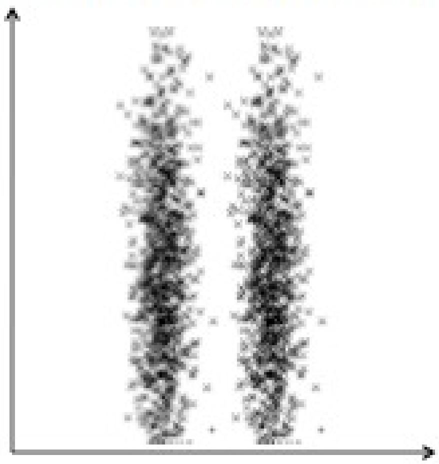
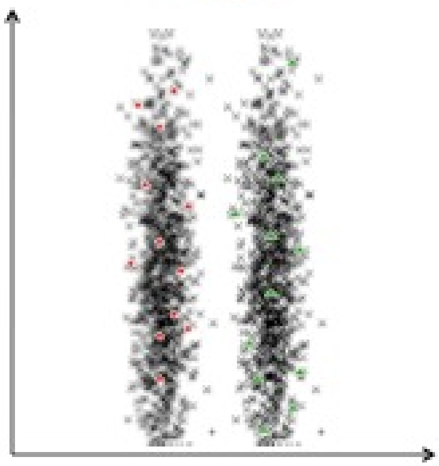
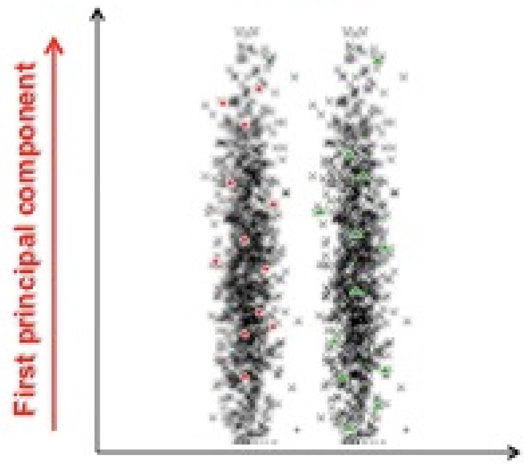
Example: PCA does not always work



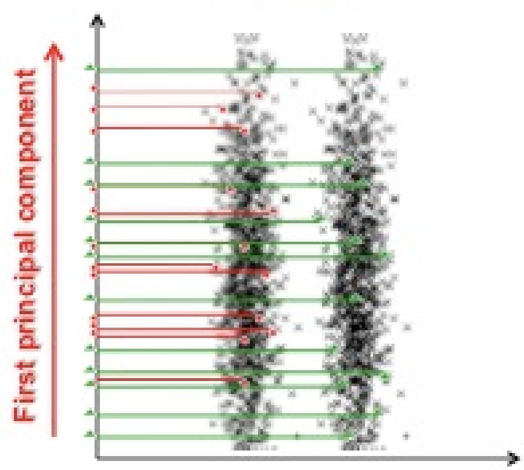
Example

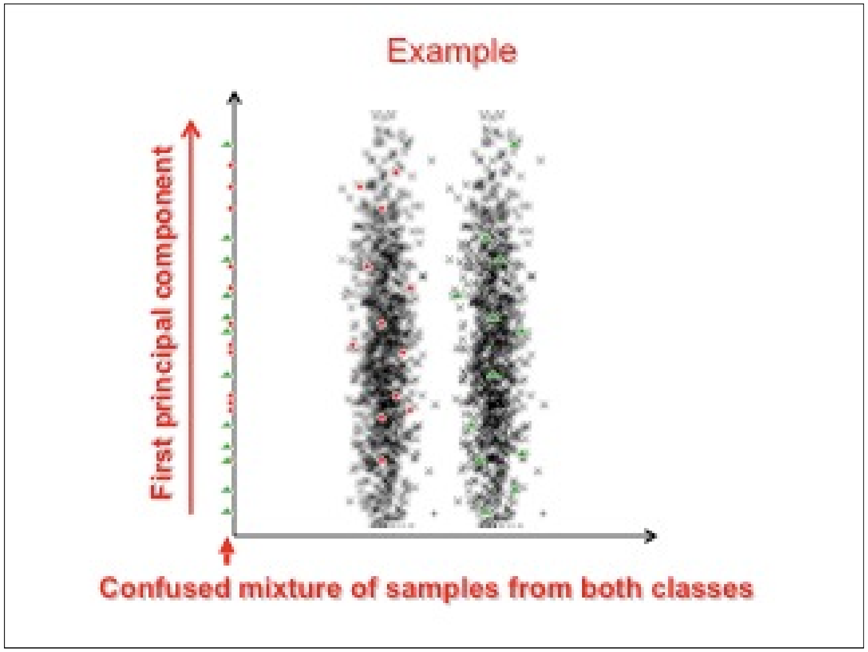


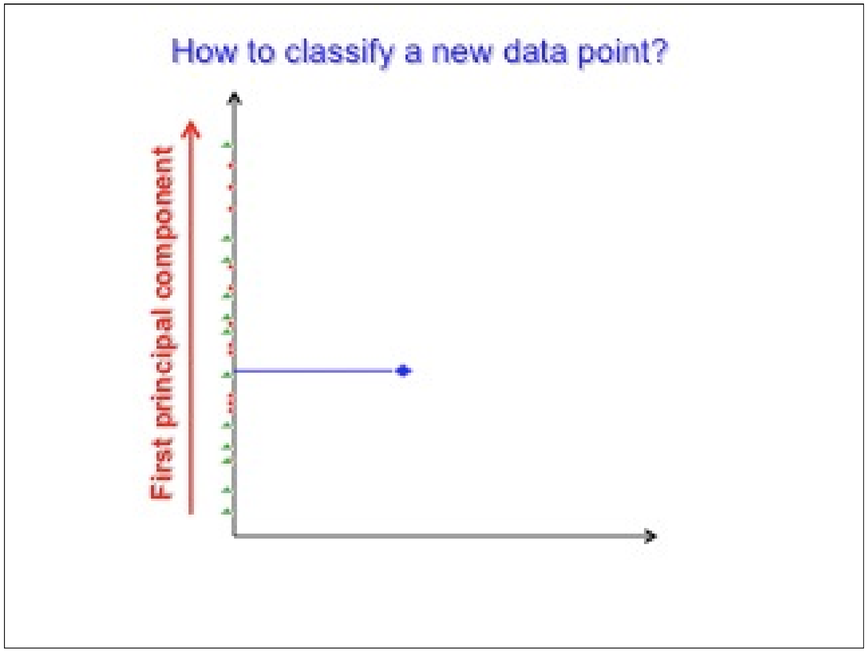
Example



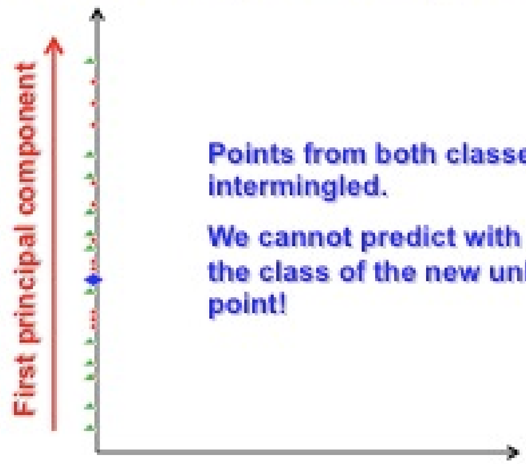
Example




Example
First principal component
Confused mixture of samples from both classes


How to classify a new data point?
First principal component

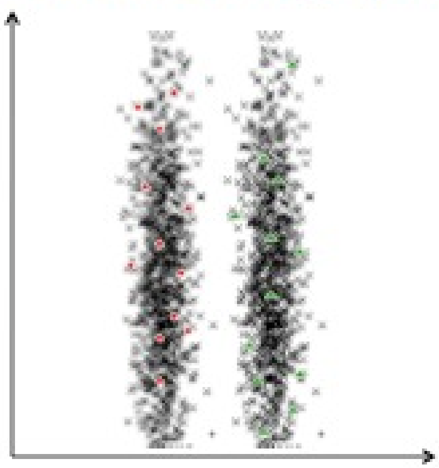
How to classify a new data point?



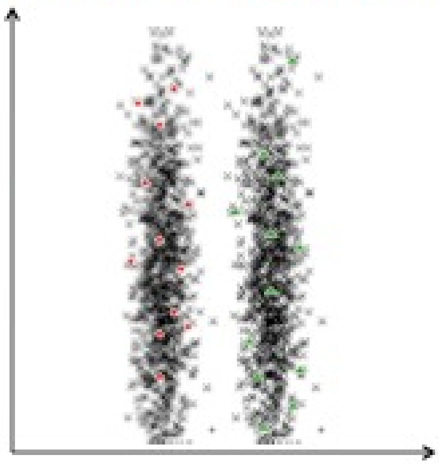
Points from both classes are intermingled.

We cannot predict with accuracy the class of the new unlabeled point!

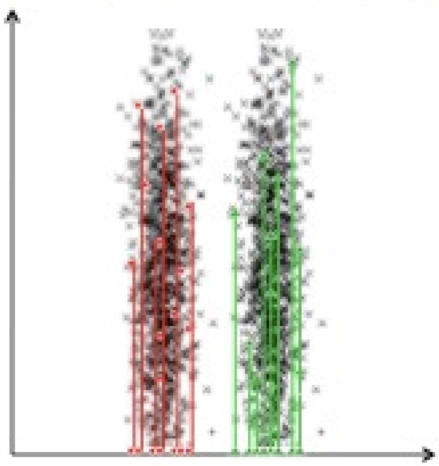
Can we find a better projection?



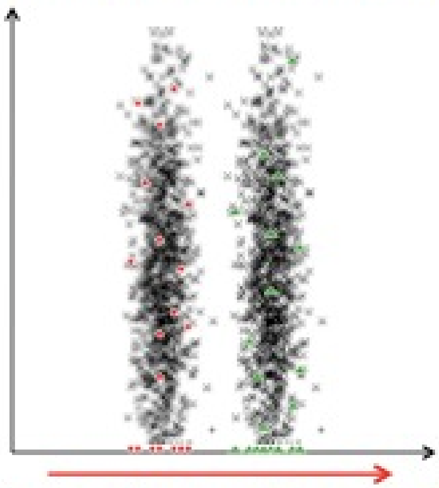
How about the horizontal dimension?



How about the horizontal dimension?

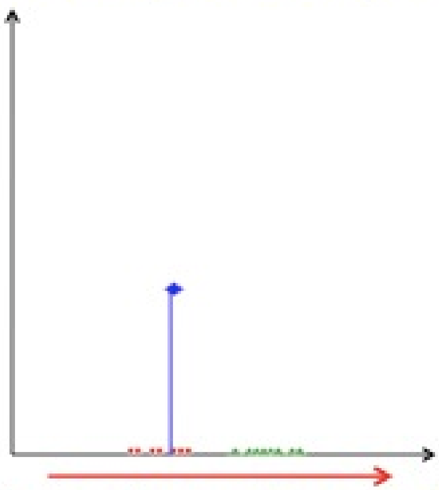


How about the horizontal dimension?

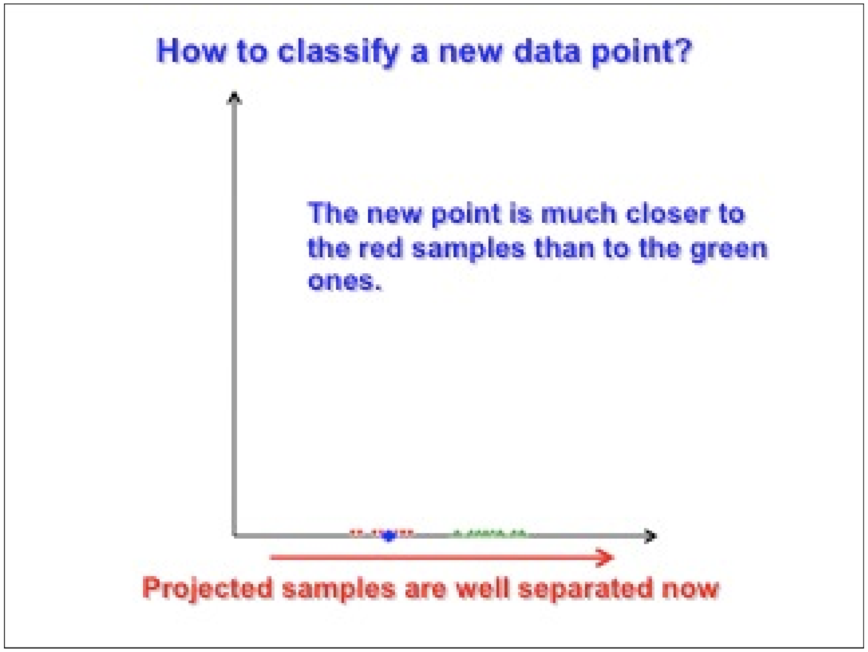


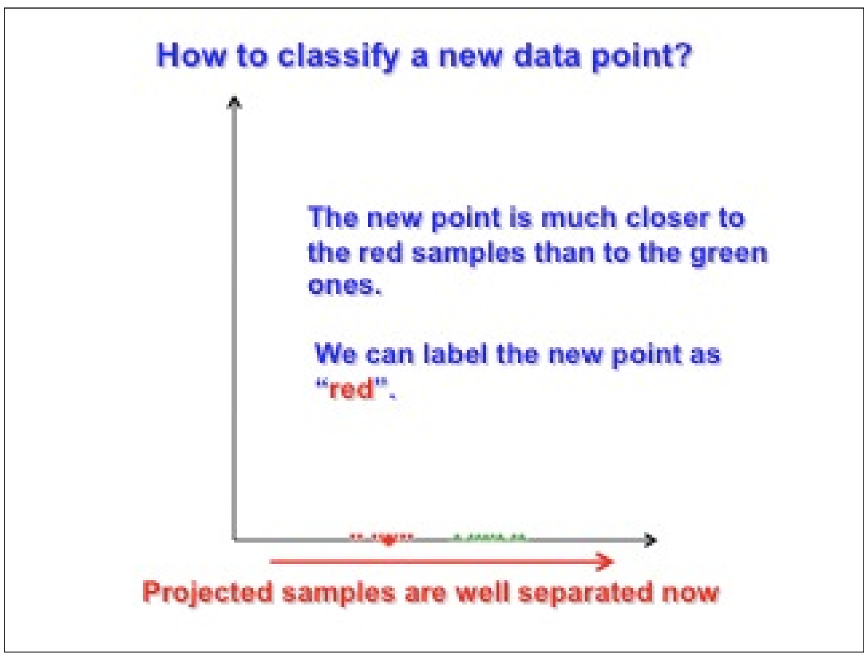
Projected samples are well separated now

How to classify a new data point?



Projected samples are well separated now


How to classify a new data point?
The new point is much closer to the red samples than to the green ones.
Projected samples are well separated now


How to classify a new data point?
The new point is much closer to the red samples than to the green ones.
We can label the new point as "red".
Projected samples are well separated now

What did we do?

- Find an orientation along which the projected samples are well separated;
- This is exactly the goal of linear discriminant analysis (LDA);
- In other words: we are after the linear projection that best separate the data, i.e. best discriminate data of different classes.

How can we find such discriminant direction?

LDA

$\{(x_n, C_i)\}_{i=1}^{N}$ $\quad x_n \in \Re^d$ $\quad C_i \in \{C_1, C_2\}$

- N_1 samples of class C_1
- N_2 samples of class C_2
- Consider $w \in \Re^d$ with $\|w\| = 1$
- Then: $w^T x$ is the projection of x along the direction of w
- We want the projections $w^T x$ where $x \in C_1$ separated from the projections $w^T x$ where $x \in C_2$

LDA

- A measure of the separation between the projected points is the difference of the sample means:

$$m_i = \frac{1}{N_i} \sum_{x \in C_i} x$$ Sample mean of class C_i

$$m_i = \frac{1}{N_i} \sum_{x \in C_i} w^T x = w^T m_i$$ Sample mean for the projected points

$$\Rightarrow \quad |m_1 - m_2| = |w^T (m_1 - m_2)|$$

We wish to make the above difference as large as we can. In addition…

LDA

- To obtain good separation of the projected data we really want the difference between the means to be large relative to some measure of the standard deviation of each class:

$$s_i^2 = \sum_{x \in C_i} \left(w^T x - m_i\right)^2$$ Scatter for the projected samples of class C_i

$$s_1^2 + s_2^2$$ Total **within-class scatter** of the projected samples

$$\arg\max_w \frac{|m_1 - m_2|^2}{s_1^2 + s_2^2}$$ **Fisher linear discriminant analysis**

LDA

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

To obtain $J(w)$ as an explicit function of w we define the following matrices :

$$S_i = \sum_{x \in X_i} (x - m_i)(x - m_i)^T$$

$S_W = S_1 + S_2$ **Within-class scatter matrix**

Then :

$$s_i^2 = \sum_{x \in X_i} \left(w^T x - m_i\right)^2 = \sum_{x \in X_i} \left(w^T x - w^T m_i\right)^2 =$$

$$\sum_{x \in X_i} w^T (x - m_i)(x - m_i)^T w = w^T S_i w$$

LDA

So: $s_1^2 = w^T S_1 w$ and $s_2^2 = w^T S_2 w$

Thus: $s_1^2 + s_2^2 = w^T S_1 w + w^T S_2 w =$

$$w^T (S_1 + S_2) w = w^T S_W w$$

Similarly:

$$(m_1 - m_2)^2 = \left(w^T m_1 - w^T m_2\right)^2 =$$

$$w^T (m_1 - m_2)(m_1 - m_2)^T w =$$

$$w^T S_B w$$

where $S_B = (m_1 - m_2)(m_1 - m_2)^T$ **Between-class scatter matrix**

LDA

We have obtained :

$s_1^2 + s_2^2 = w^T S_W w$

$(m_1 - m_2)^2 = w^T S_B w$

$$\rightarrow \quad J(w) = \frac{|m_1 - m_2|^2}{s_1^2 + s_2^2} = \frac{w^T S_B w}{w^T S_W w}$$

$$\arg \max_w \frac{w^T S_B w}{w^T S_W w}$$

LDA

It can be shown tha t a vector t hat maximizes J must satisfy :

$S_B w = \lambda S_W w \quad \Leftrightarrow \quad S_W^{-1} S_B w = \lambda w$

We observe that :

$S_B w = (m_1 - m_2) \underbrace{(m_1 - m_2)^T w}_{\text{scalar}}$

Always in the direction of $(m_1 - m_2)$

$$\rightarrow \quad \boxed{w = S_W^{-1}(m_1 - m_2)}$$

LDA

$$w = S_W^{-1}(m_1 - m_2)$$

- Gives the linear function with the maximum ratio of between-class scatter to within-class scatter.
- The problem, e.g. classification, has been reduced from a *q*-dimensional problem to a more manageable one-dimensional problem.
- Optimal for multivariate normal class conditional densities.

LDA

- The analysis can be extended to multiple classes.
- Both PCA and LDA are ***linear*** techniques for dimensionality reduction: they project the data along directions that can be expressed as ***linear combination*** of the input features.
- The "appropriate" transformation depends on the data and on the **task** we want to perform on the data. Note that LDA uses class labels, PCA does **not**.
- Non-linear extensions of PCA and LDA exist (e.g., Kernel-PCA, generalized LDA).