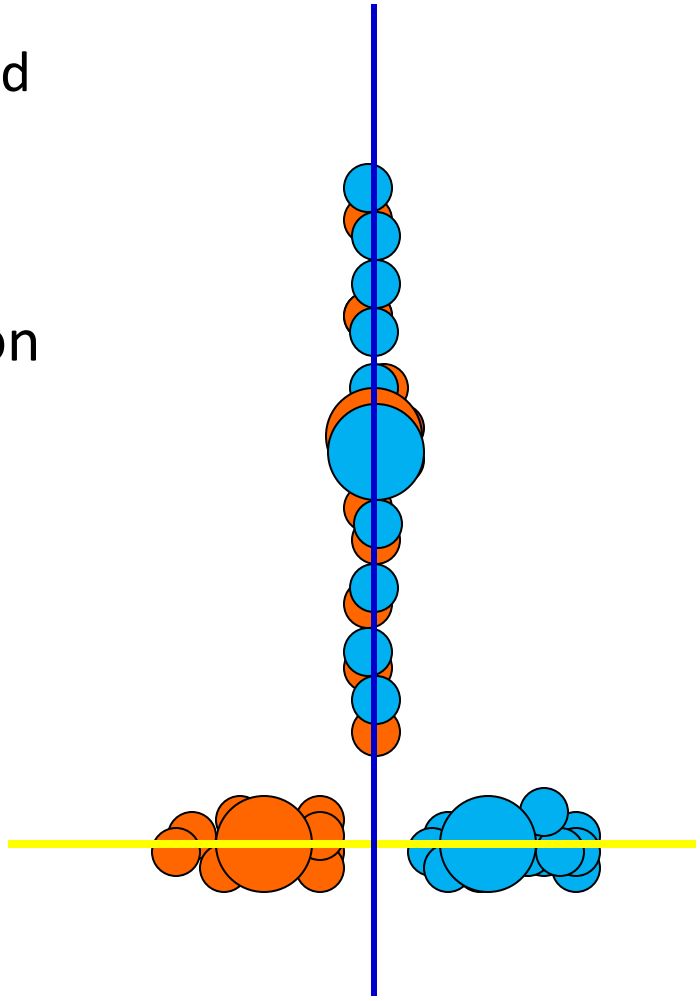
- Linear discriminant analysis (LDA) finds most discriminant projection by
  - maximising between-class distance and
  - minimising within-class distance



Linear Discriminant Analysis

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

- Thus, perform dimensionality reduction while preserving as much of the class discriminatory information as possible

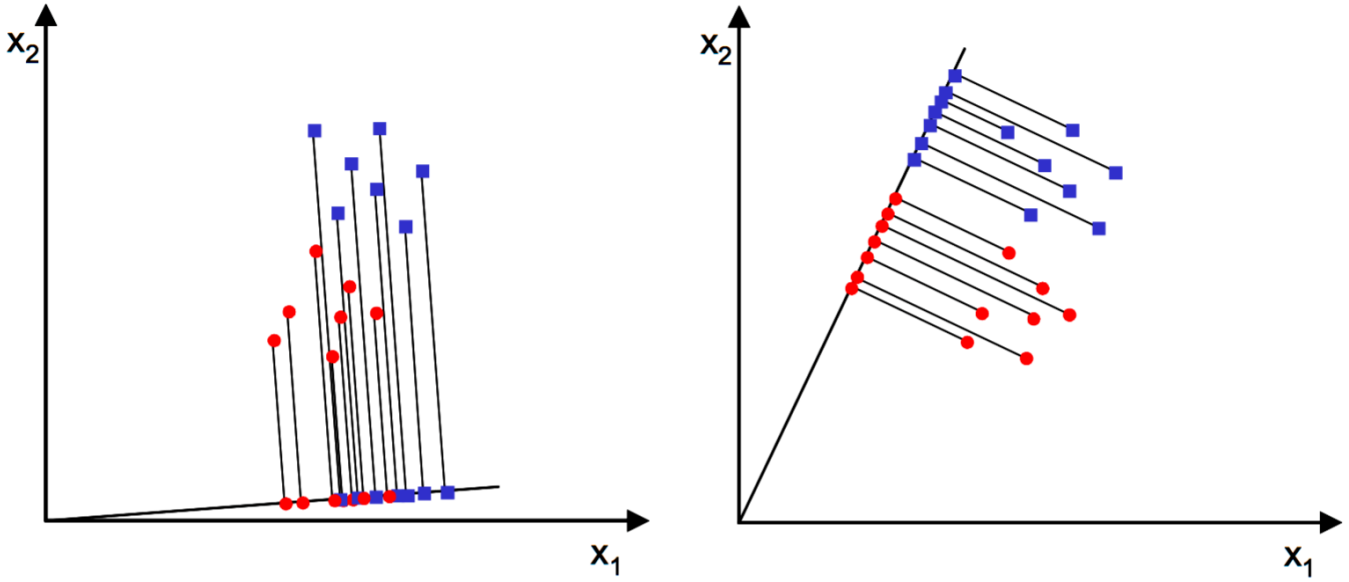


Linear Discriminant Analysis

- Assume we have a set of 2-dimensional samples $\{x^{(1}, x^{(2}, \ldots, x^{(N}\}$
  - N_1 of these samples belong to class ω_1, and N_2 belong to class ω_2.
- We seek to obtain a scalar y by projecting the samples x onto a line (w defines the projection):

$$y = w^T x$$

  - Of all possible lines we would like to select the one that maximises the separability of the scalars



Linear Discriminant Analysis

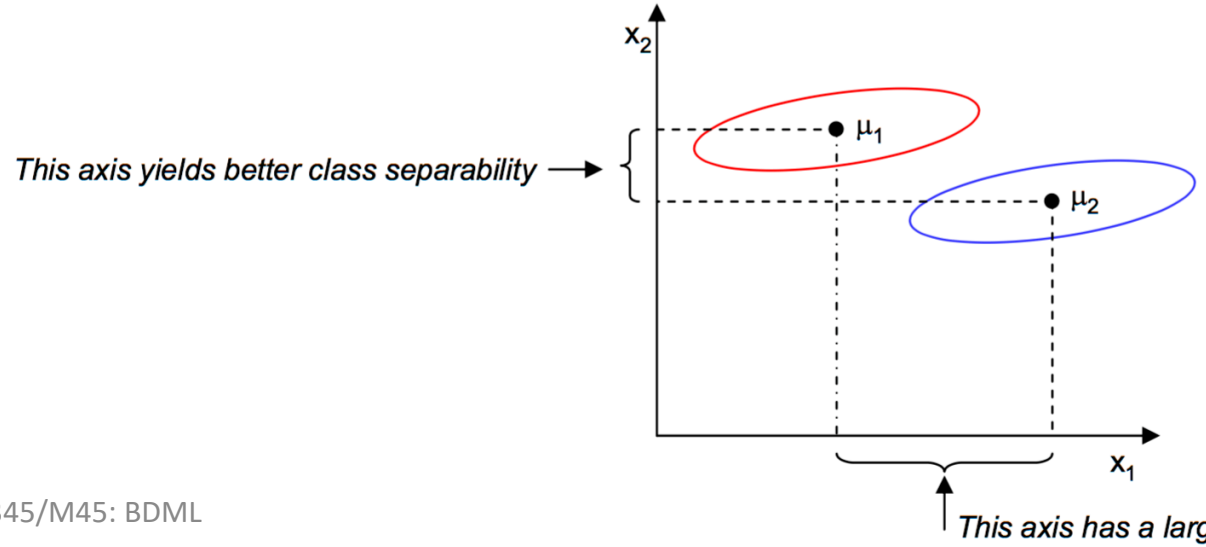
- In order to find a good projection vector, we need to define a measure of separation between the projections
  - The mean vector of each class in x and y feature space is:

$$\mu_i = \frac{1}{N_i}\sum_{x\in\omega_i} x \qquad \tilde{\mu}_i = \frac{1}{N_i}\sum_{y\in\omega_i} y = \frac{1}{N_i}\sum_{x\in\omega_i} w^T x = w^T \mu_i$$

  - The distance between the projected mean values gives a measure of separability of two classes after projection:

$$\left|\tilde{\mu}_1 - \tilde{\mu}_2\right| = \left|w^T(\mu_1 - \mu_2)\right|$$

    - But on its own, it is not sufficient, see an example below:



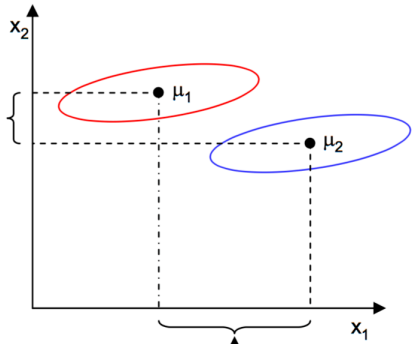
Linear Discriminant Analysis

- The solution proposed by Fisher is to maximise a function that represents the difference between the means, and normalised by a measure of the within-class scatter.
  - For each class, we define the **scatter** (an equivalent of the variance) as:

$$\tilde{s}_i^2 = \sum_{y \in \omega_i} (y - \tilde{\mu}_i)^2$$

    - Where the quantity $\left(\tilde{s}_1^2 + \tilde{s}_2^2\right)$ is called the **within-class scatter** of the projected samples.
  - The Fisher's linear discriminant is defined as the linear function $\mathbf{w}^T\mathbf{x}$ that maximises the criterion function J:

$$J(w) = \frac{\left|\tilde{\mu}_1 - \tilde{\mu}_2\right|^2}{\tilde{s}_1^2 + \tilde{s}_2^2}$$

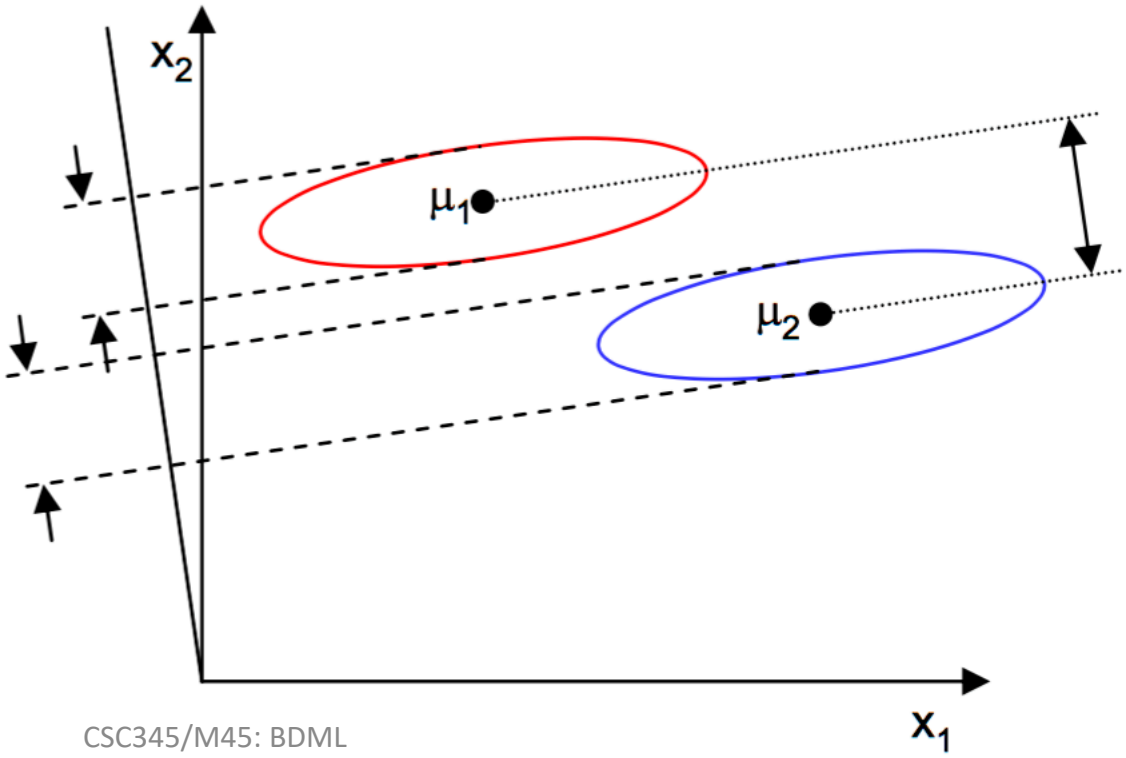


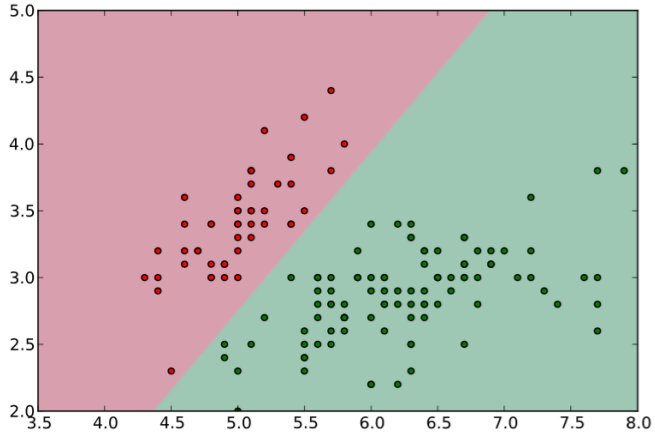
Linear Discriminant Analysis

- Fisher's LDA criterion:

$$J(w) = \frac{\left|\tilde{\mu}_1 - \tilde{\mu}_2\right|^2}{\tilde{s}_1^2 + \tilde{s}_2^2}$$

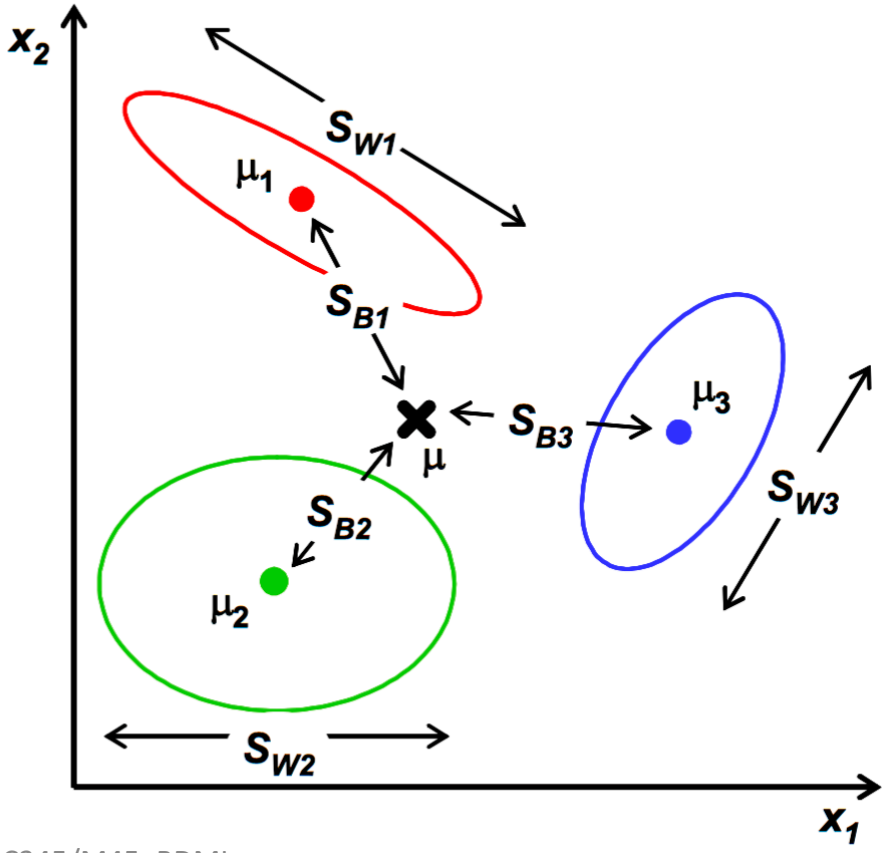
  - We are looking for a projection where samples from the same class are projected very close to each other and, at the same time, the projected means are as farther apart as possible

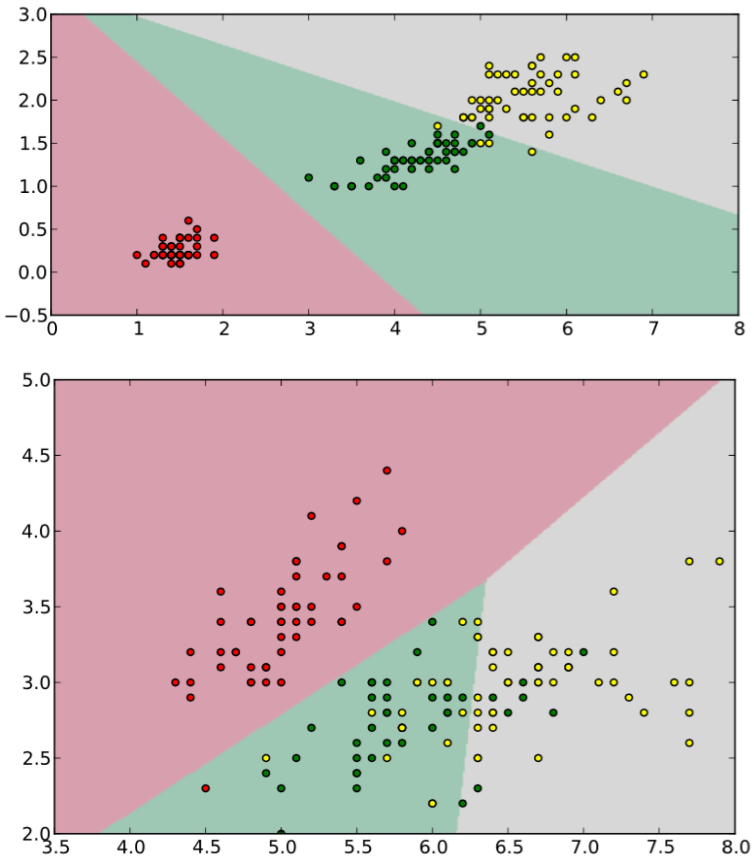




Linear Discriminant Analysis

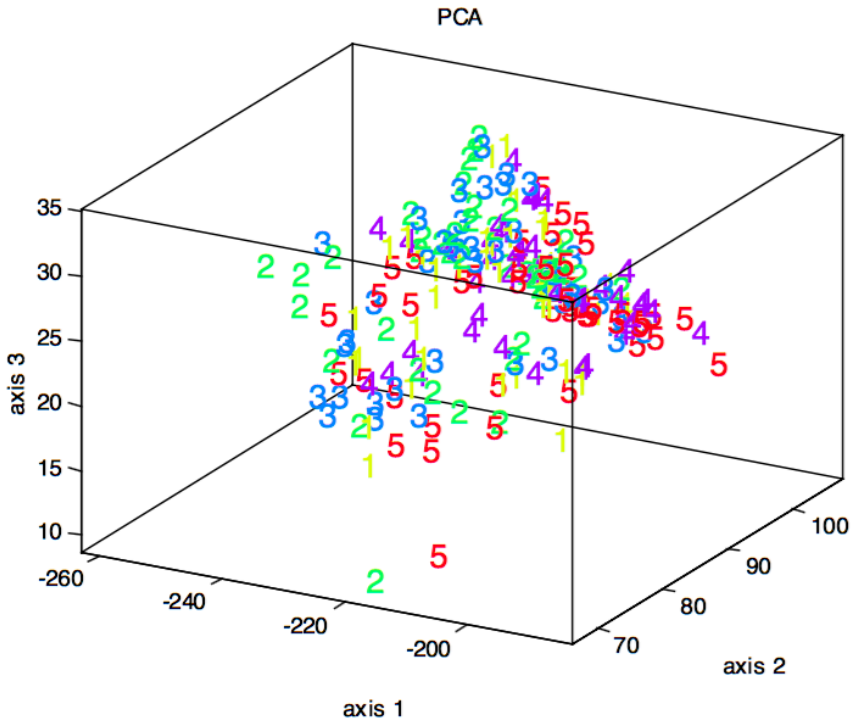
- LDA can be generalised for multi-class problems (C-class)
  - Instead of one projection, we seek C-1 projections
  - The same principle as in 2-class LDA

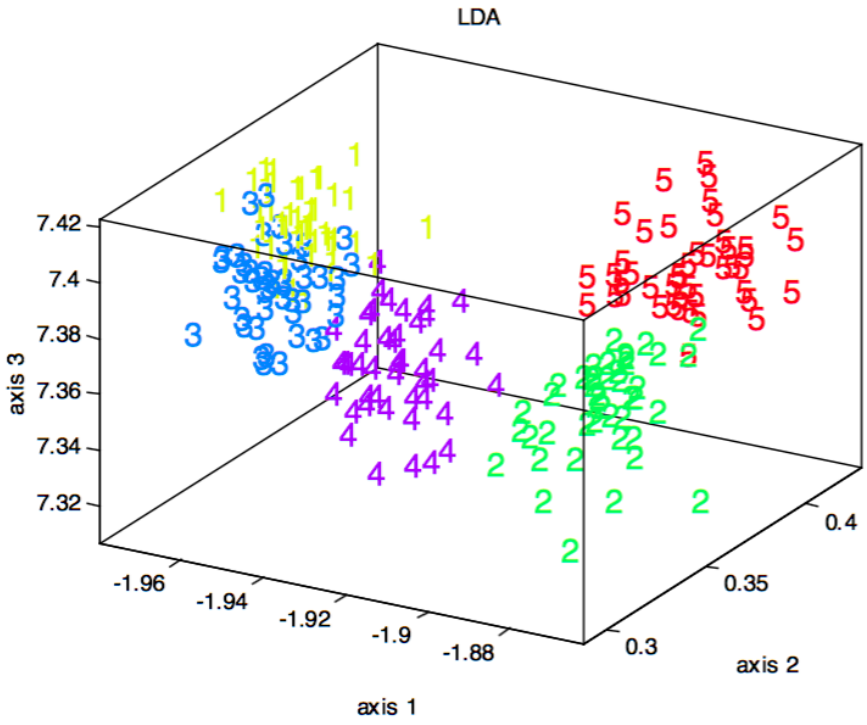




LDA vs. PCA

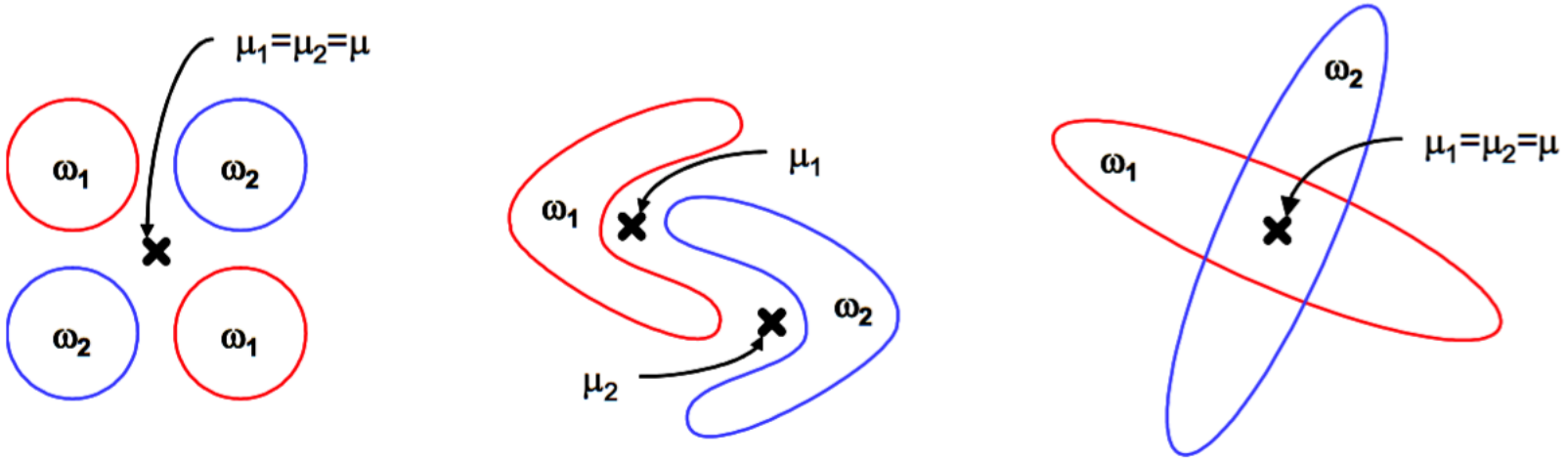
- LDA is supervised and PCA is unsupervised
  - Class information gives LDA discriminative power
    - E.g. digit recognition with dimensionality reduction





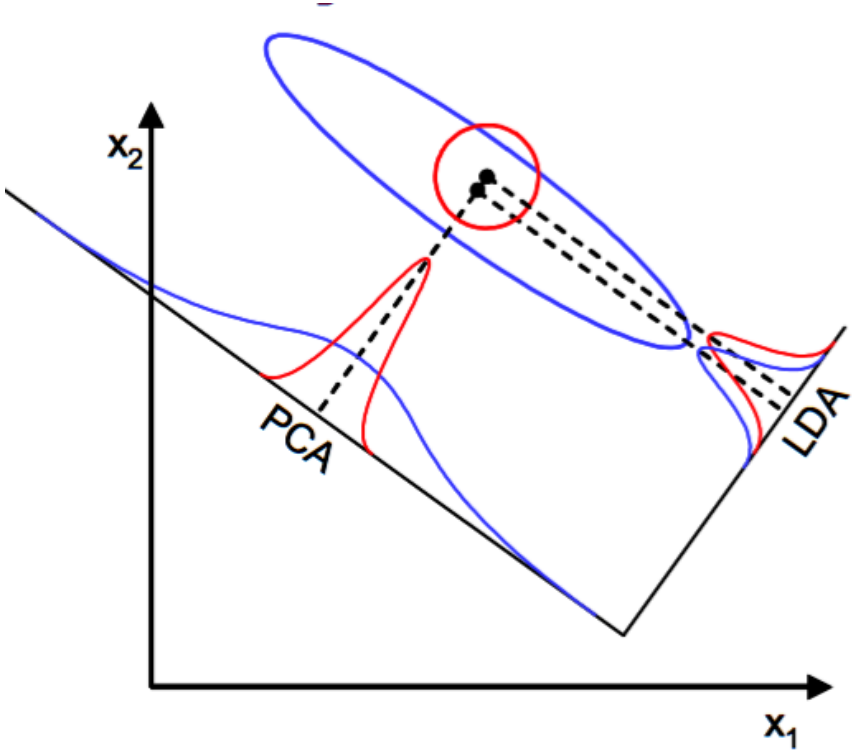
LDA Limitations

- LDA is a parametric method since it assumes uni-modal Gaussian likelihoods
  - If the distributions are significantly non-Gaussian, the LDA projections will not able to preserve any complex structure of the data, which may be needed for classification



LDA Limitations

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



Quiz

- 2 classes represented by two Gaussian functions
  - A testing sample (green) lies with an equal distance to both class centres
  - Which class should this sample belong?