第五章

主成分分析（PCA）及LDA算法

任课教师：柳欣老师

email：starxliu@163.com

6.1 主成分原理

- 每个人都会遇到有很多变量的数据。
- 比如全国或各个地区的带有许多经济和社会变量的数据；各个学校的研究数据，生物特征数据等等。
- 这些数据的共同特点是变量很多，在如此多的变量之中，有很多是相关的。人们希望能够找出它们的少数“代表”来对它们进行描述。
- 本章就介绍两种把变量维数降低以便于描述、理解和分析的方法：主成分分析（principal component analysis）和因子分析（factor analysis）。实际上主成分分析可以说是因子分析的一个特例。在引进主成分分析之前，先看下面的例子。

成绩数据

- 100个学生的数学、物理、化学、语文、历史、英语的成绩如下表（部分）。

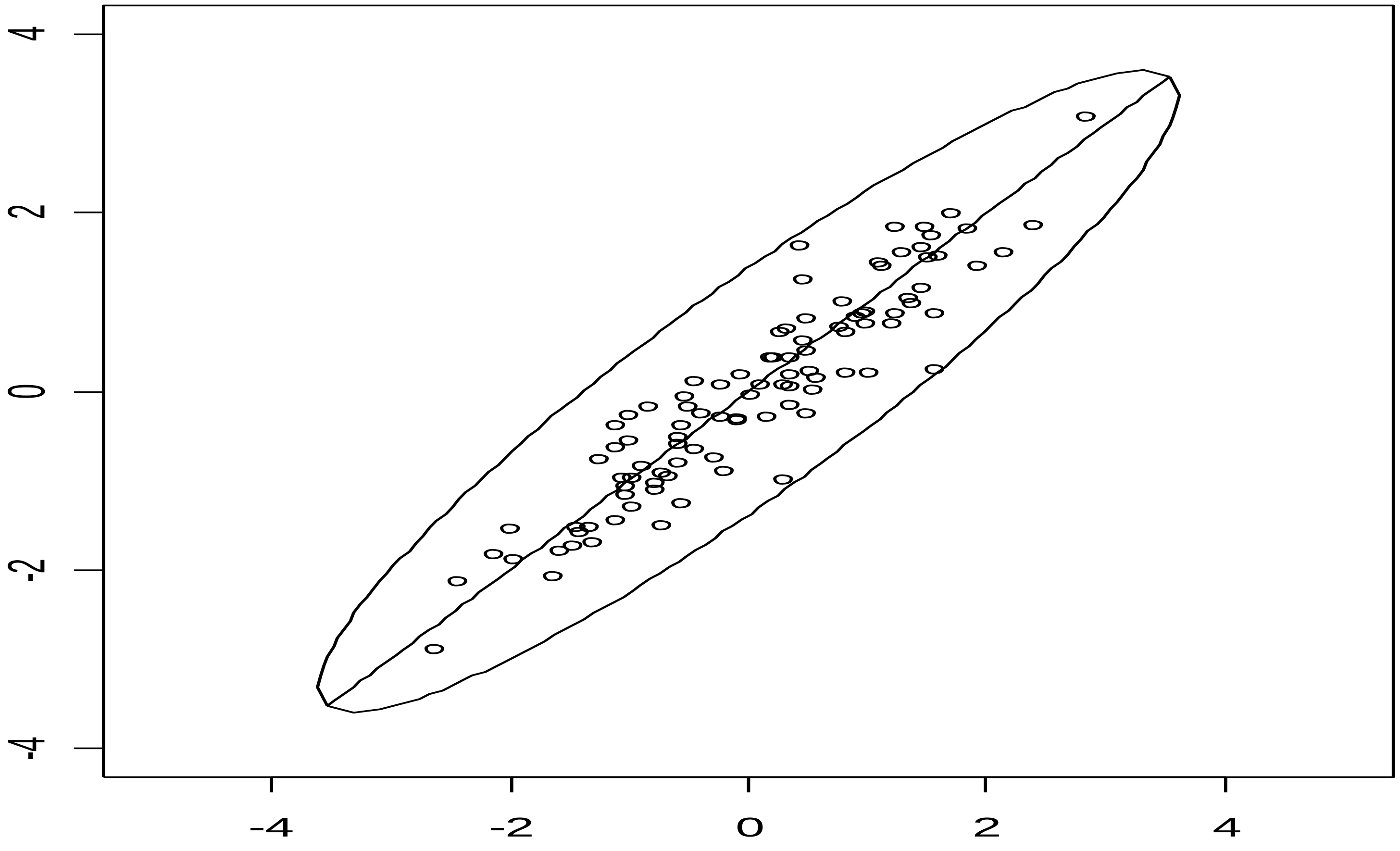
学生代码	数学	物理	化学	语文	历史	英语
1	65	61	72	84	81	79
2	77	77	76	64	70	55
3	67	63	49	65	67	57
4	80	69	75	74	74	63
5	74	70	80	84	81	74
6	78	84	75	62	71	64
7	66	71	67	52	65	57
8	77	71	57	72	86	71
9	83	100	79	41	67	50
…	…	…	…	…	…	…

从本例可能提出的问题

- 目前的问题是，能不能把这个数据的6个变量用一两个综合变量来表示呢？
- 这一两个综合变量包含有多少原来的信息呢？
- 能不能利用找到的综合变量来对学生排序呢？这一类数据所涉及的问题可以推广到对企业，对学校进行分析、排序、判别和分类等问题。

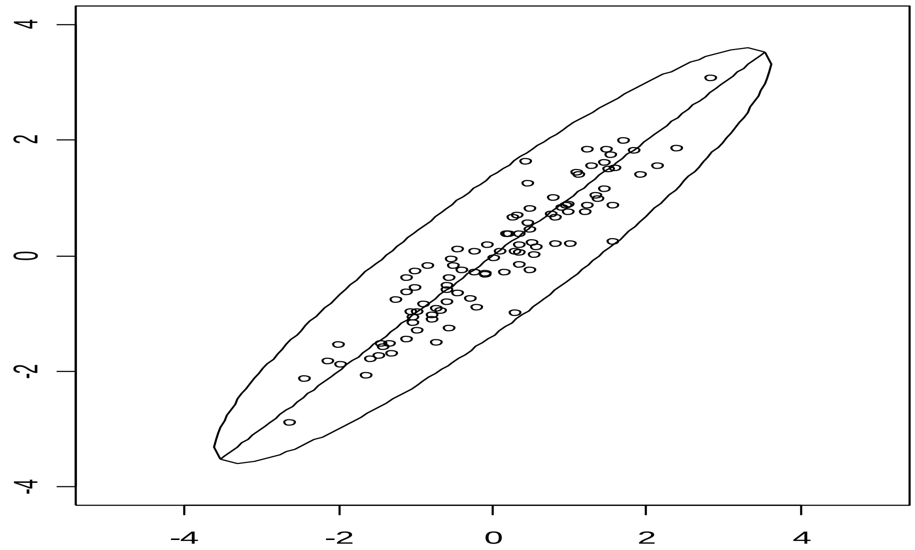
空间的点

- 例中的的数据点是六维的；也就是说，每个观测值是6维空间中的一个点。我们希望把**6维空间用低维空间表示**。
- 先假定只有二维，即只有两个变量，它们由横坐标和纵坐标所代表；因此每个观测值都有相应于这两个坐标轴的两个坐标值；如果这些数据形成一个椭圆形状的点阵（这在变量的二维正态的假定下是可能的）
- 那么这个椭圆有一个长轴和一个短轴。在短轴方向上，数据变化很少；在极端的情况，短轴如果退化成一点，那只有在长轴的方向才能够解释这些点的变化了；这样，由二维到一维的降维就自然完成了。



4
2
0
-2
-4
-4
-2
0
2
4

椭球的长短轴



- 当坐标轴和椭圆的长短轴平行，那么代表长轴的变量就描述了数据的主要变化，而代表短轴的变量就描述了数据的次要变化。
- 但是，坐标轴通常并不和椭圆的长短轴平行。因此，需要寻找椭圆的长短轴，并进行变换，使得新变量和椭圆的长短轴平行。
- 如果长轴变量代表了数据包含的大部分信息，就用该变量代替原先的两个变量（舍去次要的一维），降维就完成了。
- 椭圆（球）的长短轴相差得越大，降维也越有道理。

主轴和主成分

- 对于多维变量的情况和二维类似，也有高维的椭球，只不过无法直观地看见罢了。
- 首先把高维椭球的主轴找出来，再用代表大多数数据信息的最长的几个轴作为新变量；这样，主成分分析就基本完成了。
- 注意，和二维情况类似，高维椭球的主轴也是互相垂直的。这些互相正交的新变量是原先变量的线性组合，叫做主成分 (**principal component**)。

主成分选取

- 正如二维椭圆有两个主轴，三维椭球有三个主轴一样，有几个变量，就有几个主成分。
- 选择越少的主成分，降维就越好。什么是标准呢？那就是这些被选的主成分所代表的主轴的长度之和占了主轴长度总和的大部分。有些文献建议，所选的主轴总长度占所有主轴长度之和的大约**85%**即可，其实，这只是一个大体的说法；具体选几个，要看实际情况而定。

主成分分析原理

- **Principal Component Analysis（PCA）**
- 主成分分析**(Principal Component Analysis,** 简称**PCA)**是一种常用的基于变量协方差矩阵对信息进行处理、压缩和抽提的有效方法。

基于PCA算法的人脸识别

- **PCA**方法由于其在降维和特征提取方面的有效性，在人脸识别领域得到了广泛
  的应用。
- **PCA**方法的基本原理是:利用**K-L**变换抽取人脸的主要成分，构成特征脸空间，识别时将测试图像投影到此空间，得到一组投影系数，通过与各个人脸图像比较进行识别。
- 利用特征脸法进行人脸识别的过程由训练阶段和识别阶段两个阶段组成
- 其具体步骤如下：

训练阶段

- 第一步：假设训练集有**200**个样本，由灰度图组成，每个样本大小为**M*N**

- 写出训练样本矩阵：

$$x = \left(x_1, x_2, \ldots, x_{200}\right)^T$$

- 其中向量**x**$_i$为由第**i**个图像的每一列向量堆叠成一列的**MN**维列向量,即把矩阵向量化,如下图所示：

训练阶段

- 如：第i个图像矩阵为

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

- 则x_i为

$$\begin{bmatrix} 1 \\ 4 \\ 7 \\ 2 \\ 5 \\ 8 \\ 3 \\ 6 \\ 9 \end{bmatrix}$$

训练阶段

- 第二步：计算平均脸

计算训练图片的平均脸：

$$\Psi = \frac{1}{200}\sum_{i=1}^{i=200} x_i$$

- 第三步：计算差值脸

计算每一张人脸与平均脸的差值

$$d_i = x_i - \Psi, i = 1,2,...,200$$

训练阶段

- 第四步：构建协方差矩阵

$$C=\frac{1}{200}\sum_{i=1}^{200}d_i d_i^{\,T}=\frac{1}{200}AA^T$$

$$A=\left(d_1,d_2,\ldots,d_{200}\right)$$

训练阶段

- 第五步：求协方差矩阵的特征值和特征向量，构造特征脸空间

协方差矩阵的维数为**MN*MN**，考虑其维数较大，计算量比较大，所以采用奇异值分解**(SingularValue Decomposition ,SVD)**定理，通过求解$A^T A$的特征值和特征向量来获得AA^T的特征值和特征向量。

v_i

训练阶段

- 求出$A^T A$的特征值 λ_i及其正交归一化特征向量 v_i
- 根据特征值的贡献率选取前p个最大特征值及其对应的特征向量
- 贡献率是指选取的特征值的和与占所有特征值的和比，即：

$$\varphi = \frac{\sum_{i=1}^{i=p} \lambda_i}{\sum_{i=1}^{i=200} \lambda_i} \geq a$$

训练阶段

- 一般取 $a = 99\%$ 即使训练样本在前**p**个特征向量集上的投影有**99%**的能量

求出原协方差矩阵的特征向量

$$u_i = \frac{1}{\sqrt{\lambda_i}} A v_i \ (i = 1,2,\ldots,p)$$

则“特征脸”空间为：

$$w = (u_1, u_2, \ldots, \ \ u_p)$$

训练阶段

- 第六步
- 将每一幅人脸与平均脸的差值脸矢量投影到“特征脸”空间，即

$$\Omega_i = w^T d_i \left(i = 1,2,\ldots,200 \right)$$

识别阶段

- 第一步：将待识别的人脸图像 Γ与平均脸的差值脸投影到特征脸空间，得到其特征向量表示：

$$\Omega^{\Gamma} = w^{T}(\Gamma - \Psi)$$

- 第二部：采用欧式距离来计算 Ω^{Γ}与每个人脸的距离 ε_i

$$\varepsilon_i^{\ 2} = \left\| \Omega_i - \Omega^{\Gamma} \right\|^2 (i = 1,2,\ldots,200)$$

1：其他距离形式，核空间距离情况。
2：一般情况下，求最小值对应的训练集合中的标签号作为识别结果

6.2 线性判别分析（LDA）

线性判别分析

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

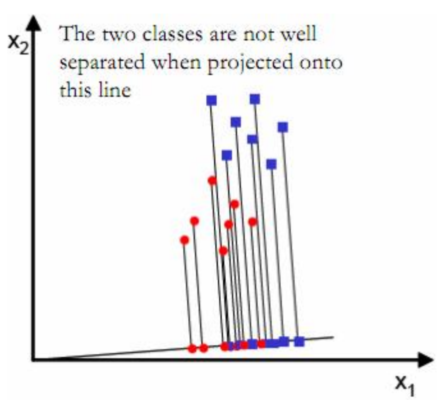
LDA与PCA(主成分分析)都是常用的降维技术。PCA主要是从特征的协方差角度，去找到比较好的投影方式。LDA更多的是考虑了标注，即希望投影后不同类别之间数据点的距离更大，同一类别的数据点更紧凑。

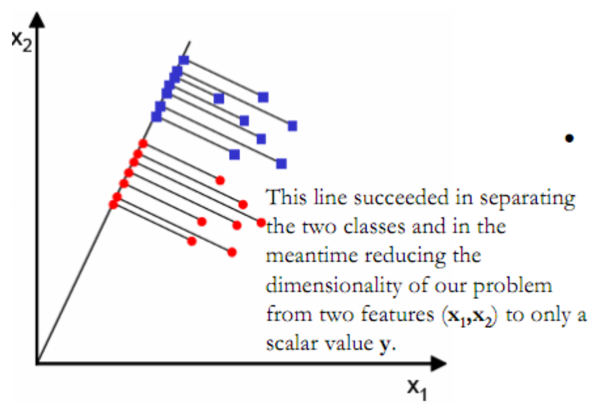
线性判别分析（LDA）

基本思想

线性判别分析的基本思想是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。

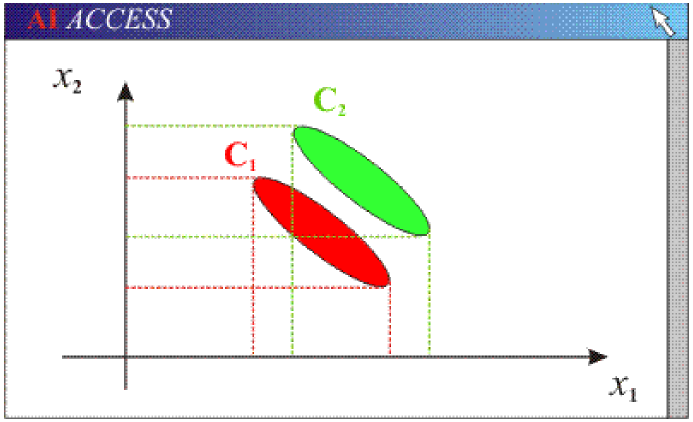
因此，它是一种有效的特征抽取方法。使用这种方法能够使投影后模式样本的类间散布矩阵最大，并且同时类内散布矩阵最小。

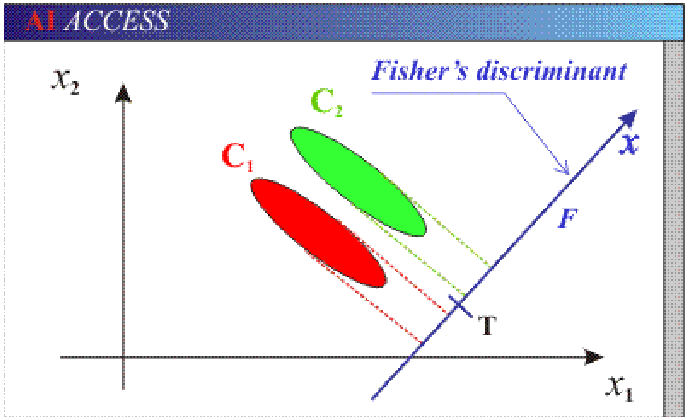




线性判别分析（LDA）

下面给出一个例子，说明**LDA**的目标：





可以看到两个类别，一个绿色类别，一个红色类别。左图是两个类别的原始数据，现在要求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。右图映射到的直线就是用LDA方法计算得到的，可以看到，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）。

线性判别分析（LDA）

要说明白LDA，首先得弄明白线性分类器(Linear Classifier)：因为LDA是一种线性分类器。对于K-分类的一个分类问题，会有K个线性函数：

$$y_k(x) = w_k^T x + w_{k0}$$

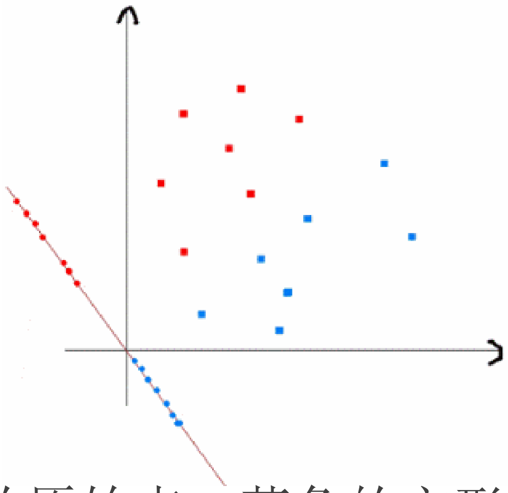
权向量（weight vector）
法向量（normal vector）

阈值（threshold）
偏置（bias）

当满足条件：对于所有的j，都有$Y_k > Y_j$,的时候，我们就说x属于类别k。对于每一个分类，都有一个公式去算一个分值，在所有的公式得到的分值中，找一个最大的，就是所属的分类。

线性判别分析（LDA）

上式实际上就是一种投影，是将一个高维的点投影到一条高维的直线上，LDA最求的目标是，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，如下图所示：



红色的方形的点为0类的原始点、蓝色的方形点为1类的原始点，经过原点的那条线就是投影的直线，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，这个数据只是随便画的，如果在高维的情况下，看起来会更好一点。下面我来推导一下二分类LDA问题的公式：

线性判别分析（LDA）

假设用来区分二分类的直线（投影函数)为： $y = w^T x$

LDA分类的一个目标是使得不同类别之间的距离越远越好，同一类别之中的距离越近越好，所以我们需要定义几个关键的值:

线性判别分析（LDA）

类别i的原始中心点(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： $\widetilde{m_i} = w^T m_i$

衡量类别i投影后，类别点之间的分散程度（方差）为：$\tilde{s}_i = \sum_{y \in Y_i} (y - \widetilde{m}_i)^2$

最终我们可以得到一个下面的公式，表示LDA投影到w后的目标优化函数：

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

线性判别分析（LDA）

我们分类的目标是，使得类别内的点距离越近越好（集中），类别间的点越远越好。

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，分子为两个类别各自的中心点的距离的平方，我们最大化$J(w)$就可以求出最优的w

线性判别分析（LDA）

我们定义一个投影前的各类别分散程度的矩阵，这个矩阵看起来有一点麻烦，其实意思是，如果某一个分类的输入点集**Di**里面的点距离这个分类的中心店m_i越近，则S_i里面元素的值就越小，如果分类的点都紧紧地围绕着m_i，则S_i里面的元素值越更接近**0.**

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

带入S_i，将**J(w)**分母化为：

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

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

线性判别分析（LDA）

同样的将**J(w)**分子化为：

这样目标优化函数可以化成下面的形式：

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

推导过程忽略了，最后推导结果如下：

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

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

线性判别分析（LDA）

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W_i。

LDA在人脸识别中的应用

奇异值分解(singular value decomposition,简称SVD)是一种有效的代数特征提取方法.由于奇异值特征在描述图像时是稳定的,且具有转置不变性、旋转不变性、位移不变性、镜像变换不变性等重要性质,因此奇异值特征可以作为图像的一种有效的代数特征描述。

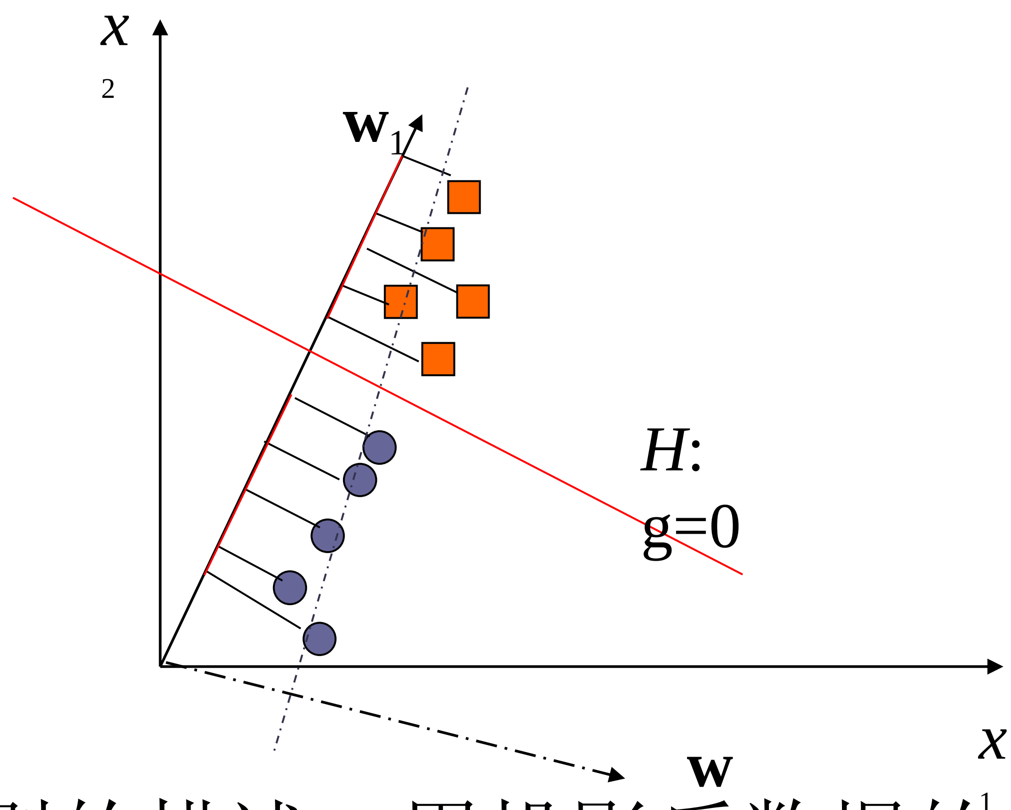
基于主成分分析(principal component analysis,简称PCA)该方法将人脸图像按行(列)展开所形成的一个高维向量看作是一种随机向量,因此可以采用K-L变换获得其正交K-L基底.对应于其中较大特征值的基底具有与人脸相似的形状,故称其为特征脸.利用相对较小的Eigenface集描述人脸,这样每幅人脸图像就对应于一个维数较低的权向量,因此,人脸识别可以在降维后的空间上进行.然而,该方法的缺点是,得到的特征在一般情况下是最佳描述特征(the most expressive features,简称MEFs),而不是最佳分类特征(the most discriminating features,简称MDFs).

主要应用方法

Fisher线性判别方法(Fisher linear discriminant analysis,简称FLD)使投影后的模式样本的类间散布矩阵最大而类内散布矩阵最小,也就是说,投影后保证模式样本在新的空间中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.

Fisher线性判别分析提取的特征向量集强调的是不同人脸的差异而不是照明条件、人脸表情和方向的变化.因而,采用FLD方法对光照条件、人脸姿态等的变化不太敏感,从而有助于提高识别效果。然而,由于在正常情况下人脸识别问题总是一个小样本问题,故其类内散布矩阵总为奇异阵而使此方法的求解变得很困难.

Fisher线性判别图例



Fisher准则的描述：用投影后数据的统计性质（均值和离散度的函数）作为判别优劣的标准。

D 维空间样本分布的描述量

- 各类样本均值向量$\mathbf{m}_i$　$$\mathbf{m}_i=\frac{1}{N_i}\sum_{x\in K_i}\mathbf{x}\quad i=1,2$$
- 样本类内离散度矩阵$\mathbf{S}_i$与总类内离散度矩阵$\mathbf{S}_w$

$$\mathbf{S}_i=\sum_{\mathbf{x}\in\chi_i}(\mathbf{x}-\mathbf{m}_i)(\mathbf{x}-\mathbf{m}_i)^T,\quad i=1,2 \qquad \mathbf{S}_w=\mathbf{S}_1+\mathbf{S}_2$$

- 样本类间离散度矩阵$\mathbf{S}_b$　$$\mathbf{S}_b=(\mathbf{m}_1-\mathbf{m}_2)(\mathbf{m}_1-\mathbf{m}_2)^T$$
  ：

离散度矩阵在形式上与协方差矩阵很相似，但协方差矩阵是一种总体期望值，而离散矩阵只是表示有限个样本在空间分布的离散程度

一维Y空间样本分布的描述量

- **各类样本均值** $\tilde{m}_i = \frac{1}{N_i}\sum_{y\in\psi_i} y, \quad i=1,2$

- 样本类内离散度和总类内离散度

$$\tilde{S}_i = \sum_{y\in\psi_i}(y-\tilde{m}_i)^2, \quad i=1,2 \qquad \tilde{S}_w = \tilde{S}_1 + \tilde{S}_2$$

- 样本类间离散度 $\tilde{S}_b = (\tilde{m}_1 - \tilde{m}_2)^2$

以上定义描述*d*维空间样本点到一向量投影后的**分散情况**。样本离散度的定义与随机变量方差相类似

样本与其投影统计量间的关系

- 样本x与其投影y的统计量之间的关系：

$$\tilde{m}_i = \frac{1}{N_i}\sum_{y\in\psi_i} y = \frac{1}{N_i}\sum_{\mathbf{y}\in K_i} \mathbf{w}^T\mathbf{x} = \mathbf{w}^T\mathbf{m}_i, \quad i = 1,2$$

$$\tilde{S}_b = (\tilde{m}_1 - \tilde{m}_2)^2 = (\mathbf{w}^T\mathbf{m}_1 - \mathbf{w}^T\mathbf{m}_2)^2$$

$$= \mathbf{w}^T(\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T\mathbf{w} = \mathbf{w}^T S_b \mathbf{w}$$

样本与其投影统计量间的关系

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

$$= \sum_{x \in K_i} (\mathbf{w}^T \mathbf{x} - \mathbf{w}^T \mathbf{m}_i)^2$$

$$= \mathbf{w}^T \left[\sum_{x \in K_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T \right] \mathbf{w}$$

$$= \mathbf{w}^T S_i \mathbf{w}$$

$$\tilde{S}_w = \tilde{S}_1 + \tilde{S}_1 = \mathbf{w}^T (S_1 + S_2) \mathbf{w} = \mathbf{w}^T S_w \mathbf{w}$$

Fisher准则函数

- 评价投影方向w的原则，使原样本向量在该方向上的投影能兼顾类间分布尽可能分开，类内尽可能密集的要求
- **Fisher准则函数的定义：**

$$J_F(w)=\frac{\tilde{S}_b}{\tilde{S}_1+\tilde{S}_2}=\frac{\mathbf{w}^T S_b \mathbf{w}}{\mathbf{w}^T S_w \mathbf{w}}$$

- Fisher最佳投影方向的求解

$$\mathbf{w}^*=\underset{\mathbf{w}}{\arg\max}\, J_F(\mathbf{w})$$

Fisher最佳投影方向的求解

- 采用拉格朗日乘子算法解决

$$\mathbf{w}^* = S_w^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$$

$\mathbf{m}_1$-$\mathbf{m}_2$是一向量，对与($\mathbf{m}_1$-$\mathbf{m}_2$)平行的向量投影可使两均值点的距离最远。但是如果从使类间分得较开，同时又使类内密集程度较高这样一个综合指标来看，则需根据两类样本的分布离散程度对投影方向作相应的调整，这就体现在对$\mathbf{m}_1$-$\mathbf{m}_2$ 向量按S_w^{-1}作一线性变换，从而使Fisher准则函数达到极值点

判别函数的确定

- 前面讨论了使Fisher准则函数极大的*d*维向量**w***的计算方法，判别函数中的另一项$\mathbf{w}_0$（阈值）可采用以下几种方法确定：

$$w_0 = -\frac{\tilde{m}_1 + \tilde{m}_2}{2} \qquad w_0 = -\frac{N_1\tilde{m}_1 + N_2\tilde{m}_2}{N_1 + N_2} = \tilde{m}$$

$$w_0 = \frac{\tilde{m}_1 + \tilde{m}_2}{2} - \frac{\ln\left[P(\omega_1)/P(\omega_2)\right]}{N_1 + N_2 - 2}$$

- 分类规则:

$$y = \mathbf{w}^T\mathbf{x} + w_0 > 0 \rightarrow \mathbf{x} \in \omega_1$$

$$y = \mathbf{w}^T\mathbf{x} - w_0 < 0 \rightarrow \mathbf{x} \in \omega_2$$

Fisher公式的推导

$$J_F(w)=\frac{\tilde{S}_b}{\tilde{S}_1+\tilde{S}_2}=\frac{\mathbf{w}^T S_b \mathbf{w}}{\mathbf{w}^T S_w \mathbf{w}}$$

令 $\mathbf{w}^T S_w \mathbf{w} = \mathrm{c} \neq 0$

定义Lagrange函数: $L(\mathbf{w},\lambda)=\mathbf{w}^T S_b \mathbf{w}-\lambda(\mathbf{w}^T S_w \mathbf{w}-c)$

令: $\dfrac{\partial L(\mathbf{w},\lambda)}{\partial \mathbf{w}}=S_b\mathbf{w}-\lambda S_w \mathbf{w}=0$

$$S_w^{-1}S_b\mathbf{w}=\lambda\mathbf{w}$$

$$\lambda\mathbf{w}=S_w^{-1}S_b\mathbf{w}=S_w^{-1}(\mathbf{m}_1-\mathbf{m}_2)(\mathbf{m}_1-\mathbf{m}_2)^T\mathbf{w}$$

$$=S_w^{-1}(\mathbf{m}_1-\mathbf{m}_2)R$$

$$\mathbf{w}^*=\frac{R}{\lambda}S_w^{-1}(\mathbf{m}_1-\mathbf{m}_2)\simeq S_w^{-1}(\mathbf{m}_1-\mathbf{m}_2)$$

Fisher准则举例

例：设两类样本的类内离散矩阵分别为$\mathbf{S}_1,\mathbf{S}_2$，各类样本均值分别为$\mathbf{m}_1=(2, 0)^t$, $\mathbf{m}_2=(2, 2)^t$, 试用Fisher准则求其决策面方程。

$$\mathbf{S}_1=\begin{bmatrix}1 & \frac{1}{2}\\ \frac{1}{2} & 1\end{bmatrix}$$

$$\mathbf{S}_2=\begin{bmatrix}1 & -\frac{1}{2}\\ -\frac{1}{2} & 1\end{bmatrix}$$

答：

$$\mathbf{S}_W=\mathbf{S}_1+\mathbf{S}_2=\begin{bmatrix}2 & 0\\ 0 & 2\end{bmatrix}$$

Fisher准则最佳投影

$$\mathbf{w}^*=\mathbf{S}_w^{-1}(\mathbf{m}_1-\mathbf{m}_2)=\begin{bmatrix}0.5 & 0\\ 0 & 0.5\end{bmatrix}\begin{bmatrix}0\\ -2\end{bmatrix}=\begin{bmatrix}0\\ -1\end{bmatrix}$$

由于两类样本分布形状是相同的（只是方向不同），因此$-w_0$应为（投影后）两类均值的中点

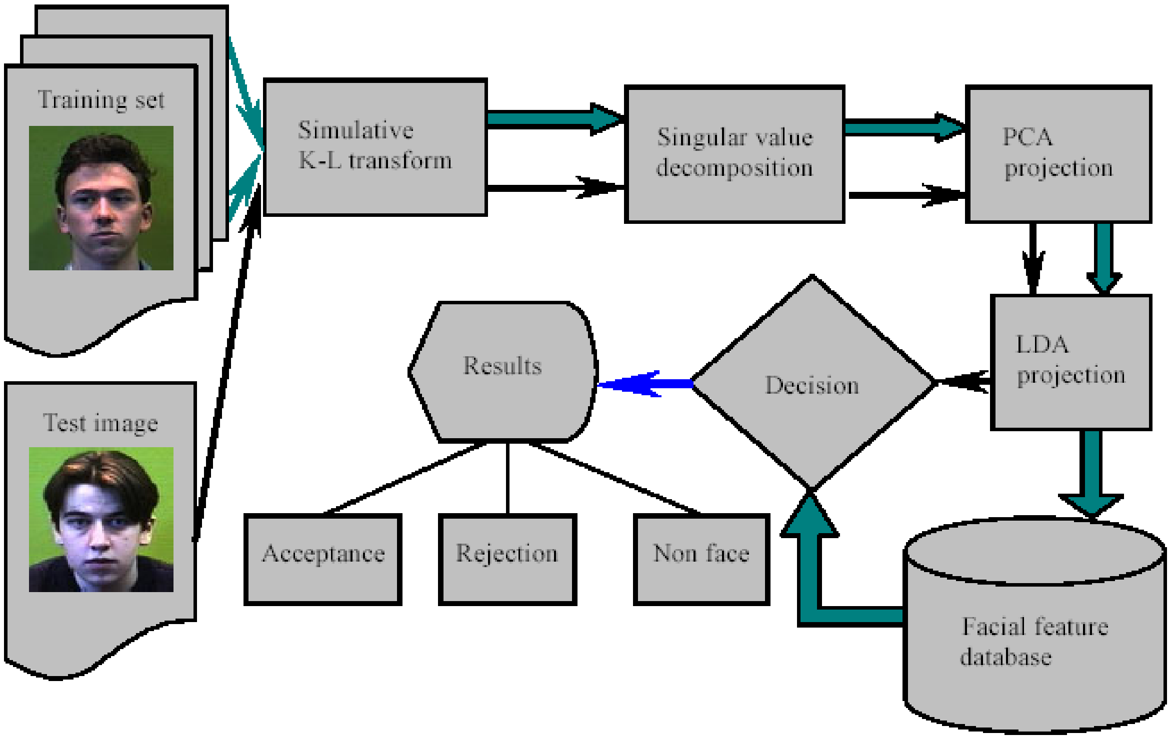
$$w_0=-\frac{\tilde{m}_1+\tilde{m}_2}{2}=-\frac{\mathbf{w}^{*T}(\mathbf{m}_1+\mathbf{m}_2)}{2}=1$$

$$\mathbf{w}^{*T}\mathbf{x}+w_0=0$$

即：$x_2=1$

Fisher准则最佳分界面

人脸识别系统结构


Training set
Test image
Simulative
K-L transform
Singular value
decomposition
PCA
projection
LDA
projection
Results
Decision
Acceptance
Rejection
Non face
Facial feature
database