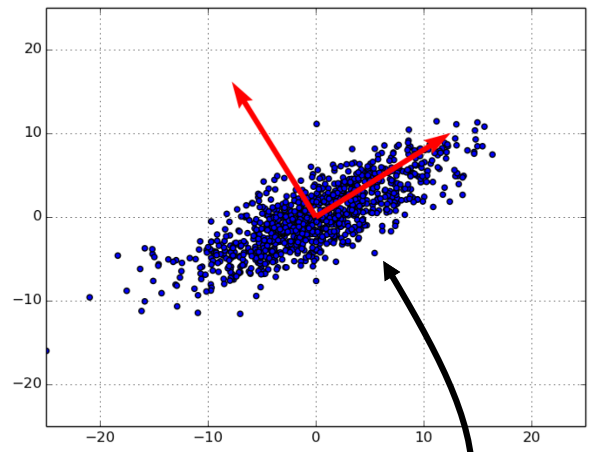
MACHINE LEARNING
MEI/1

University of Beira Interior, Department of Informatics

Hugo Pedro Proença, hugomcp@di.ubi.pt, 2019/2020

Dimensionality Reduction: PCA Summary

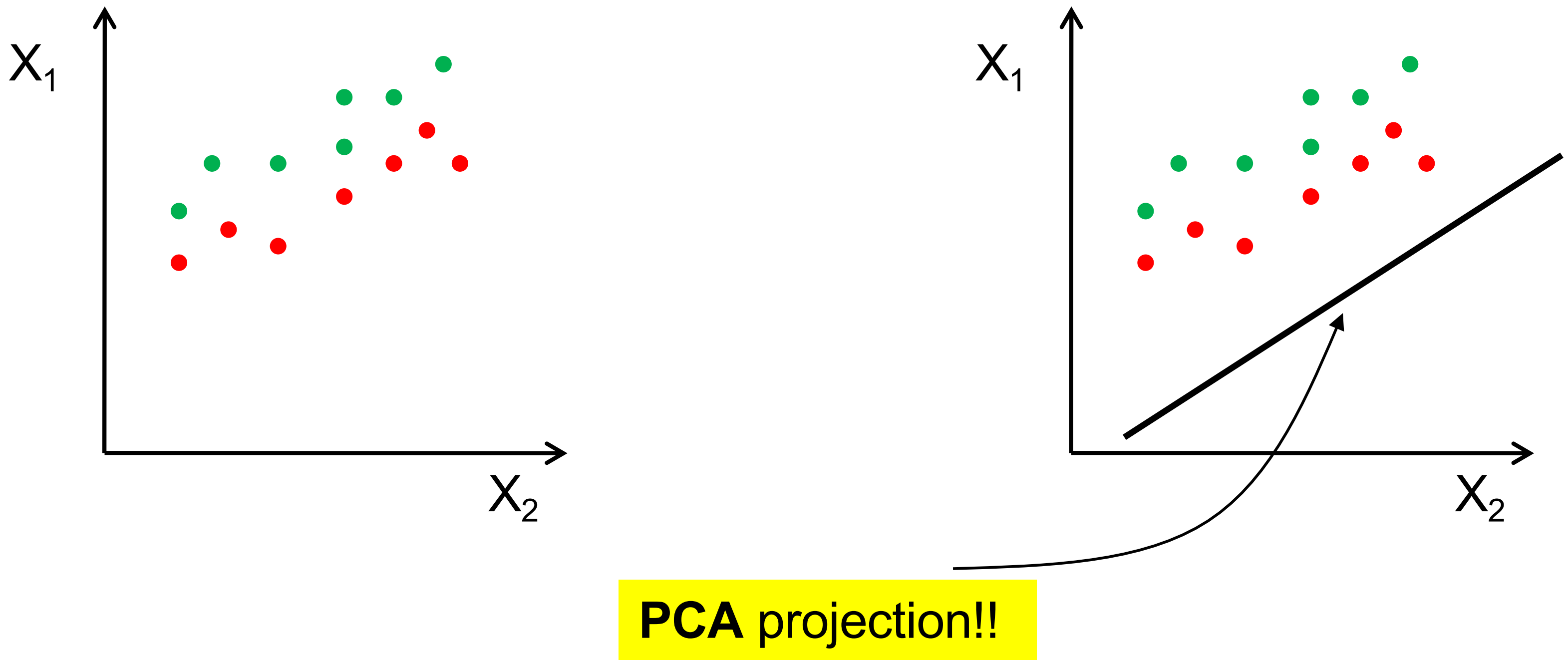
- As we saw in the last class, the idea in **PCA** is to obtain **compact representations of the data**, that keep as much as possible the amount of information.
  - Starting from an instance $x \in \mathbb{R}^n$, we obtain a $x^* \in \mathbb{R}^d$, with $d \ll n$.
- This process minimizes the effects of the **curse of dimensionality,** while also **reducing** the **computational burden** of the model inference process.
- PCA is based in the notion of **Covariance Matrix,** and in its top-d **eigenvectors**.
  - The top-d eigenvectors are those corresponding to the largest magnitude eigenvalues.
  - The set of eigenvectors is a basis of the original feature space
- The compact representations of data can be seen as the *recipes* of the original data with respect to the eigenvector, seen as the *ingredients*.
  - “This line is reconstructed by adding **0.2** $\mathbf{v}_1$, **0.1** $\mathbf{v}_2$,…



The key idea is to find the direction(s) (vector(s)) onto which data maximally **span**

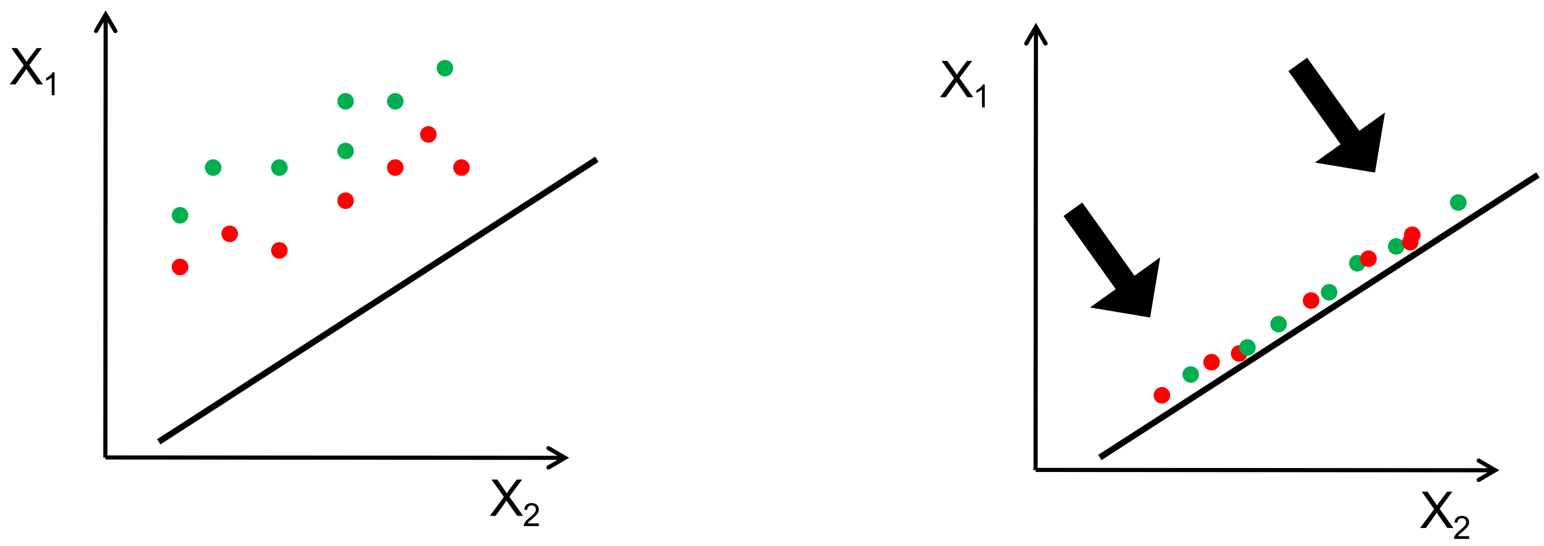
Dimensionality Reduction: PCA Summary

- However, note that **PCA** is completely agnostic with respect to the “class” information, which can be a problem in the case of supervised classification problems.



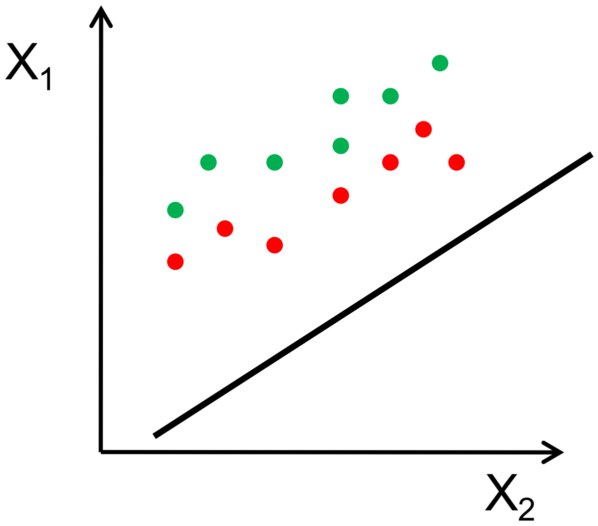
Dimensionality Reduction: PCA Summary

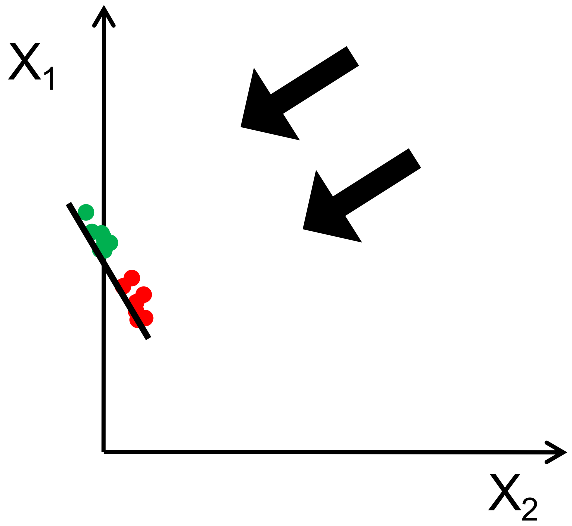
- In this example, the PCA projection will be disastrous for classification purposes.
  - Samples in the projected space will have poor separability



Dimensionality Reduction: PCA Summary

- For classification, it will be much better if the data is not projected according to the first eigenvector, but according to the second.
  - Classes have maximum separability in this representation





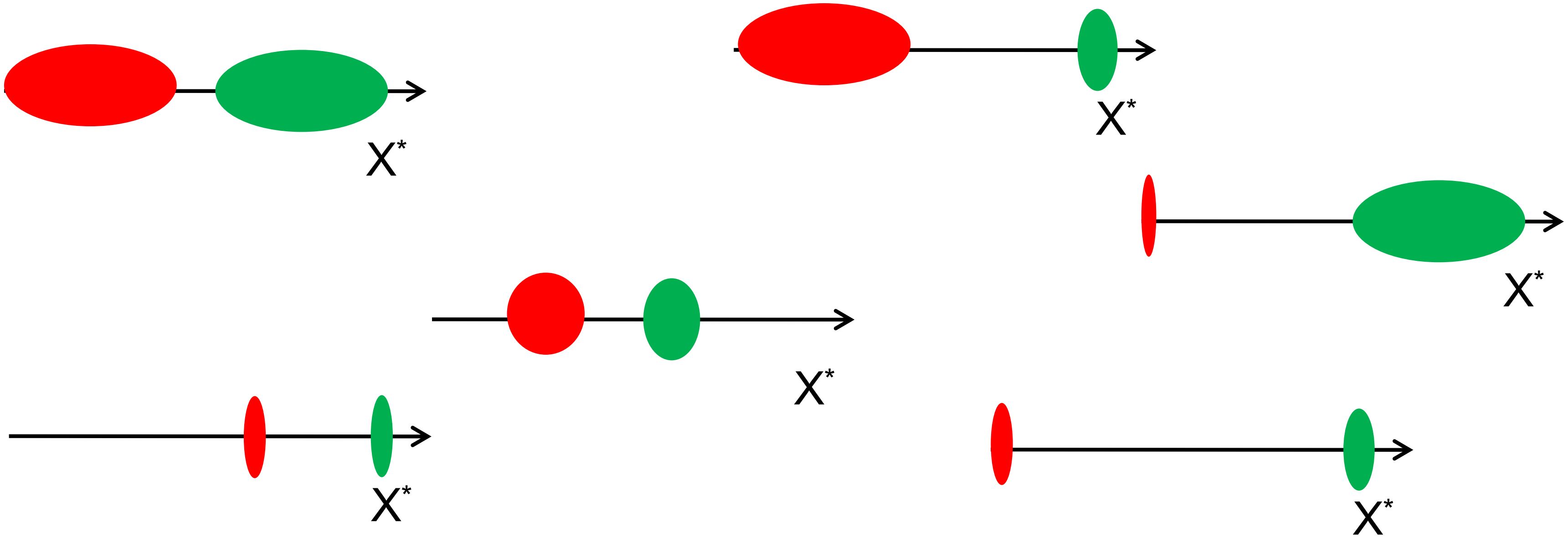
- This is exactly the goal of **LDA**: **L**inear **D**iscriminant **A**nalysis.
  - Find the data projections (i.e., compact representations) that maximize the classes separability

LDA: Linear Discriminant Analysis

- To find the best projection vector, it is importante to define a "goodness measure" for each one, so that na optimization process will enable to obtain the desired result.
    - In practice terms, this is what we have been referring to as the "**cost (objective) function**".
    - Most times, the term "cost" is used when referring to minimization problems, and "objective" is seen in maximization problems, with the semantics associated to both terms being the same.
- We are interested in obtaining a function that produces **high scores** when classes are clearly separable, and **low scores** in other cases.

LDA: Linear Discriminant Analysis

- Also, when deciding between "good" projection vectors, the best ones will be those that maximize "**inter-classes separability**", while minimizing "**intra-class spread**".
  - This is the "**Holy Grail**" of Machine Learning.



LDA: Linear Discriminant Analysis

- The Fisher Linear Discriminant was developed by Sir Ronald Fisher in 1936, and it attempts to determine whether a set of independent variables is effective in predicting the value of a discrete dependent variable (class).
- It defines a ratio between the distance between classes centroids and the sum of variances in each class.

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

where $\mu^*{}_i$ denotes the centroid of class “i”, and $s^*{}_i{}^2$ is the corresponding variance. (w) is the projection to be evaluated.

  - These are obtained in the projected space

LDA: Linear Discriminant Analysis

- The centroid in the original space is given by:

$$\mu = \frac{1}{N} \sum_{i \in Cj} x$$

- The centroid in the projected space is given by:

$$\mu^* = \frac{1}{N} \sum_{i \in \omega_j} \omega^T x$$

or equivalently:

$$\mu^* = \omega^T \mu$$

- Hence, the Fisher discriminant function can be given by:

$$J(w) = \frac{\omega^T(\mu_1 - \mu_2)}{{s^*}_1^2 + {s^*}_2^2}$$

LDA: Linear Discriminant Analysis

- Similarly, the sample variance in the original space is given by:

$$s_j = \sum_{i \in Cj} (x - \mu_j)(x - \mu_j)^T$$

- And the corresponding value in the projected space:

$$s^*{}_j = \sum_{i \in Cj} (\omega^T x - \omega^T \mu_j)^2$$

or equivalently:

$$s^*{}_j = \sum_{i \in Cj} \omega^T (x - \mu_j)(x - \mu_j)^T \omega$$

$$s^*{}_j = \omega^T s_j{}^T \omega$$

LDA: Linear Discriminant Analysis

- After Algebraic manipulation, using $s_w = s_1 + s_2$ and $s_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$, we have:

$$J(w) = \frac{\omega^T s_B \omega}{\omega^T s_w \omega}$$

  - s_w and s_B are typically designated as the **within and between scatter matrices**.

- We are now interested in finding the "w" that maximizes $J(w)$. As in the previous cases, the solution is given by obtaining the derivative with respect to ω, and find its zeros.

$$\frac{d}{d\omega} J(\omega) = 0$$

LDA: Linear Discriminant Analysis

- From previous Algebra courses, we know that:

$$\left(\frac{a}{b}\right)' = \frac{a'b \; - b'a}{b^2}$$

which is equivalent to:

$$\omega^T s_W \omega \frac{d}{d\omega} \quad \omega^T s_B \omega \; - \omega^T s_B \omega \frac{d}{d\omega} \omega^T s_W \omega = 0$$

$$(\omega^T s_W \omega) 2 s_B \omega \; - (\omega^T s_B \omega) 2 s_W \omega \qquad = 0$$

$$s_B \omega \; - \mathrm{J}(\omega) s_W \omega = 0$$

$$s_w{}^{-1} s_B \omega \; - \mathrm{J}(\omega) \omega = 0$$

- This is called a **generalized eigenvalue problem:**

$$\boldsymbol{A}\, v = \lambda\, v$$

- And the solution is given (as seen in the previous class) by the first eigenvector of "**A**", i.e., the eigenvector with the largest eigenvalue.

LDA: Linear Discriminant Analysis

- Summary of the LDA algorithm:

1. Obtain both classes means (original space)
2. Obtain both covariance matrices (original space)
3. Obtain the within class scatter matrix s_w
4. Obtain the within class scatter matrix s_B
5. Obtain the eigenvectors of $s_w^{-1} s_B$
6. Get the eigenvector corresponding to the eigenvalue with the largest magnitude. (ω)
7. Project the data in the original space (x) into the LDA space by

$$\boldsymbol{x}^{*} = \boldsymbol{\omega}^{T} \boldsymbol{x}$$

Machine Learning: LDA Exercise

- Consider the "AR.csv" dataset, available at the course web page.
  - It contains a "csv" representation of [48 x 64] face images
  - We will use it to distinguish between "Male" and "Female" genders
  - In each line, the last column gives the corresponding class (1="Male"; 0="Female")
- Implement a "Python" script that finds the LDA projection, i.e., the 1D representation of the feature space that maximizes the separability between the classes ("*man*" and "*woman*").