特征降维

翟婷婷

扬州大学
信息工程（人工智能）学院
zhtt@yzu.edu.cn

2023春

课程目标

- 理解特征降维的意义。
- 掌握主成分分析(PCA)和线性判别分析(FDA)两种特征降维方法的原理，并能实际运用。
- 理解PCA与FDA降维的差别。

引言

- 一般来说，如果用给定的一组特征获得的分类性能不够好，那么考虑添加新的特征是很自然的。

- 在实际应用中，不是特征越多分类精度越高。当特征维数增加到一个临界点后，继续增加反而会导致分类器性能变差，这种现象称为“维数灾难”问题。

- 缓解“维数灾难”经常使用的办法：

  降低特征向量的维数，即特征降维

特征降维

- **降维的可行性**：特征向量往往包含冗余信息：

  ① 有些特征可能与分类问题无关

  ② 特征之间存在着很强的相关性

- **降维方法**：

  重新设计特征：设计生成新的更少的有效特征。

  特征选择：选择现有特征集的一个子集。

  特征变换：对现有特征组合，形成新的更少的特征。

- 本节集中于通过特征变换进行降维的方法。

特征降维

- 两个经典的通过线性变换进行降维的方法：
  - ✓ 主成分分析 (Principal Component Analysis, PCA)：最小化重构误差（误差平方和）
  - ✓ 线性判别分析 (Linear Discriminant Analysis，LDA)：最大化类别可分性
- 两个方法目的都是寻找一个好的线性变换/投影：
  - ✓ PCA：寻找最能够表示(represents)原始数据的投影方法。——不考虑类别
  - ✓ LDA：寻找最能够分开(separates)各类数据的投影方法。——考虑类别

主成分分析(PCA)

- PCA：寻找数据的线性投影，将高维特征向量投影到低维空间中，是无监督的降维方法。
- **首先考虑问题1**：给定d维的样本集$\mathcal{D}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots\boldsymbol{x}_n\}$，要求采用一个向量$\boldsymbol{x}$来最好地表示该样本集，那么，$\boldsymbol{x}$是什么呢?

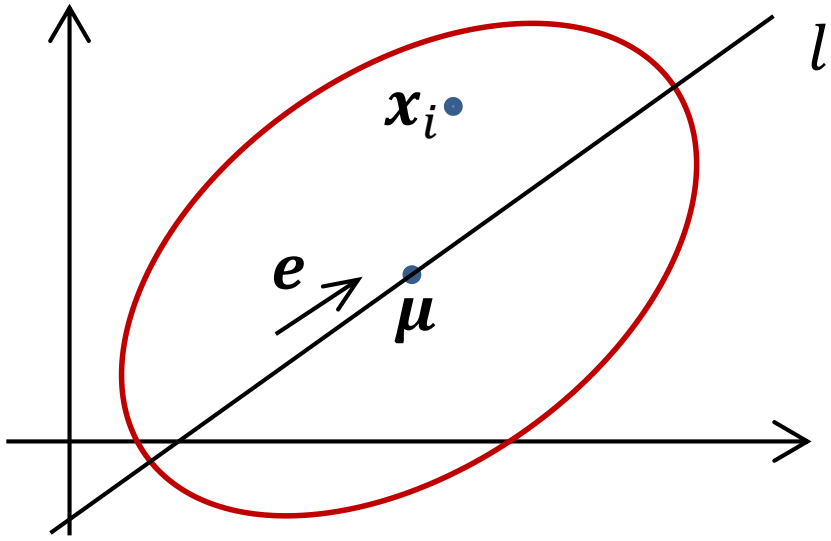
  需要一个准则函数来评价表示的好坏程度，经常使用误差平方和准则函数:

$$J(\boldsymbol{x})=\sum_{i=1}^{n}||\boldsymbol{x}-\boldsymbol{x}_i||^2$$

  使得准则函数$J(\boldsymbol{x})$取得最小值的那个$\boldsymbol{x}$，就是我们要找的那个向量。令$J(\boldsymbol{x})$关于$\boldsymbol{x}$的梯度为零，很容易得到，最优的向量$\boldsymbol{x}^*=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i$，即为样本均值向量$\boldsymbol{\mu}$。

主成分分析(PCA)

- **考虑问题2**：给定d维的样本集$\mathcal{D} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots \boldsymbol{x}_n\}$，记$\boldsymbol{\mu}$为样本集的均值向量，记$l$为通过$\boldsymbol{\mu}$的一条直线，$\boldsymbol{e}$为直线的方向向量，且$||\boldsymbol{e}|| = 1$。将$\mathcal{D}$中的样本投影到直线$l$上，如何投影能使得投影点最好地表示原始样本集？



- 记样本$\boldsymbol{x}_i$投影到直线l上的点记为$\hat{\boldsymbol{x}}_i$，则$\hat{\boldsymbol{x}}_i$可以表示为$\boldsymbol{\mu} + a_i\boldsymbol{e}$，其中$a_i$是一个实数，表示离开$\boldsymbol{\mu}$的距离。

主成分分析(PCA)

- 误差平方和可表示为：

$$J(a_1, a_2 \cdots a_n) = \sum_{i=1}^{n} ||\hat{\boldsymbol{x}}_i - \boldsymbol{x}_i||^2$$

$$= \sum_{i=1}^{n} ||\boldsymbol{\mu} + a_i \boldsymbol{e} - \boldsymbol{x}_i||^2 = \sum_{i=1}^{n} ||a_i \boldsymbol{e} - (\boldsymbol{x}_i - \boldsymbol{\mu})||^2$$

$$= \sum_{i=1}^{n} a_i{}^2 - 2\sum_{i=1}^{n} a_i \boldsymbol{e}^{\mathrm{T}}(\boldsymbol{x}_i - \boldsymbol{\mu}) + \sum_{i=1}^{n} ||\boldsymbol{x}_i - \boldsymbol{\mu}||^2$$

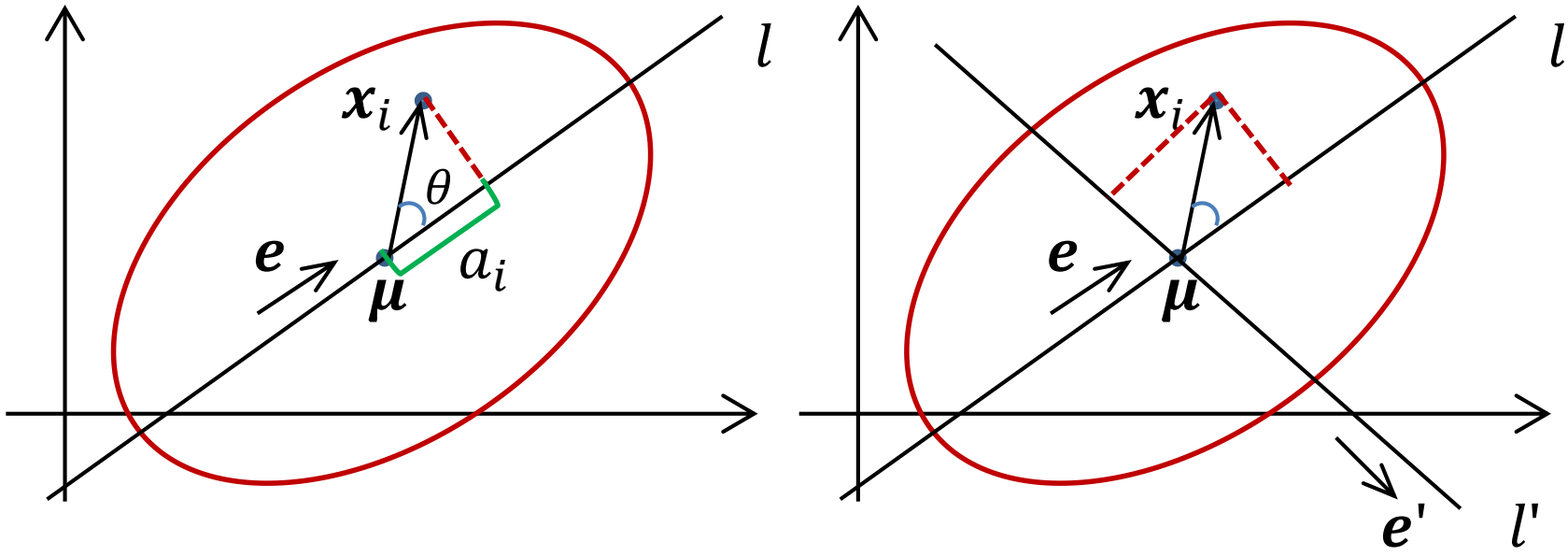
- 求出最优的系数a_i：

$$\frac{\partial J(a_1, a_2 \cdots a_n)}{\partial a_i} = 2a_i - 2\boldsymbol{e}^{\mathrm{T}}(\boldsymbol{x}_i - \boldsymbol{\mu}) = 0$$

$$\Longrightarrow \quad a_i = \boldsymbol{e}^{\mathrm{T}}(\boldsymbol{x}_i - \boldsymbol{\mu}) = ||\boldsymbol{x}_i - \boldsymbol{\mu}|| \cos\theta \qquad (||\boldsymbol{e}|| = 1)$$

- 结果表明：对每个样本垂直投影到直线l上，会产生最小的误差平方和。

主成分分析(PCA)



- **考虑问题3**:
  如何找到一个最优的投影方向$\boldsymbol{e}$，使得$\mathcal{D}$中的样本投影到该方向上的投影点能最好地表示原始样本集?

主成分分析(PCA)

- 把$a_i = \boldsymbol{e}^{\mathrm{T}}(\boldsymbol{x}_i - \boldsymbol{\mu})$带入误差平方和准则函数:

$$J(\boldsymbol{e}) = \sum_{i=1}^{n} {a_i}^2 - 2\sum_{i=1}^{n} {a_i}^2 + \sum_{i=1}^{n} ||\boldsymbol{x}_i - \boldsymbol{\mu}||^2$$

$$= -\sum_{i=1}^{n} \boldsymbol{e}^{\mathrm{T}}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{e} + \sum_{i=1}^{n} ||\boldsymbol{x}_i - \boldsymbol{\mu}||^2$$

$$= -\boldsymbol{e}^{\mathrm{T}}S\boldsymbol{e} + \sum_{i=1}^{n} ||\boldsymbol{x}_i - \boldsymbol{\mu}||^2$$

其中,

$$S = \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^{\mathrm{T}}$$

S称为散布矩阵(Scatter Matrix), 它是样本协方差矩阵的$n-1$倍。S是对称半正定矩阵。

主成分分析(PCA)

- 要找到满足如下条件的$\boldsymbol{e}$:

$$\min_{\boldsymbol{e}} J(\boldsymbol{e}) \text{ 满足} ||\boldsymbol{e}|| = 1$$

等价于: $\min_{\boldsymbol{e}} -\boldsymbol{e}^{\mathrm{T}} S\boldsymbol{e}$ 满足$\boldsymbol{e}^{\mathrm{T}}\boldsymbol{e} = 1$

- 应用拉格朗日乘子法:

$$f(\boldsymbol{e}, \lambda) = -\boldsymbol{e}^{\mathrm{T}} S\boldsymbol{e} + \lambda(\boldsymbol{e}^{\mathrm{T}}\boldsymbol{e} - 1)$$

令对$\boldsymbol{e}$求偏导为零得到:

$$\frac{\partial f(\boldsymbol{e})}{\partial \boldsymbol{e}} = -2S\boldsymbol{e} + 2\lambda\boldsymbol{e} = 0$$

$$\Longrightarrow \quad S\boldsymbol{e} = \lambda\boldsymbol{e} \Longrightarrow \boldsymbol{e}^{\mathrm{T}} S\boldsymbol{e} = \lambda\boldsymbol{e}^{\mathrm{T}}\boldsymbol{e} = \lambda$$

λ为S的特征值，$\boldsymbol{e}$为S对应于特征值λ的特征向量

- 为了最小化误差平方和函数$J(\boldsymbol{e})$，选取散布矩阵S最大的特征值对应的特征向量$\boldsymbol{e}$作为投影直线的方向。

主成分分析(PCA)

- 通过选取散布矩阵S最大的特征值对应的特征向量$\boldsymbol{e}$作为投影方向，可以将一个d维的训练样本$\boldsymbol{x}_i$投影到一维的空间中，它的坐标值为：

$$a_i = \boldsymbol{e}^{\mathrm{T}}(\boldsymbol{x}_i - \boldsymbol{\mu})$$

- 考虑更一般的情况：投影到d'维的空间中，$d' < d$

  设$\boldsymbol{e}^1, \boldsymbol{e}^2 \cdots \boldsymbol{e}^{d'}$分别为以$\boldsymbol{\mu}$为原点的$d'$个方向向量，$\boldsymbol{e}^i \perp \boldsymbol{e}^j$，$\forall i \neq j$，$||\boldsymbol{e}^i|| = 1$，则一个样本$\boldsymbol{x}_i$投影到$d'$维子空间中的点$\hat{\boldsymbol{x}}_i$可以表示为：

$$\hat{\boldsymbol{x}}_i = \boldsymbol{\mu} + \sum_{j=1}^{d'} a_i{}^j \boldsymbol{e}^j$$

  带入误差平方和函数，先求出投影后在每个方向上的坐标值$a_i{}^j$，然后再求出最优的投影方向$\boldsymbol{e}^1, \boldsymbol{e}^2 \cdots \boldsymbol{e}^{d'}$。

主成分分析(PCA)

- 误差平方和函数为:

$$J=\sum_{i=1}^{n}||\widehat{\boldsymbol{x}}_i-\boldsymbol{x}_i||^2=\sum_{i=1}^{n}||\sum_{j=1}^{d'}a_i{}^j\boldsymbol{e}^j-(\boldsymbol{x}_i-\boldsymbol{\mu})||^2$$

$$=\sum_{i=1}^{n}\left(||\sum_{j=1}^{d'}a_i{}^j\boldsymbol{e}^j||^2-2\left(\sum_{j=1}^{d'}a_i{}^j\boldsymbol{e}^j\right)^{\mathsf{T}}(\boldsymbol{x}_i-\boldsymbol{\mu})+||\boldsymbol{x}_i-\boldsymbol{\mu}||^2\right)$$

$$=\sum_{i=1}^{n}\left(\sum_{j=1}^{d'}(a_i{}^j)^2-2\sum_{j=1}^{d'}a_i{}^j\,\boldsymbol{e}^{j\mathsf{T}}(\boldsymbol{x}_i-\boldsymbol{\mu})+||\boldsymbol{x}_i-\boldsymbol{\mu}||^2\right)$$

其中:

$$||\sum_{j=1}^{d'}a_i{}^j\boldsymbol{e}^j||^2=\left(\sum_{j=1}^{d'}a_i{}^j\boldsymbol{e}^j\right)^{\mathsf{T}}\left(\sum_{j=1}^{d'}a_i{}^j\boldsymbol{e}^j\right)=\sum_{j=1}^{d'}(a_i{}^j)^2$$

- 令准则函数对$a_i{}^j$求偏导数为零:

$$\frac{\partial J}{\partial a_i{}^j}=2a_i{}^j-2\boldsymbol{e}^{j\mathsf{T}}(\boldsymbol{x}_i-\boldsymbol{\mu})=0$$

$$a_i{}^j=\boldsymbol{e}^{j\mathsf{T}}(\boldsymbol{x}_i-\boldsymbol{\mu})$$

主成分分析(PCA)

- 代回误差平方和函数:

$$J=\sum_{i=1}^{n}\left(\sum_{j=1}^{d'}(a_i{}^j)^2-2\left(\sum_{j=1}^{d'}a_i{}^j\boldsymbol{e}^j\right)^{\mathsf{T}}(\boldsymbol{x_i}-\boldsymbol{\mu})+||\boldsymbol{x_i}-\boldsymbol{\mu}||^2\right)$$

$$=\sum_{i=1}^{n}\left(-\sum_{j=1}^{d'}(a_i{}^j)^2+||\boldsymbol{x_i}-\boldsymbol{\mu}||^2\right)$$

$$=-\sum_{i=1}^{n}\sum_{j=1}^{d'}\boldsymbol{e}^{j\mathsf{T}}(\boldsymbol{x_i}-\boldsymbol{\mu})(\boldsymbol{x_i}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{e}^j+\sum_{i=1}^{n}||\boldsymbol{x_i}-\boldsymbol{\mu}||^2$$

$$=-\sum_{j=1}^{d'}\boldsymbol{e}^{j\mathsf{T}}\boldsymbol{S}\boldsymbol{e}^j+\sum_{i=1}^{n}||\boldsymbol{x_i}-\boldsymbol{\mu}||^2 \quad (\boldsymbol{S}\text{为散布矩阵})$$

- 也就是说，我们要找的最优投影方向$\boldsymbol{e}^1,\boldsymbol{e}^2\cdots\boldsymbol{e}^{d'}$为:

$$\min_{\boldsymbol{e}^1,\boldsymbol{e}^2\cdots\boldsymbol{e}^{d'}}-\sum_{j=1}^{d'}\boldsymbol{e}^{j\mathsf{T}}\boldsymbol{S}\boldsymbol{e}^j \text{ 满足}\boldsymbol{e}^{j\mathsf{T}}\boldsymbol{e}^j=1,\ \forall j=1,2,\cdots d'$$

- 同样使用拉格朗日乘子法求解:

主成分分析(PCA)

$$f(\boldsymbol{e}^1, \boldsymbol{e}^2 \cdots \boldsymbol{e}^{d'}, \lambda^1 \cdots \lambda^{d'}) = -\sum_{j=1}^{d'} \boldsymbol{e}^{j\top} \boldsymbol{S} \boldsymbol{e}^j + \sum_{j=1}^{d'} \lambda^j (\boldsymbol{e}^{j\top} \boldsymbol{e}^j - 1)$$

- 令f对$\boldsymbol{e}^j$求偏导数为零:

$$\frac{\partial f}{\partial \boldsymbol{e}^j} = -2\boldsymbol{S}\boldsymbol{e}^j + 2\lambda^j \boldsymbol{e}^j = 0$$

$$\Rightarrow \boldsymbol{S}\boldsymbol{e}^j = \lambda^j \boldsymbol{e}^j \Rightarrow \boldsymbol{e}^{j\top} \boldsymbol{S} \boldsymbol{e}^j = \lambda^j$$

代回误差平方和函数得到:

$$J = -\sum_{j=1}^{d'} \lambda^j + \sum_{i=1}^{n} ||\boldsymbol{x_i} - \boldsymbol{\mu}||^2$$

因此，为了使得误差平方和函数取得最小值，选择的d'个投影方向应该是散布矩阵$\boldsymbol{S}$的前d'个最大的特征值对应的d'个特征向量。由于散布矩阵$\boldsymbol{S}$是实对称矩阵(d维)，一定可以找到d'个正交的特征向量。

主成分分析(PCA)

- 求出的$\boldsymbol{e}^1, \boldsymbol{e}^2 \cdots \boldsymbol{e}^{d'}$称为主成分(Principal component)，它们构成一个新的坐标系。一个d维的向量$\boldsymbol{x}_i$在该坐标系中新的表示是一个d'维的向量$(a_i{}^1, a_i{}^2, \cdots a_i{}^{d'})^{\mathsf{T}}$

  其中$a_i{}^j = \boldsymbol{e}^{j\mathsf{T}}(\boldsymbol{x}_i - \boldsymbol{\mu})$是向量$\boldsymbol{x}_i - \boldsymbol{\mu}$在$\boldsymbol{e}^j$上投影的长度。

- 重新表示$\boldsymbol{x}_i$降维后的向量为：

$$\begin{bmatrix} a_i{}^1 \\ a_i{}^2 \\ \vdots \\ a_i{}^{d'} \end{bmatrix} = \begin{bmatrix} \boldsymbol{e}^{\mathbf{1}\mathsf{T}}(\boldsymbol{x}_i - \boldsymbol{\mu}) \\ \boldsymbol{e}^{\mathbf{2}\mathsf{T}}(\boldsymbol{x}_i - \boldsymbol{\mu}) \\ \vdots \\ \boldsymbol{e}^{d'\mathsf{T}}(\boldsymbol{x}_i - \boldsymbol{\mu}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{e}^{\mathbf{1}\mathsf{T}} \\ \boldsymbol{e}^{\mathbf{2}\mathsf{T}} \\ \vdots \\ \boldsymbol{e}^{d'\mathsf{T}} \end{bmatrix} (\boldsymbol{x}_i - \boldsymbol{\mu})$$

$$= [\boldsymbol{e}^1\ \boldsymbol{e}^2 \cdots \boldsymbol{e}^{d'}]^{\mathsf{T}}(\boldsymbol{x}_i - \boldsymbol{\mu}) = Q^{\mathsf{T}}(\boldsymbol{x}_i - \boldsymbol{\mu})$$

其中：$Q = [\boldsymbol{e}^1, \boldsymbol{e}^2 \cdots \boldsymbol{e}^{d'}]$称为投影矩阵。对任意一个$\boldsymbol{x}_i$，通过$\boldsymbol{Q}^{\mathsf{T}}(\boldsymbol{x}_i - \boldsymbol{\mu})$可以将$\boldsymbol{x}_i$从$d$维降到$d'$维，其中$d' < d$。

主成分分析(PCA)

- **主成分分析步骤（d维降为d'维）：**

1、计算散布矩阵S (或样本的协方差矩阵)

$$S=\sum_{i=1}^{n}(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^{\mathsf{T}}，\ 其中\boldsymbol{\mu}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i$$

2、计算S的特征值和特征向量;

3、按照特征值从大到小对相应的特征向量排序;

4、选择特征值最大的前d'个特征向量作为投影向量，构成$d*d'$维的**标准正交**的投影矩阵Q;

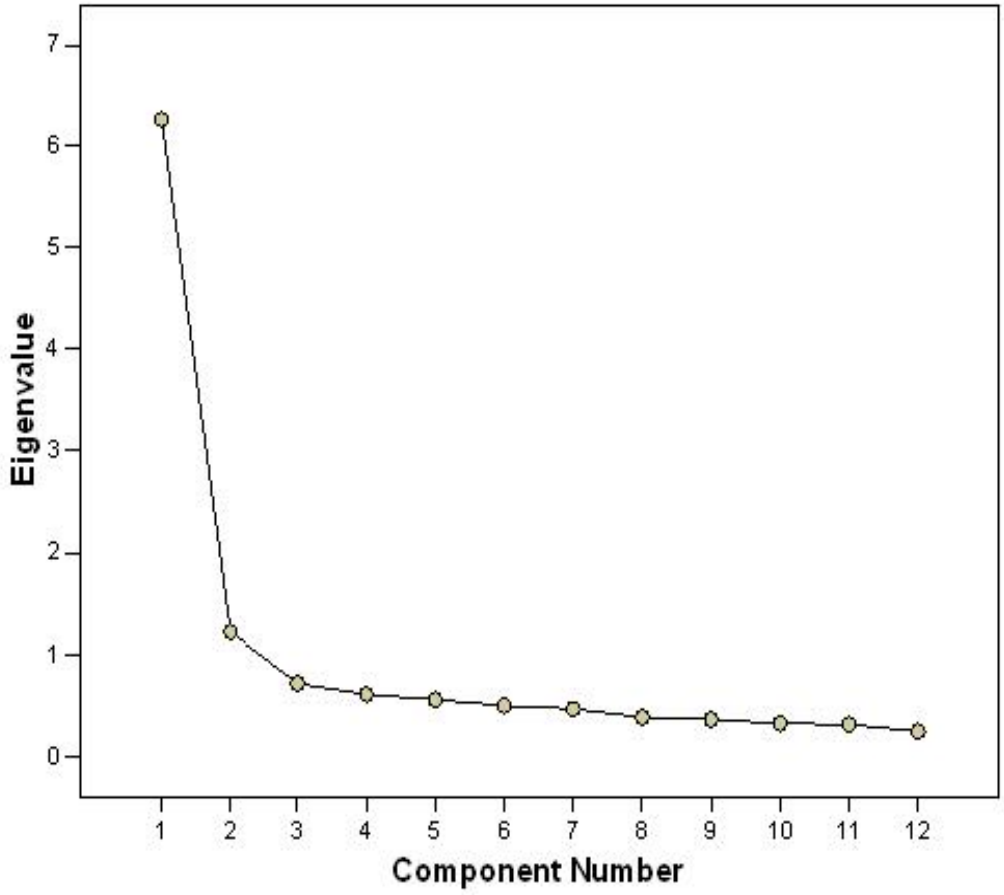
5、对任意的d维样本$\boldsymbol{x}_i$，用PCA降维后的d'维向量为:

$$\boldsymbol{y}_i=Q^{\mathsf{T}}(\boldsymbol{x}_i-\boldsymbol{\mu})$$

重构公式：$\boldsymbol{x}_i\approx Q\boldsymbol{y}_i+\boldsymbol{\mu}$

主成分分析(PCA)

- 通常，最大的几个特征值占据了所有特征值之和的绝大部分。将数据投影到少数几个最大特征值对应的特征向量方向上即可保留原数据中的绝大部分信息。



剩下的小部分信息，即较小的特征值对应的特征向量所表示的信息，通常可以认为是数据噪声而丢掉。

主成分分析(PCA)

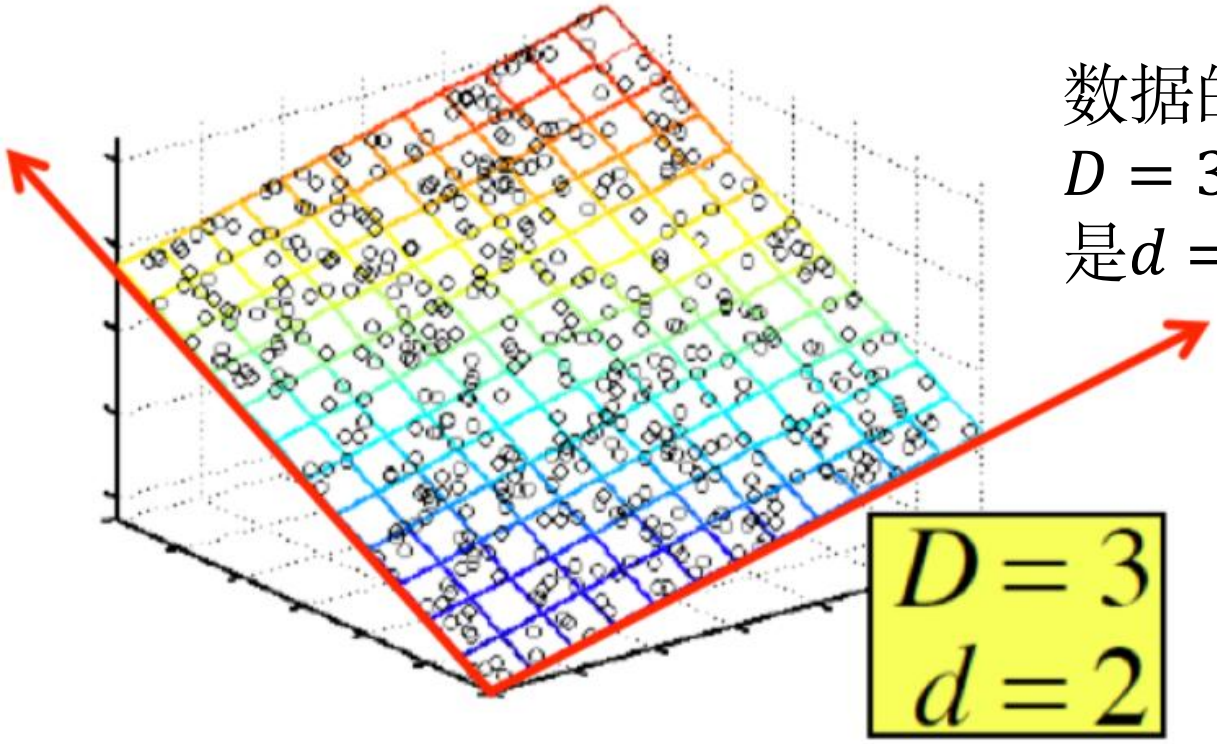
- 原始数据的**本征维度**(intrinsic dimensionality)决定了能否在d'**维**$(d' < d)$的子空间中**充分表示**给定的d**维**数据。

- 数据集的**本征维度**可以看作是表示数据集所需要的**最小变量数**。因为任意低维数据可简单地通过增加(如复制)不必要的维度或随机维将其转换至更高维空间中，所以数据集的本征维度小于等于数据集的原始维度。

- 给定一个d**维**的数据集，若数据集的**本征维度为**d'**维**，$d' < d$，可以将该数据集从d**维**削减至d'**维**，不会丢失重要信息。

主成分分析(PCA)

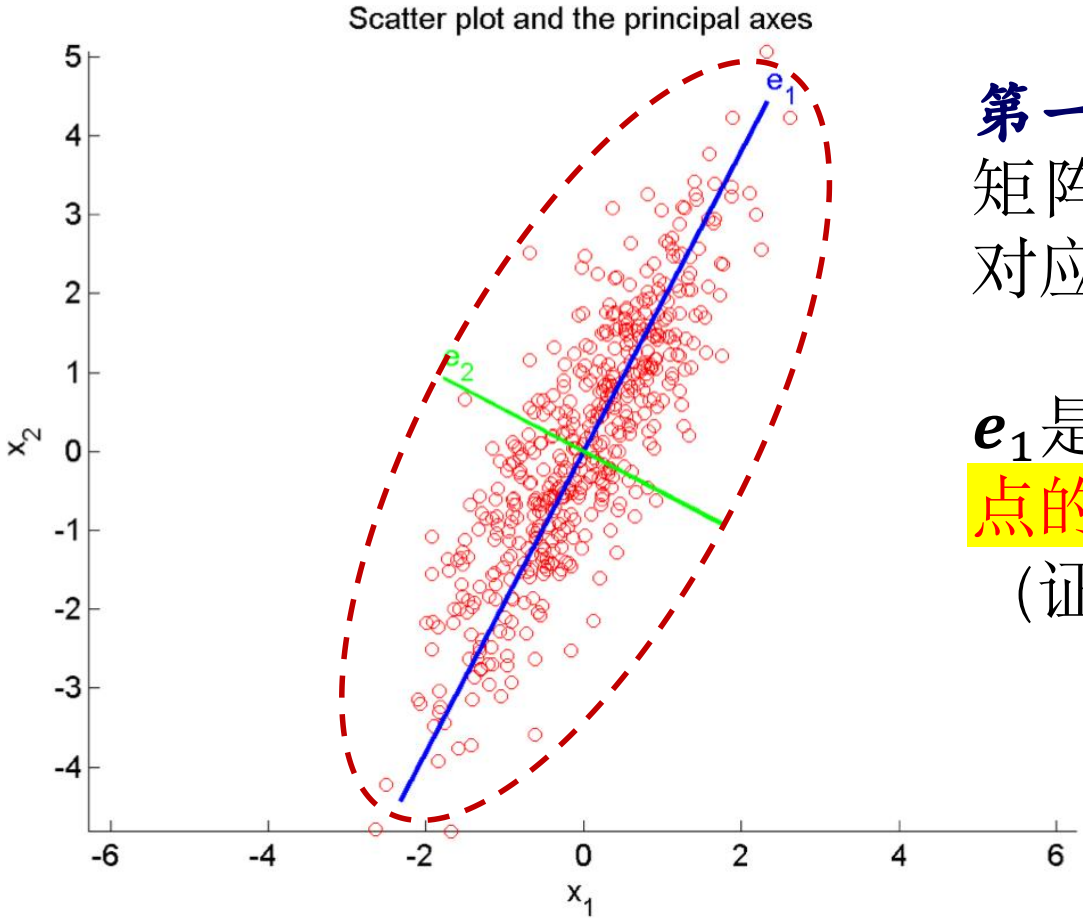
- 本征维度的几何解释是整个数据集中的样本位于d'**维**超曲面拓扑上。



数据的原始维度是$D = 3$，本征维度是$d = 2$。

主成分分析(PCA)

- PCA的几何解释：样本在d维空间中形成一个椭球形的云团，散布矩阵S的特征向量就是这个云团的主轴。



第一主成分e_1是散布矩阵S的最大特征值对应的特征向量。

e_1是所有样本的投影点的方差最大的方向。

（证明-课后作业）

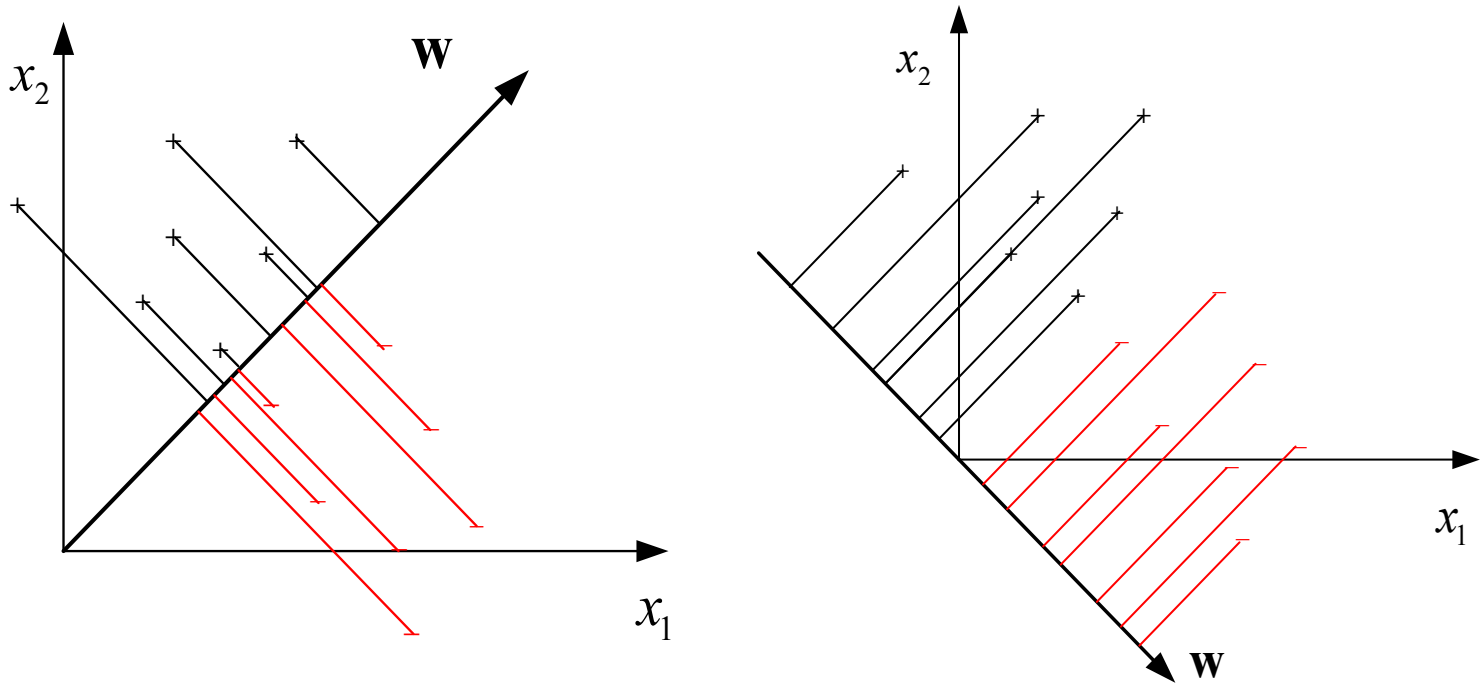
Fisher线性判别分析(FDA)

- PCA寻找的是用来有效表示数据的主轴方向(主成分)，是无监督的特征降维方法，没有考虑样本的类别信息。

- Fisher线性判别分析(FDA)寻找的是能够有效分类的方向，是有监督的降维方法。

Fisher线性判别分析(FDA)

- 对于两类别的样本集合，FDA希望找到一个投影方向，使得不同类别的样本在该方向上的投影尽量的分开。



对于两类别的样本集合，向不同的方向投影，产生的数据点的可分程度是不同的。

Fisher线性判别分析(FDA)

- 给定一个样本集合$\mathcal{D}=\{\mathcal{D}_1,\mathcal{D}_2\}$：$\mathcal{D}_1$中的样本属于$\omega_1$，样本数为$n_1$；$\mathcal{D}_2$中的样本属于$\omega_2$，样本数为$n_2$。
- FDA的目标：找到一个直线方向$\boldsymbol{w}$，**将两类样本最好地分开**，其中$\|\boldsymbol{w}\|=1$。样本$\boldsymbol{x}$投影后的点的坐标为$y=\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}$。

FDA定义一个准则函数衡量两类样本分开的程度：

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

分子：不同类的投影点尽量分开

分母：同类的投影点尽量靠近

其中，$\tilde{m}_i$是$\mathcal{D}_i$中样本投影后的点的均值；

$\tilde{s}_i{}^2$是$\mathcal{D}_i$中样本投影后的点的分散程度，即方差。

Fisher线性判别分析(FDA)

- $\tilde{m}_i$是$\mathcal{D}_i$中样本投影后的点的均值:

$$\tilde{m}_1 = \frac{1}{n_1}\sum_{\boldsymbol{x}_i \in \mathcal{D}_1} \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{m}_1, \quad 其中\boldsymbol{m}_1 = \frac{1}{n_1}\sum_{\boldsymbol{x}_i \in \mathcal{D}_1} \boldsymbol{x}_i$$

$$\tilde{m}_2 = \frac{1}{n_2}\sum_{\boldsymbol{x}_i \in \mathcal{D}_2} \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{m}_2 \ , \quad 其中\boldsymbol{m}_2 = \frac{1}{n_2}\sum_{\boldsymbol{x}_i \in \mathcal{D}_2} \boldsymbol{x}_i$$

$$(\tilde{m}_1 - \tilde{m}_2)^2 = \boldsymbol{w}^{\mathrm{T}}(\boldsymbol{m}_1 - \boldsymbol{m}_2)(\boldsymbol{m}_1 - \boldsymbol{m}_2)^{\mathrm{T}}\boldsymbol{w} = \boldsymbol{w}^{\mathrm{T}}S_{\mathrm{B}}\boldsymbol{w}$$

其中,

$$S_{\mathrm{B}} = (\boldsymbol{m}_1 - \boldsymbol{m}_2)(\boldsymbol{m}_1 - \boldsymbol{m}_2)^{\mathrm{T}}$$

称S_{B}为类间散布矩阵 **(between-class scatter matrix)** ,
它是对称半正定的。

Fisher线性判别分析(FDA)

- $\tilde{s}_i^{\,2}$是$\mathcal{D}_i$中样本投影后的点的方差：

$$\tilde{s}_1^{\,2} = \frac{1}{n_1}\sum_{\boldsymbol{x}_i \in \mathcal{D}_1} (\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i - \tilde{m}_1)^2 = \frac{1}{n_1}\sum_{\boldsymbol{x}_i \in \mathcal{D}_1} (\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{m}_1)^2$$

$$= \frac{1}{n_1}\sum_{\boldsymbol{x}_i \in \mathcal{D}_1} \boldsymbol{w}^{\mathrm{T}}(\boldsymbol{x}_i - \boldsymbol{m}_1)(\boldsymbol{x}_i - \boldsymbol{m}_1)^{\mathrm{T}}\boldsymbol{w} = \boldsymbol{w}^{\mathrm{T}} S_1 \boldsymbol{w}$$

其中，$S_1 = \frac{1}{n_1}\sum_{\boldsymbol{x}_i \in \mathcal{D}_1} (\boldsymbol{x}_i - \boldsymbol{m}_1)(\boldsymbol{x}_i - \boldsymbol{m}_1)^{\mathrm{T}}$

称为ω_1类的类内散布矩阵。

- 同理可得：$\tilde{s}_2^{\,2} = \boldsymbol{w}^{\mathrm{T}} S_2 \boldsymbol{w}$

其中，$S_2 = \frac{1}{n_2}\sum_{\boldsymbol{x}_i \in \mathcal{D}_2} (\boldsymbol{x}_i - \boldsymbol{m}_2)(\boldsymbol{x}_i - \boldsymbol{m}_2)^{\mathrm{T}}$，称为$\omega_2$类的类内散布矩阵。

Fisher线性判别分析(FDA)

- 定义**总的类内散布矩阵(within-class scatter matrix):**

$$S_w = S_1 + S_2$$

则可以得到:

$$\tilde{s}_1^{\,2} + \tilde{s}_2^{\,2} = \boldsymbol{w}^{\mathrm{T}} S_1 \boldsymbol{w} + \boldsymbol{w}^{\mathrm{T}} S_2 \boldsymbol{w} = \boldsymbol{w}^{\mathrm{T}} S_{\mathrm{w}} \boldsymbol{w}$$

- 代入准则函数得到:

$$J(\boldsymbol{w}) = \frac{\boldsymbol{w}^{\mathrm{T}} S_{\mathrm{B}} \boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}} S_{\mathrm{w}} \boldsymbol{w}}$$

广义瑞利熵
generalizad
Rayleigh quotient

FDA的目标是找到一个最大化$J(\boldsymbol{w})$的$\boldsymbol{w}$。

根据瑞利商函数的性质，$J(\boldsymbol{w})$的最大值为$S_{\mathrm{w}}^{-1} S_{\mathrm{B}}$的最大特征值，最小值为$S_{\mathrm{w}}^{-1} S_{\mathrm{B}}$的最小特征值。

Fisher线性判别分析(FDA)

- 问题可以重写为

$$\min_w -\boldsymbol{w}^{\mathrm{T}} S_{\mathrm{B}} \boldsymbol{w} \quad s.t. \boldsymbol{w}^{\mathrm{T}} S_{\mathrm{w}} \boldsymbol{w} = 1$$

- 拉格朗日乘子法: $f(\boldsymbol{w}, \lambda) = -\boldsymbol{w}^{\mathrm{T}} S_{\mathrm{B}} \boldsymbol{w} + \lambda(\boldsymbol{w}^{\mathrm{T}} S_{\mathrm{w}} \boldsymbol{w} - 1)$
- 对$\boldsymbol{w}$求偏导，令梯度为0:

$$\frac{\partial f(\boldsymbol{w}, \lambda)}{\partial \boldsymbol{w}} = -2S_{\mathrm{B}} \boldsymbol{w} + 2\lambda S_{\mathrm{w}} \boldsymbol{w} = 0$$

- 解得: $S_{\mathrm{B}} \boldsymbol{w} = \lambda S_{\mathrm{w}} \boldsymbol{w}$ (S_{B}的秩最多为1，S_{w}在样本数大于维数时经常是非奇异矩阵)

$$\Longrightarrow {S_{\mathrm{w}}}^{-1} S_{\mathrm{B}} \boldsymbol{w} = \lambda \boldsymbol{w}$$

$$\Longrightarrow \boldsymbol{w} = \frac{1}{\lambda} {S_{\mathrm{w}}}^{-1} S_{\mathrm{B}} \boldsymbol{w} = \frac{1}{\lambda} {S_{\mathrm{w}}}^{-1} (\boldsymbol{m}_1 - \boldsymbol{m}_2) \boxed{(\boldsymbol{m}_1 - \boldsymbol{m}_2)^{\mathrm{T}} \boldsymbol{w}}$$

标量

$$\Longrightarrow \boldsymbol{w} = \alpha {S_{\mathrm{w}}}^{-1} (\boldsymbol{m}_1 - \boldsymbol{m}_2)$$

Fisher线性判别分析(FDA)

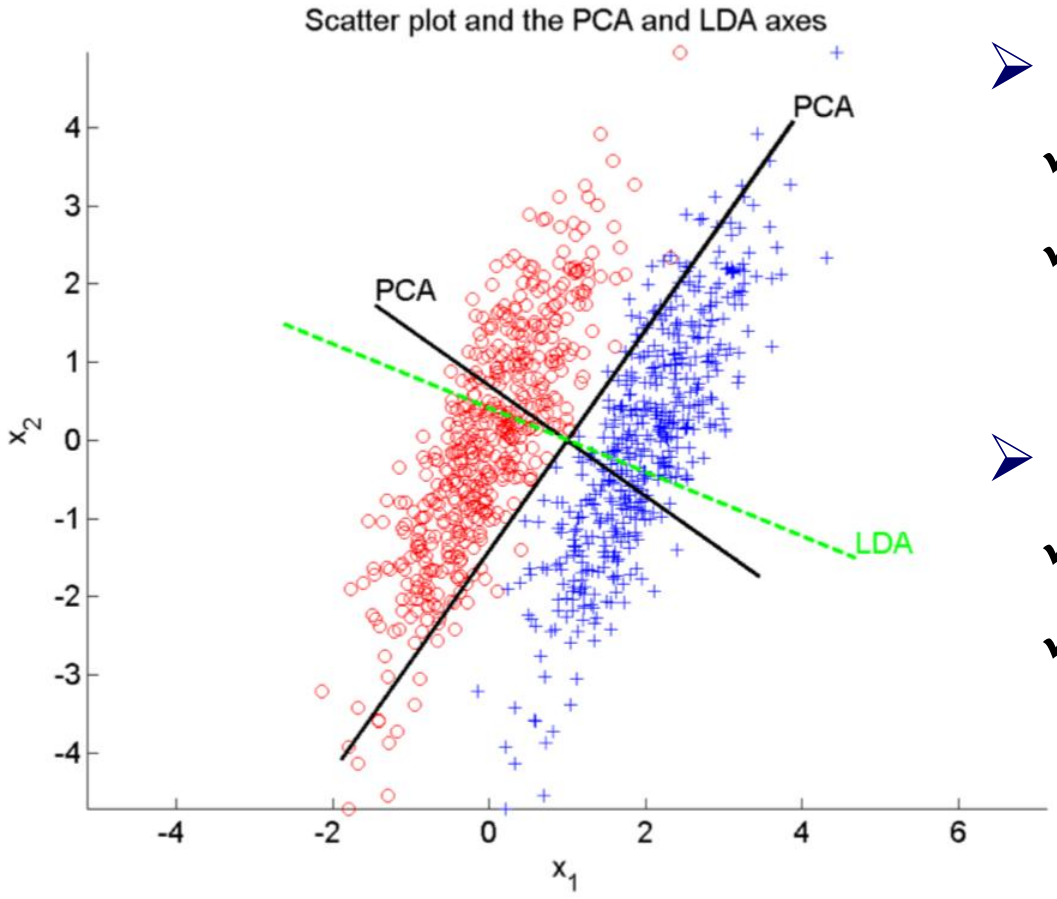
- 由于$\boldsymbol{w}$的模对问题本身无关紧要，因此使得准则函数最大化的$\boldsymbol{w}$取值为：

$$\boldsymbol{w} = {S_{\mathrm{w}}}^{-1}(\boldsymbol{m}_1 - \boldsymbol{m}_2)$$

- 一个d维的样本$\boldsymbol{x}_i$投影到方向$\boldsymbol{w}$上，成为一个1维的点：$y_i = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i$。

Fisher线性判别分析(FDA)

PCA vs. FDA：



- **主成分分析(PCA)：**
  - ✓ 无监督
  - ✓ 寻找用来有效表示数据的投影
- **线性判别分析(FDA)：**
  - ✓ 有监督
  - ✓ 寻找用来有效分类的投影

多重FDA（课下了解）

- 如何将FDA拓展到多类别的数据集上？

  多重FDA：

  两类问题需要一个Fisher投影方向

  c类问题需要$c-1$个Fisher投影方向

- 给定一个样本集$\mathcal{D}=\{\mathcal{D}_1,\mathcal{D}_2\cdots\mathcal{D}_c\}$：$\mathcal{D}_i$中的样本属于$\omega_i$类，样本数为$n_i$，$\mathcal{D}$中的总样本数为$n=\sum_i n_i$。

- 多重FDA的目的是找到$c-1$个投影方向$\boldsymbol{w}_1,\boldsymbol{w}_2\cdots\boldsymbol{w}_{c-1}$，使得不同类别的样本投影后的点尽可能分开。

  记$\boldsymbol{W}=[\boldsymbol{w}_1,\boldsymbol{w}_2\cdots\boldsymbol{w}_{c-1}]$，$\boldsymbol{W}$是$d*(c-1)$维投影矩阵。

多重FDA（课下了解）

- 对ω_i类的类内散布矩阵进行拓展:

$$S_i = \sum_{\boldsymbol{x}\in\mathcal{D}_i} (\boldsymbol{x}-\boldsymbol{m}_i)(\boldsymbol{x}-\boldsymbol{m}_i)^{\mathsf{T}}, \quad 其中\boldsymbol{m}_i = \frac{1}{n_i}\sum_{\boldsymbol{x}\in\mathcal{D}_i}\boldsymbol{x}$$

总的类内散布矩阵为:

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

- 定义类间散布矩阵为:

$$S_B = \sum_{i=1}^{c} n_i\,(\boldsymbol{m}_i-\boldsymbol{m})(\boldsymbol{m}_i-\boldsymbol{m})^{\mathsf{T}}$$

其中，$\boldsymbol{m} = \frac{1}{n}\sum_{\boldsymbol{x}\in\mathcal{D}}\boldsymbol{x}$为总的均值向量。

多重FDA（课下了解）

- 准则函数变为：

$$J(\boldsymbol{W}) = \frac{|\boldsymbol{W}^{\top} S_{\mathrm{B}} \boldsymbol{W}|}{|\boldsymbol{W}^{\top} S_{\mathrm{w}} \boldsymbol{W}|}$$

  其中$|\cdot|$表示行列式。

- 可以得到：使得$J(\boldsymbol{W})$最大化的$\boldsymbol{W}$的列向量是矩阵${S_{\mathrm{w}}}^{-1} S_{\mathrm{B}}$的最大的特征值对应的特征向量。

- 一旦使用PCA或LDA完成从d维原始特征空间到低维子空间的转换，就可以使用降维后的数据集来训练分类器。

小结

- 降低特征向量的维数，即特征降维，是缓解“维数灾难”经常的办法。
- 主成分分析 (PCA)和线性判别分析(LDA)是两种通过特征变换进行降维的方法。
- 两个方法目的都是寻找一个好的线性变换/投影：
  - ✓ PCA：寻找最能够表示(represents) 原始数据的投影方向。——不考虑类别
  - ✓ LDA：寻找最能够分开(separates)各类数据的投影方向。——考虑类别
- 使用PCA和LDA进行特征降维的步骤。