线性判别分析（LDA）

23:17:01

简介

线性判别分析(Linear Discriminant Analysis, LDA)，于1936年由Ronald Fisher首次提出，因此，也叫做Fisher线性判别(Fisher Linear Discriminant ,FLD），或者Fisher Discriminant Analysis （FDA），是模式识别的经典算法，并在1996年由Belhumeur引入模式识别和人工智能领域。

基本思想

将高维的模式样本投影到最佳鉴别矢量空间，以达到抽取分类信息和压缩特征空间维数的效果。

投影后保证模式样本在新的子空间有最大的类间距离和最小的类内距离，即模式在该空间中有最佳的可分离性。

因此，它是一种有效的特征抽取方法。使用这种方法能够使投影后模式样本的类间散布矩阵最大，并且同时类内散布矩阵最小。

LDA与PCA

LDA与PCA都是常用的降维技术

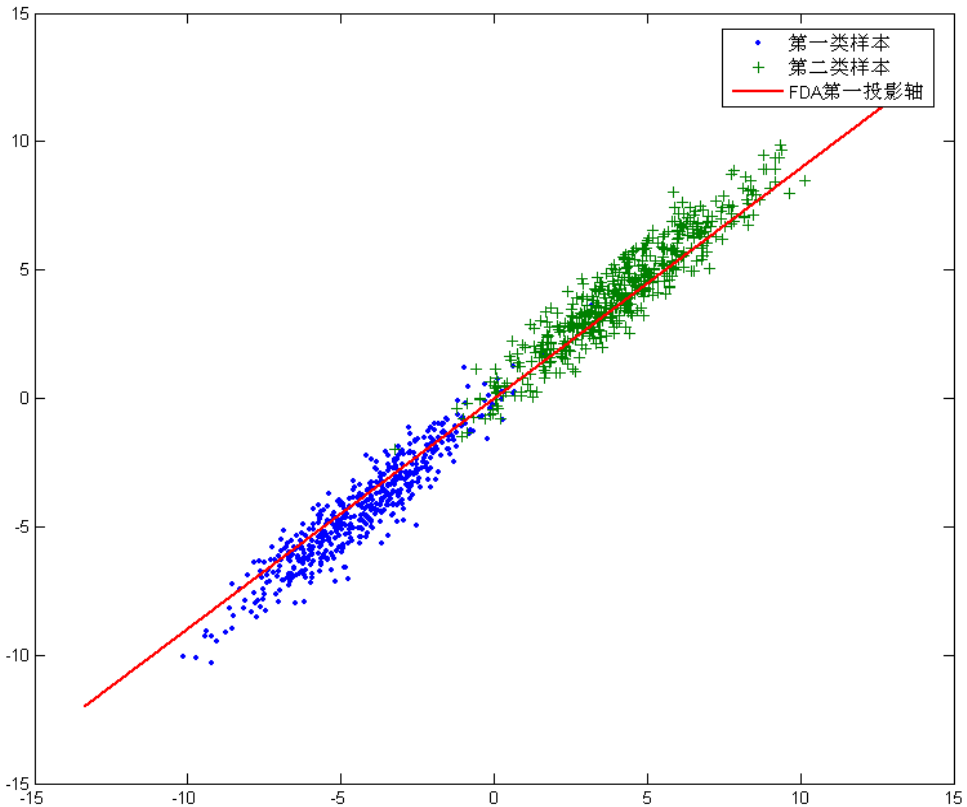
PCA是一种无监督的降维方法；

主要是从特征的协方差角度，将数据投影到方差最大的几个相互正交的方向上。

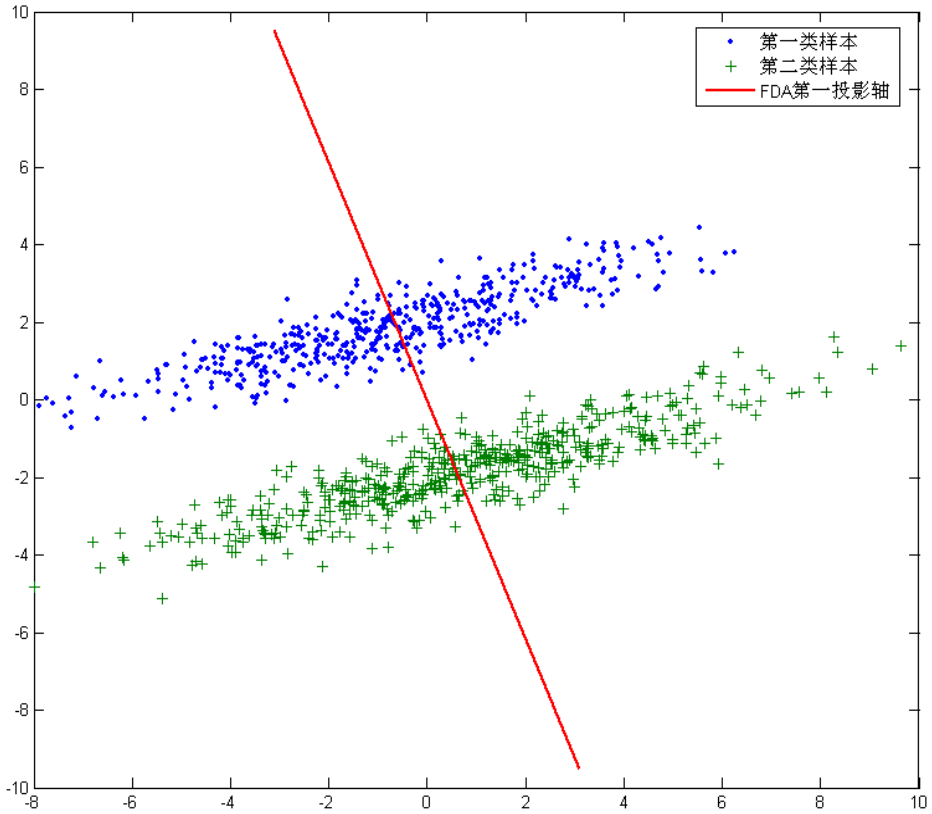
LDA是一种有监督的降维方式；

希望投影后不同类别之间数据点的距离更大，同一类别的数据点更紧凑。

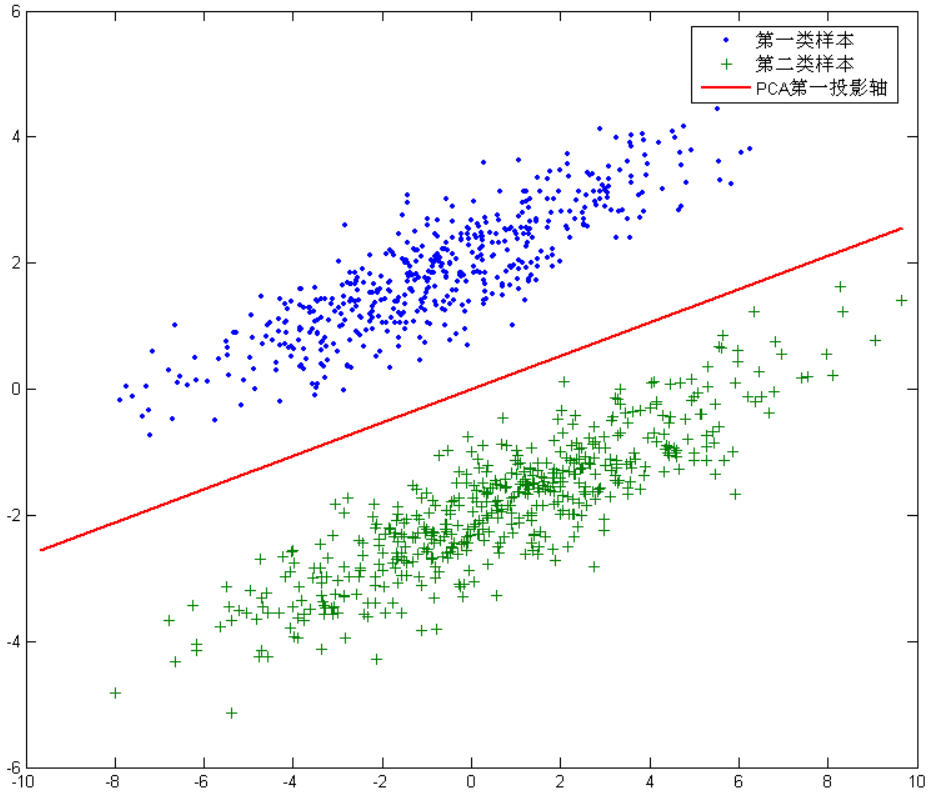
Linear discriminant analysis(LDA,FDA)



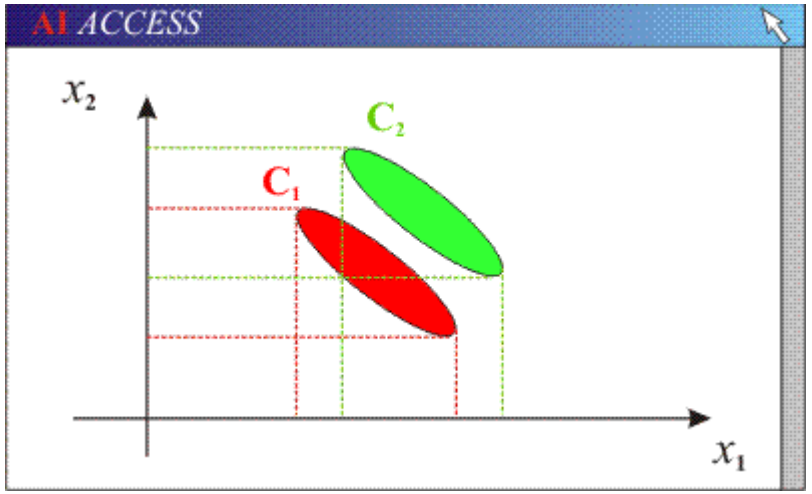
Projection axis of LDA

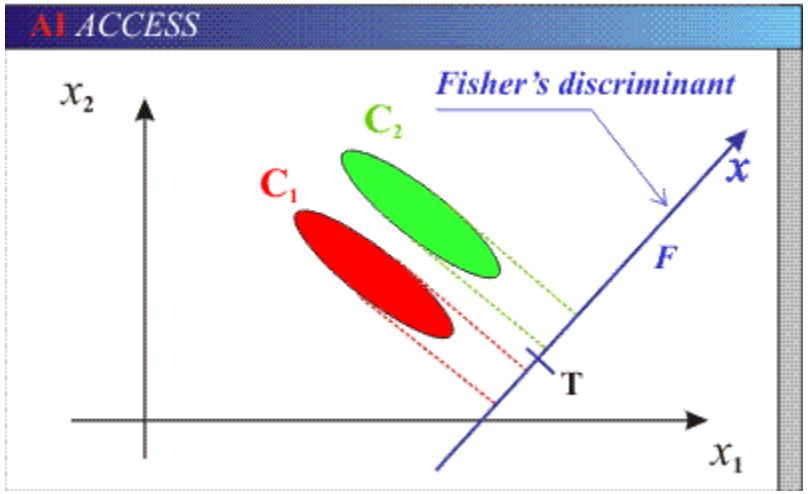


However, PCA performs badly in this case



LDA的目标：



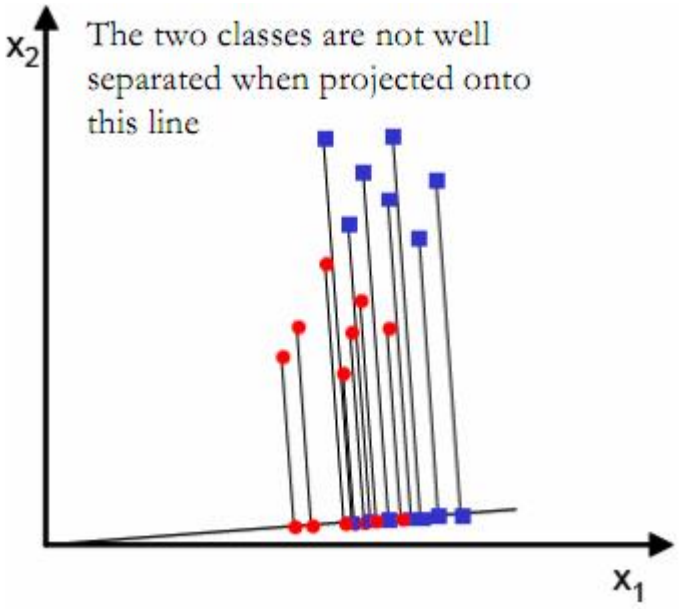


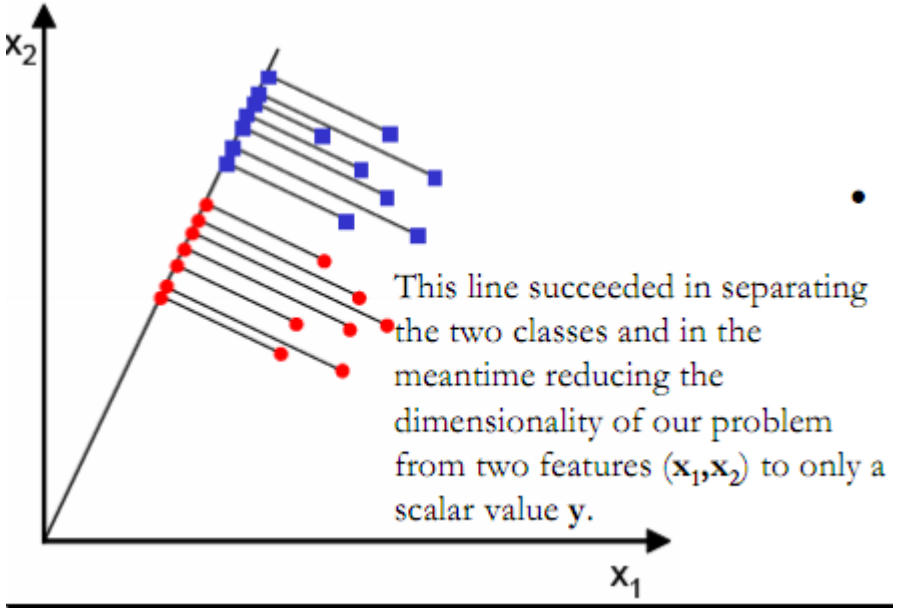
有绿色、红色两个类别。上图是两个类别的原始数据。

采用LDA方法将其映射到上图直线上

如果将数据从二维降维到一维。直接投影到x1轴或者x2轴，不同类别之间 会有重复，导致分类效果下降。

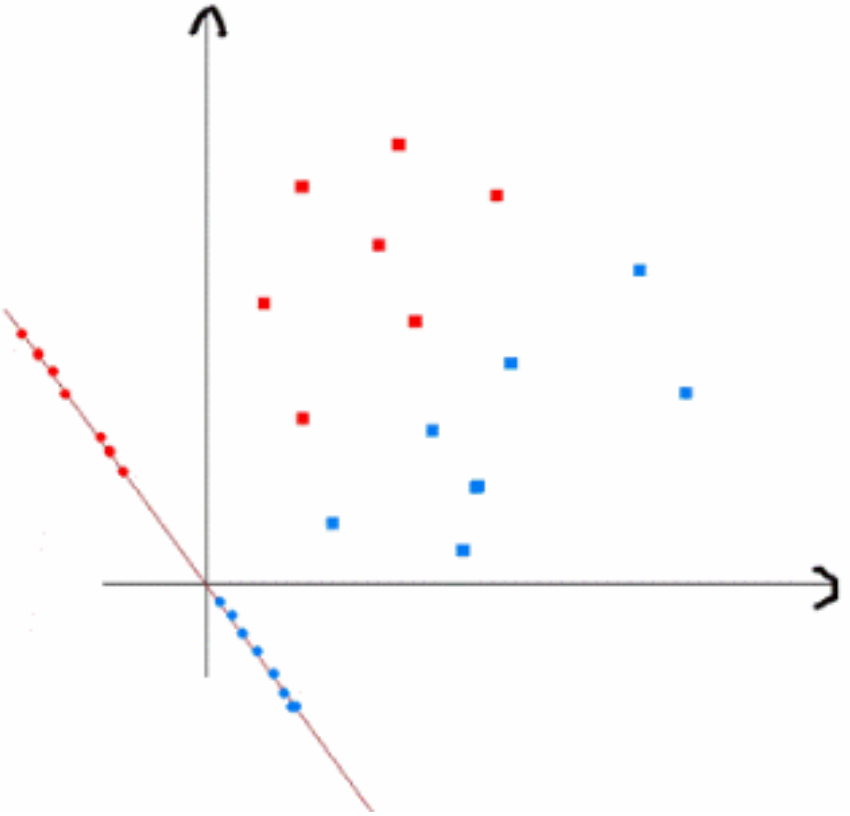
红色类别和绿色类别在映射之后之间的距离是最大的，而且每个类别内 部点的离散程度是最小的（或者说聚集程度是最大的）。


The two classes are not well separated when projected onto this line
x2
x1


x2
This line succeeded in separating the two classes and in the meantime reducing the dimensionality of our problem from two features (x1,x2) to only a scalar value y.
x1

二分类中

LDA的目标是：给出一个标注了类别的数据集，投影到了一条直线之后，能够使得点尽量的按类别区分开，当k=2即二分类问题的时候，如图所示：



$$y = w^T x$$

经过原点的直线就是投影的直线，从图上可以清楚的看到，红色的点和蓝色的点被原点明显的分开了。

用来区分二分类的直线（投影函数)为：

$$y = w^T x$$

类别i的中心点(均值)为：（Di表示属于类别i的点):

$$m_i = \frac{1}{n_i} \sum_{x \in D_i} x$$

类别i投影后的中心点为：

$$\tilde{m}_i = w^T m_i$$

衡量类别i投影后，类别点之间的分散程度（方差）为：

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

不忘初心，方得始终

LDA分类的目标是使得不同类别之间的距离越大越好，同一类别之中的距离越小越好。

可以得到如下一个公式，表示LDA投影到w后的最大化目标优化函数：

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

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

分母表示每一个类别内的方差之和，方差越大表示一个类别内的点越分散。（目标是类内投影点尽可能接近）

分子为两个类别各自的中心点的距离的平方。（目标是类间投影点尽可能远）

所以，最大化J(w)就可以求出最优的w

定义一个投影前的各类别分散程度的矩阵Si：

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

则J(w)分母化为：

$$\tilde{s}_i^2 = \sum_{x \in D_i} (w^T x - w^T m_i)^2 = \sum_{x \in D_i} w^T (x - m_i)(x - m_i)^T w = w^T S_i w$$

$$\tilde{s}_1^2 + \tilde{s}_2^2 = w^T (S_1 + S_2) w = w^T S_w w$$

S_W被称为**类内散度矩阵**

同样的将J(w)分子化为：

S_B被称为**类间散度矩阵**

$$|\widetilde{m}_1 - \widetilde{m}_2|^2 = w^T (m_1 - m_2)(m_1 - m_2)^T w = w^T S_B w$$

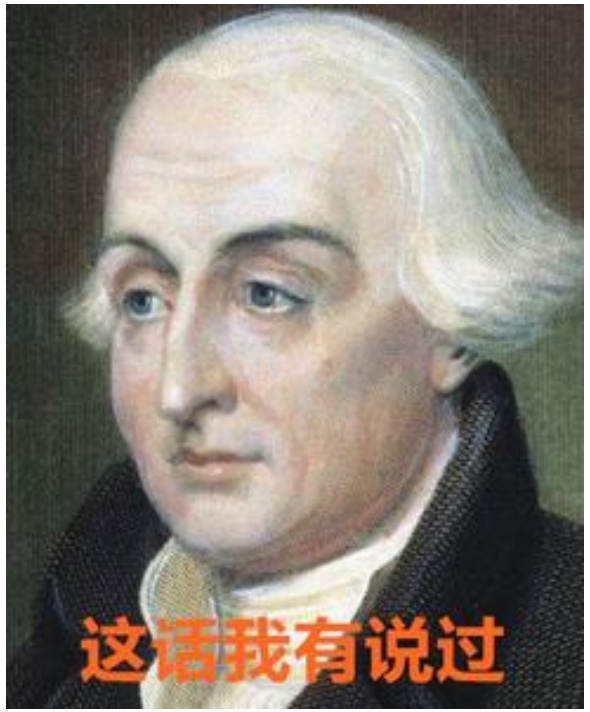
目标优化函数可以化成下面的形式：

$$J(w) = \frac{w^T S_B w}{w^T S_w w}$$

如何得到W呢？

最棒的拉格朗日乘子法来了

但是还有一个问题，如果分子、分母是都可以取任意值的，那就会使得有无穷解，不失一般性，我们将分母限制为长度为1（这是用拉格朗日乘子法一个很重要的技巧，如果忘记了，请复习一下高数），并作为拉格朗日乘子法的限制条件。



$$J(w) \equiv \frac{w^T S_B w}{w^T S_w w}$$

可得到：

$$c(w) = w^T S_B w - \lambda(w^T S_w w - 1)$$

$$\Rightarrow \frac{dc}{dw} = 2S_B w - 2\lambda S_w w = 0$$

$$\Rightarrow S_B w = \lambda S_w w$$

直接可得出

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

w就是矩阵 $S_w^{-1} S_B$ 的特征向量。

别忘啦 $$|\widetilde{m}_1 - \widetilde{m}_2|^2 = w^T(m_1 - m_2)(m_1 - m_2)^T w = w^T S_B w$$

得到 $S_B W = （m_1 - m_2）（m_1 - m_2）^T W$

$（m_1 - m_2）^T W$ 是一个标量哦

so！！！ $S_B W$ 与 $（m_1 - m_2）$ 同方向

可令 $S_B W = \lambda（m_1 - m_2）$

$$S_B W = \lambda (m_1 - m_2)$$

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

$$S_B w = \lambda S_w w$$

对于N(N>2)分类的问题，就可以直接写出以下的结论：

$$S_W = \sum_{i=1}^{c} S_i$$

$$S_B = \sum_{i=1}^{c} n_i (m_i - m)(m_i - m)^T$$

$$S_B w_i = \lambda S_w w_i$$

这同样是一个求特征值的问题，求出的第i大的特征向量，即为对应的Wi。

S_W^{-1}一直都存在吗？

S_W为奇异矩阵时可用奇异值分解得到的。

- 同样我们可以先使用PCA

1) PCA 用来降低数据的维度

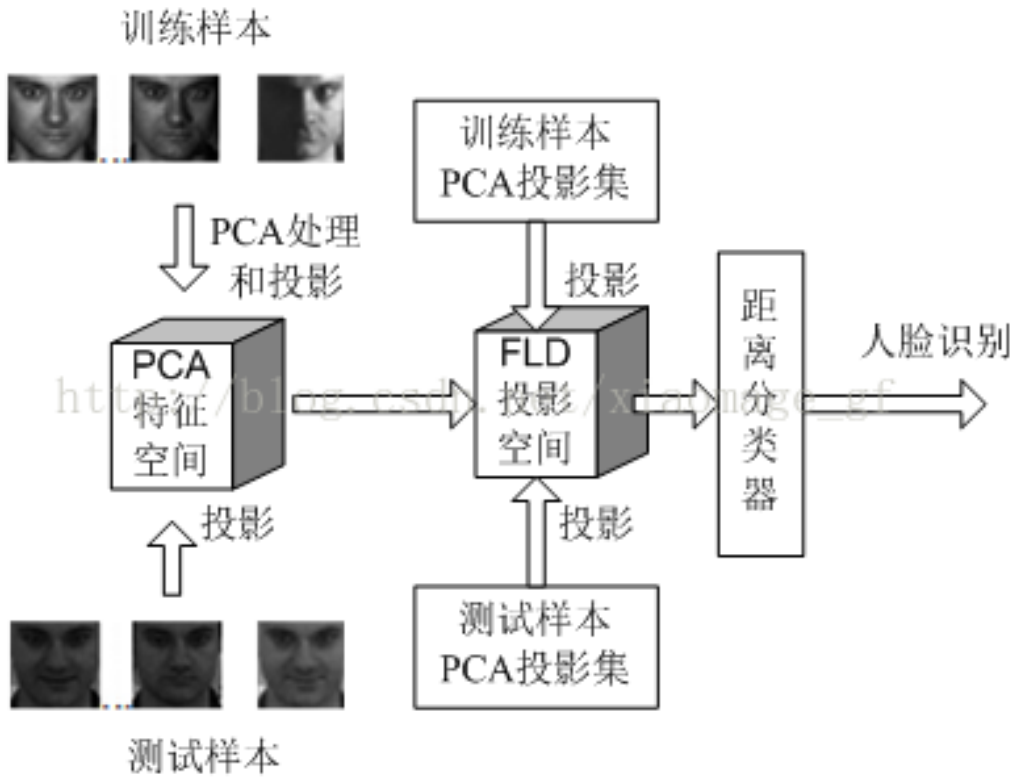
$$\begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_N \end{bmatrix} --> PCA --> \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_K \end{bmatrix}$$

2) LDA 寻找最具有判别性的方向

$$\begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_K \end{bmatrix} --> LDA --> \begin{bmatrix} z_1 \\ z_2 \\ \cdots \\ z_{C-1} \end{bmatrix}$$

LDA在人脸识别中的应用

具体的LDA人脸识别流程图如下：



LDA算法的主要优点

在降维过程中可以使用类别的先验知识经验，而像PCA这样的无监督学习则无法使用类别先验知识。

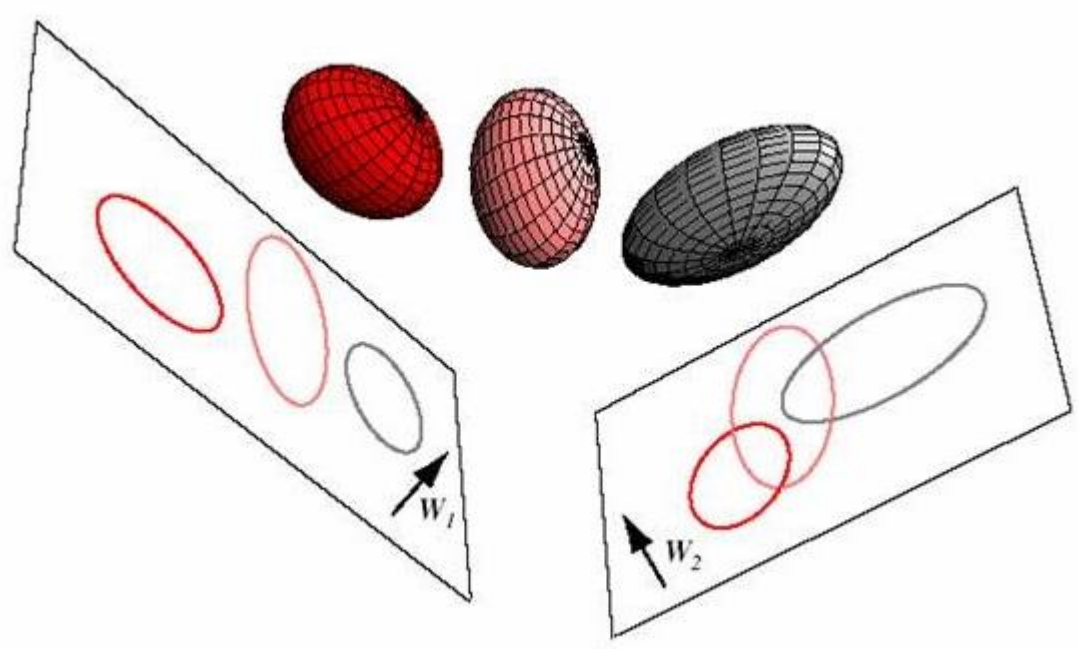
LDA算法的主要缺点

LDA不适合对非高斯分布样本进行降维，PCA也有这个问题。

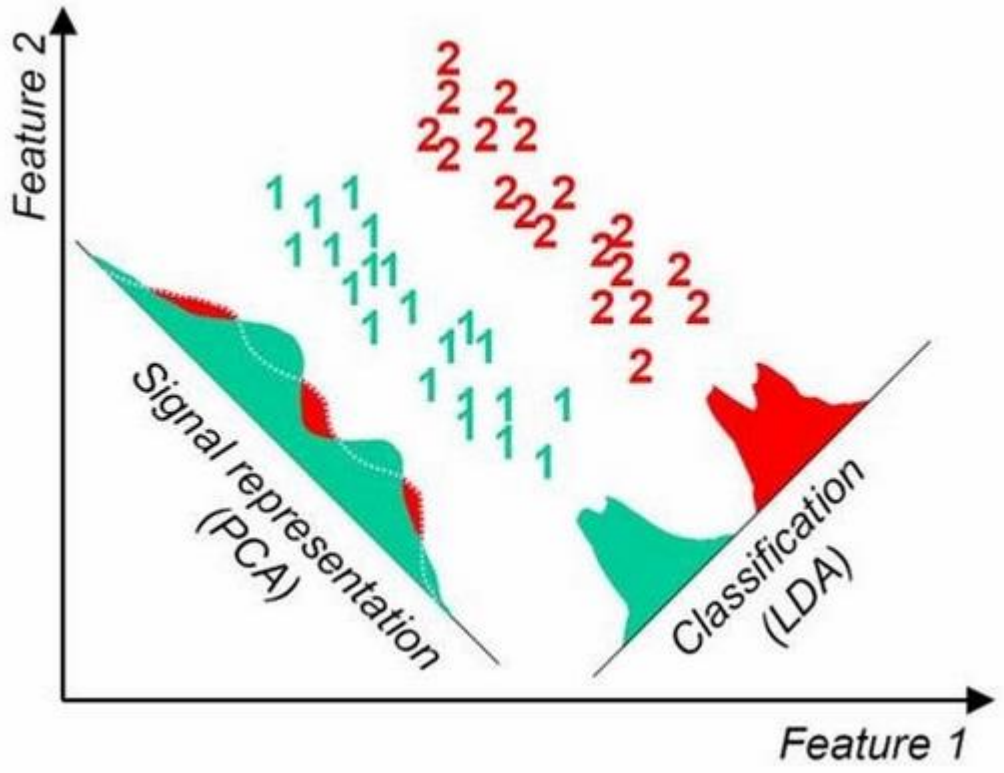
LDA在样本分类信息依赖方差而不是均值的时候，降维效果不好。

栗子

将3维空间上的球体样本点投影到二维上，W1相比W2能够获得更好的分离效果。



PCA与LDA的降维对比：



PCA选择样本点投影具有最大方差的方向，LDA选择分类性能最好的方向。