降维

主要内容

- LDA： Linear Discriminant Analysis
  - 主题模型LDA： Latent Dirichlet Allocation

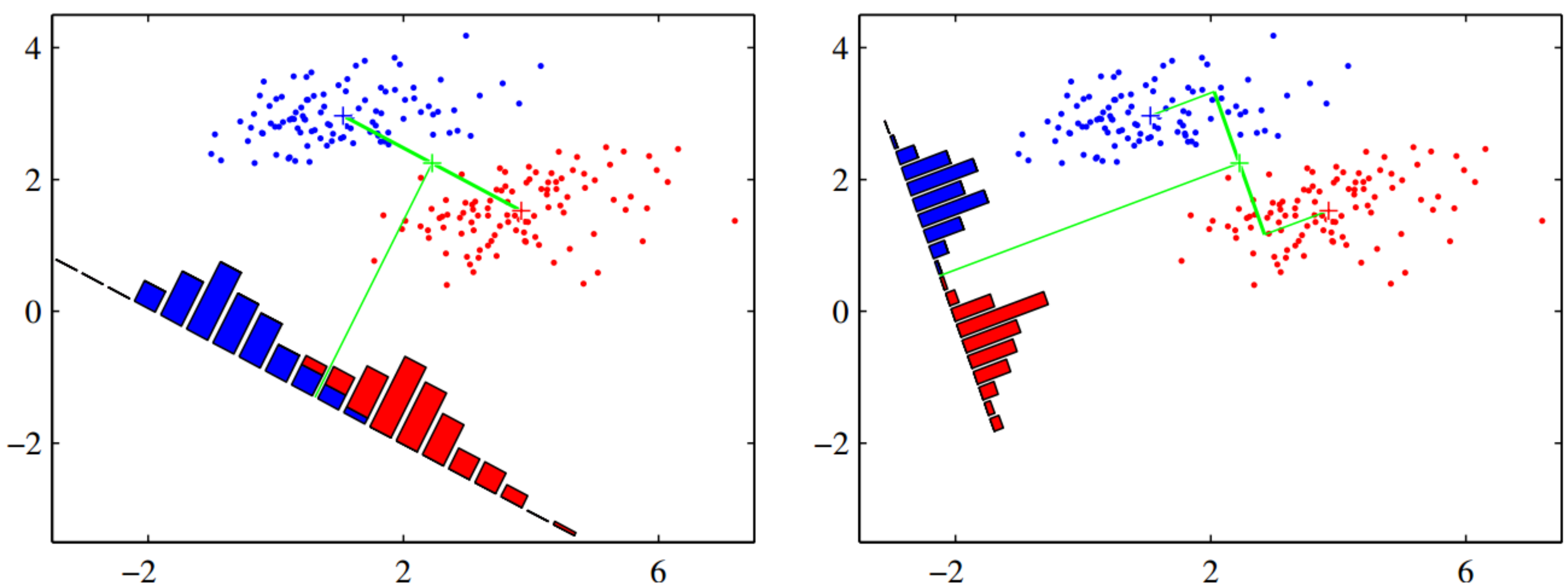
- 主成分分析PCA
  - 二阶独立性

LDA：Linear Discriminant Analysis

- 给定若干样本($\mathbf{x}_i$,c_i)，其中标记只分两类：c_i=0或者c_i=1，设计分类器，将样本分开。
- 方法：
  - Logistic回归/Softmax回归(MaxEnt)
  - SVM
  - 随机森林
  - LDA：Fisher’s linear discriminant

LDA的思路

- 假定两类数据线性可分，即：存在一个超平面，将两类数据分开。则：存在某旋转向量，将两类数据投影到1维，并且可分。



LDA的推导

- 假定旋转向量为**w**，将数据**x**投影到一维y，得到
$$y = \vec{w}^T \vec{x}$$
- 从而，可以方便的找到阈值w_0，$y \geq w_0$时为C_1类，否则为C_2类。

类内均值和方差

- 令C_1有N_1个点，C_2有N_2个点，投影前的类内均值和投影后的类内均值、松散度为：

$$\begin{cases} \vec{m}_1 = \frac{1}{N_1}\sum_{i=1}^{N_1}\vec{x}_i \\ \vec{m}_2 = \frac{1}{N_2}\sum_{i=1}^{N_2}\vec{x}_i \end{cases} \quad \begin{cases} m_1 = w^T\vec{m}_1 \\ m_2 = w^T\vec{m}_2 \end{cases} \quad \begin{cases} s_1^2 = \sum_{i=1}^{N_1}(y_i - m_1)^2 \\ s_2^2 = \sum_{i=1}^{N_2}(y_i - m_2)^2 \end{cases}$$

- 松散度(scatter)，一般称为散列值，是样本松散程度的度量，值越大，越分散。
- 严格的说，m_2应该写成：

$$\vec{m}_2 = \frac{1}{N_2}\sum_{i=N_1+1}^{N_1+N_2}\vec{x}_i$$

Fisher判别准则

- 目标函数：

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

- 向量表示：

$$(m_2 - m_1)^2 = \left(w^T \vec{m}_2 - w^T \vec{m}_1\right)^2$$

$$= \left(w^T (\vec{m}_2 - \vec{m}_1)\right)^2 = \left((\vec{m}_2 - \vec{m}_1)^T w\right)^2$$

$$= \left((\vec{m}_2 - \vec{m}_1)^T w\right)^T \left((\vec{m}_2 - \vec{m}_1)^T w\right)$$

$$= \left(w^T (\vec{m}_2 - \vec{m}_1)\right)\left((\vec{m}_2 - \vec{m}_1)^T w\right)$$

$$= w^T \left((\vec{m}_2 - \vec{m}_1)(\vec{m}_2 - \vec{m}_1)^T\right) w^T$$

$$\xleftrightarrow{\text{令} S_B = (\vec{m}_2 - \vec{m}_1)(\vec{m}_2 - \vec{m}_1)^T} w^T S_B w$$

Fisher判别准则

- 目标函数：

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

- 其中：

$$S_B=(\vec{m}_2-\vec{m}_1)(\vec{m}_2-\vec{m}_1)^T$$

$$S_W=\left(\sum_{i=1}^{N_1}(\vec{x}_i-\vec{m}_1)(\vec{x}_i-\vec{m}_1)^T\right)+\left(\sum_{i=1}^{N_2}(\vec{x}_i-\vec{m}_2)(\vec{x}_i-\vec{m}_2)^T\right)$$

- Within-class scatter matrix
- Between-class scatter
- S_w, S_b可以通过样本计算得到(已知)。

目标函数求极值

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

- 求驻点：

$$\frac{\partial J(\vec{w})}{\partial \vec{w}} = \left(\frac{\vec{w}^T S_B \vec{w}}{\vec{w}^T S_W \vec{w}} \right)'$$

$$= \frac{\left(\vec{w}^T S_B \vec{w}\right)' \left(\vec{w}^T S_W \vec{w}\right) - \left(\vec{w}^T S_W \vec{w}\right)' \left(\vec{w}^T S_B \vec{w}\right)}{\left(\vec{w}^T S_W \vec{w}\right)^2}$$

$$= \frac{2 S_B \vec{w}\left(\vec{w}^T S_W \vec{w}\right) - 2 S_W \vec{w}\left(\vec{w}^T S_B \vec{w}\right)}{\left(\vec{w}^T S_W \vec{w}\right)^2} \overset{令}{=} 0$$

$$\Rightarrow S_B \vec{w}\left(\vec{w}^T S_W \vec{w}\right) = S_W \vec{w}\left(\vec{w}^T S_B \vec{w}\right)$$

$$\Rightarrow S_B \vec{w} \propto S_W \vec{w}$$

Fisher判别投影向量公式

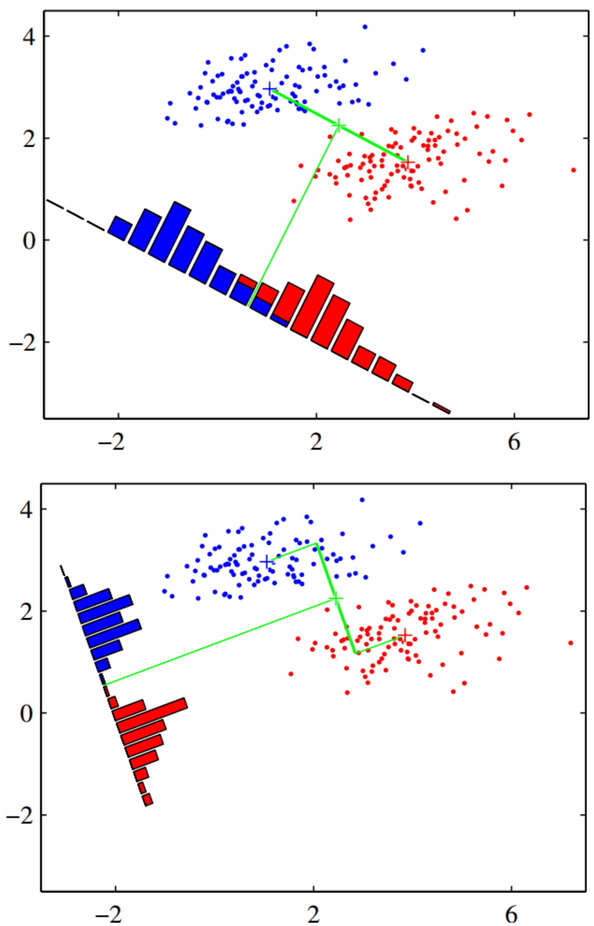
- 以上推导得到 $S_B\vec{w} \propto S_W\vec{w}$
- 根据S_B的计算公式 $S_B = (\vec{m}_2 - \vec{m}_1)(\vec{m}_2 - \vec{m}_1)^T$
- 得：

$$S_B\vec{w} = (\vec{m}_2 - \vec{m}_1)(\vec{m}_2 - \vec{m}_1)^T\vec{w}$$
$$= (\vec{m}_2 - \vec{m}_1)\left((\vec{m}_2 - \vec{m}_1)^T\vec{w}\right) \propto (\vec{m}_2 - \vec{m}_1)$$

- 从而：

$$S_W\vec{w} \propto S_B\vec{w} \propto \vec{m}_2 - \vec{m}_1$$

- 若S_W可逆，则：$\vec{w} \propto S_W^{-1}(\vec{m}_2 - \vec{m}_1)$



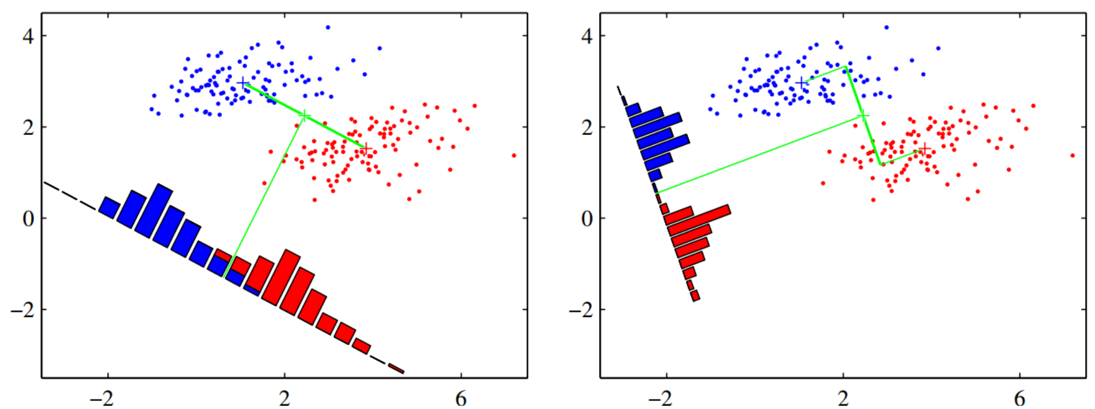

```python
def lda(data):
    n = len(data[0]) - 1
    m1 = [0 for x in range(n)]
    m2 = [0 for x in range(n)]
    m = [0 for x in range(n)]
    number1 = 0
    number2 = 0
    for d in data:
        if d[n] == 1:
            add(m1, d)
            number1 += 1
        elif d[n] == 2:
            add(m2, d)
            number2 += 1
    divide(m1, number1)
    divide(m2, number2)
    print m1,m2

    sw = [[] for x in range(n)]
    for i in range(n):
        sw[i] = [0 for x in range(n)]
    calc_sw(data, sw, m1, 1)
    calc_sw(data, sw, m2, 2)
    normal_matrix(sw)
    print "Sw矩阵: ", sw
    r = linalg.inv(sw)
    print "逆矩阵: ", r
    diff(m1, m2, m)
    normal_vector(m)
    m = multiply(r, m)
    normal_vector(m)
    return m
```

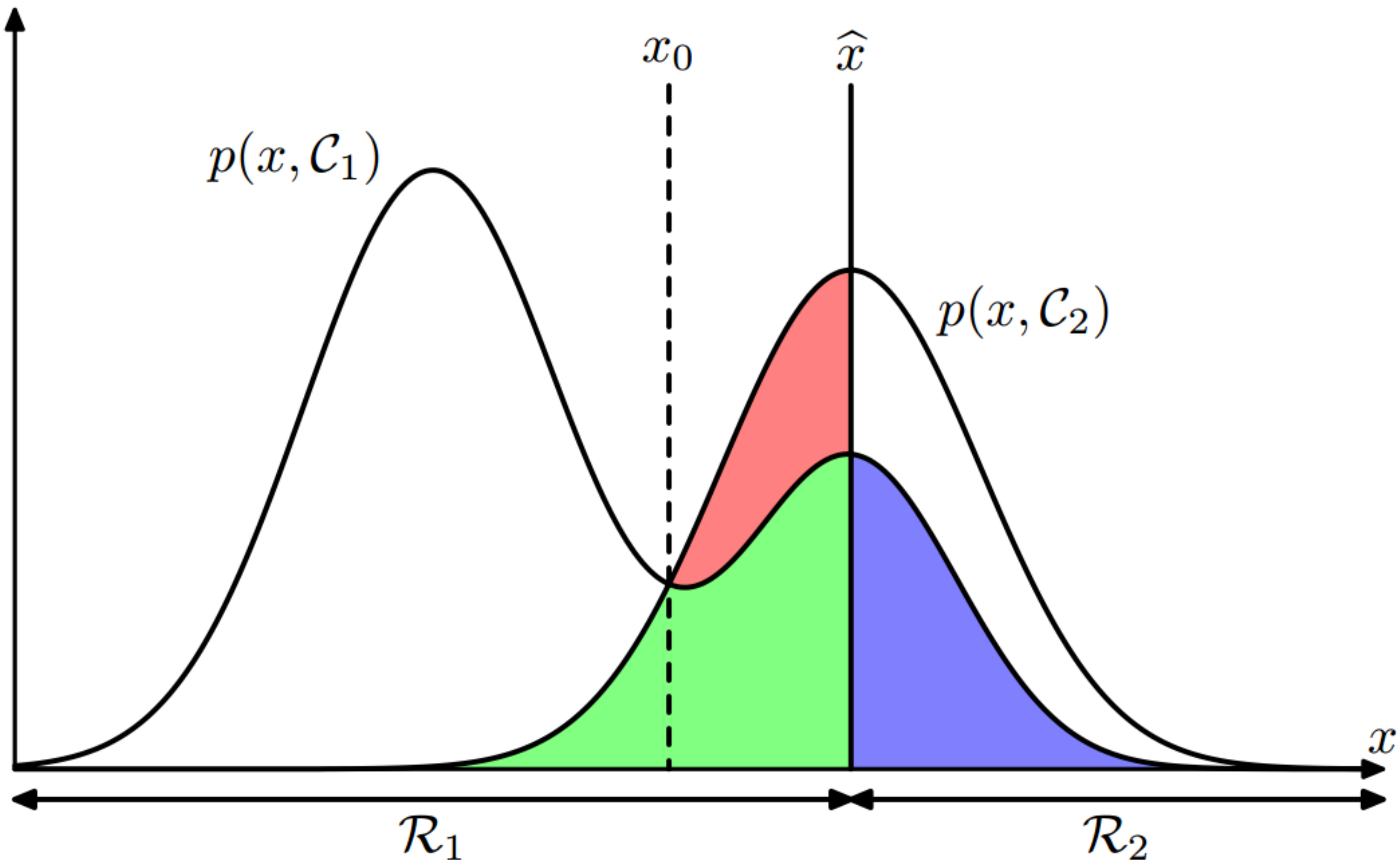
LDA与分类

$$\vec{w} \propto S_W^{-1}(\vec{m}_2 - \vec{m}_1)$$

- 线性判别分析(Fisher's linear discriminant)
  - 严格的说，它只是给出了数据的特定投影方向
- 投影后，数据可以方便的找到阈值w_0，$y \geq w_0$时为C_1类，否则为C_2类。
  - 思考：一维数据下，可以如何分类？

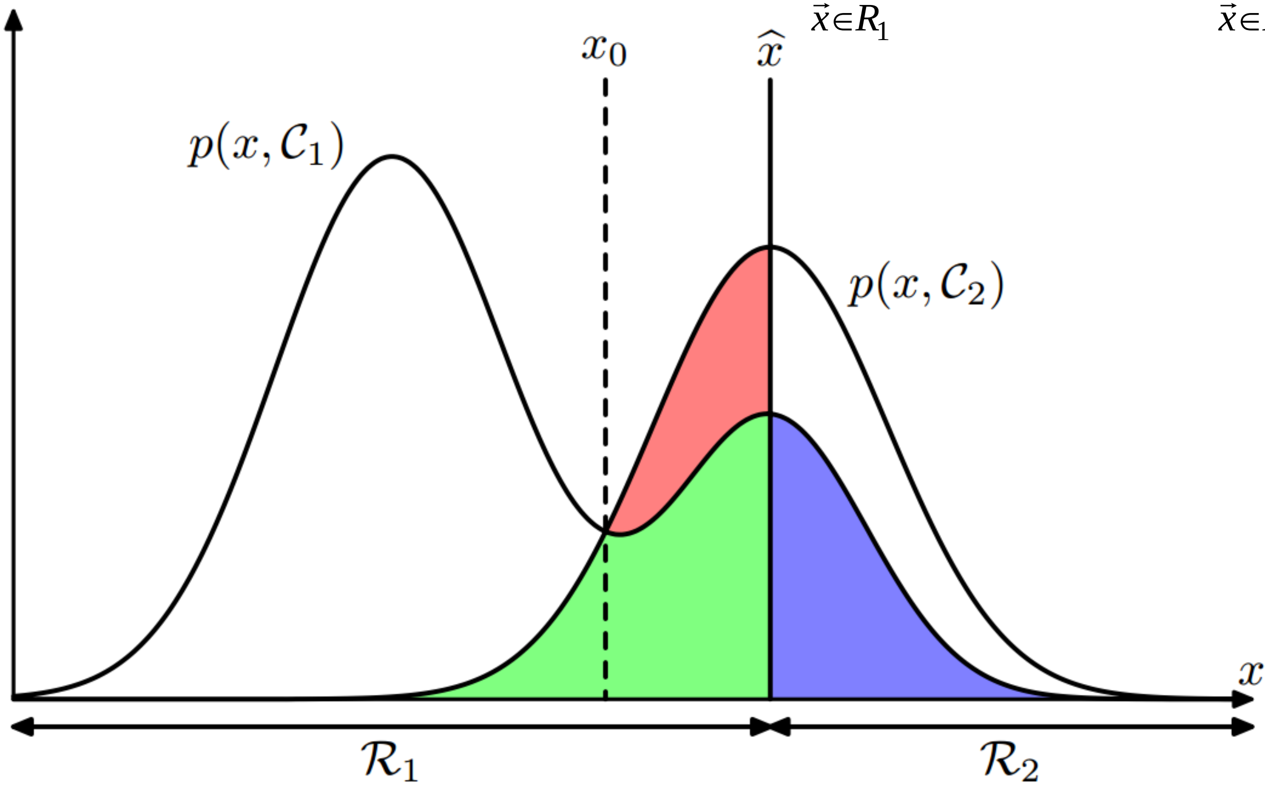


分类step1：极大似然估计

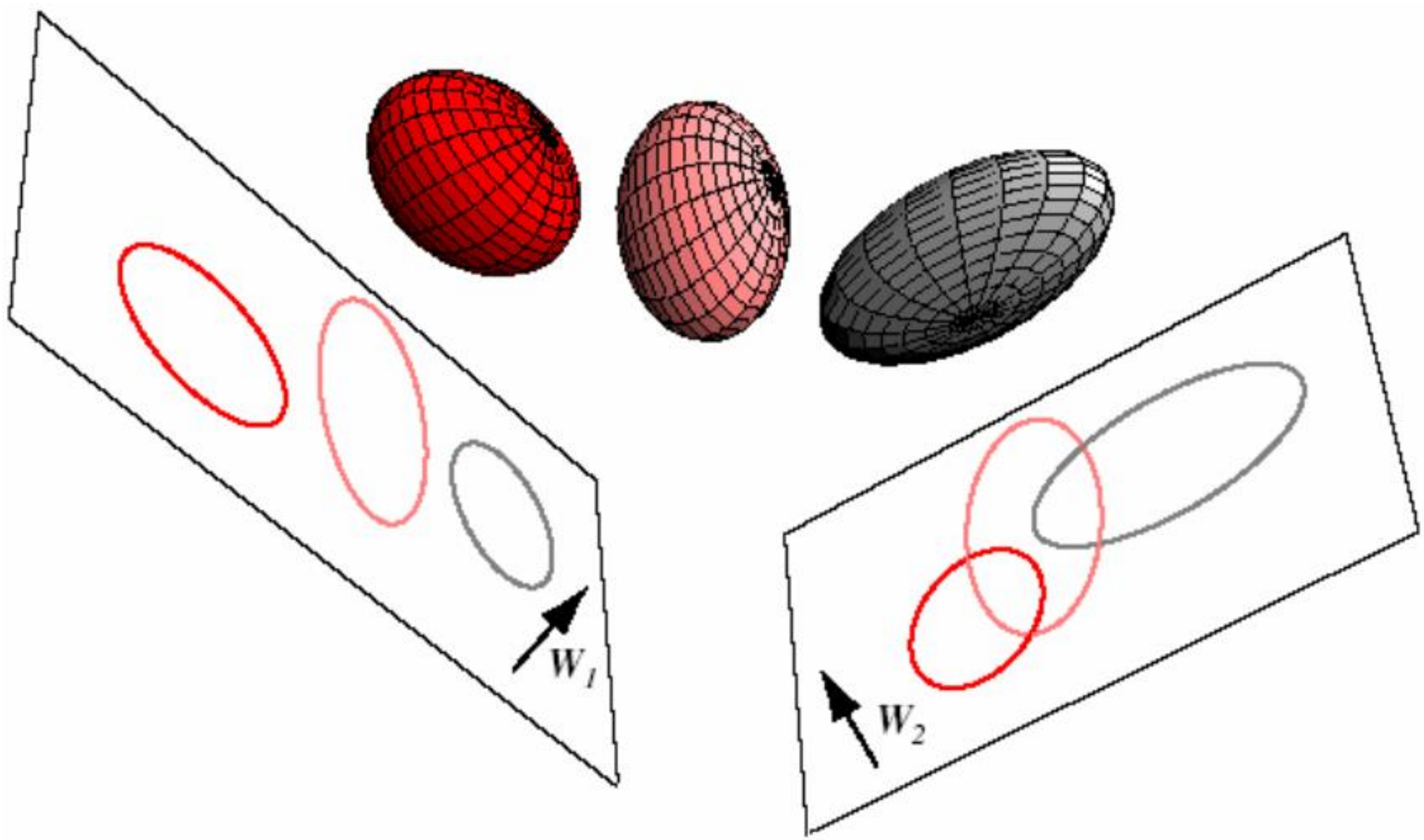


分类step2：误判率准则

$$p(error) = p(\vec{x} \in R_1, C_2) + p(\vec{x} \in R_2, C_1) = \int_{\vec{x} \in R_1} p(\vec{x}, C_2) d\vec{x} + \int_{\vec{x} \in R_2} p(\vec{x}, C_1) d\vec{x}$$

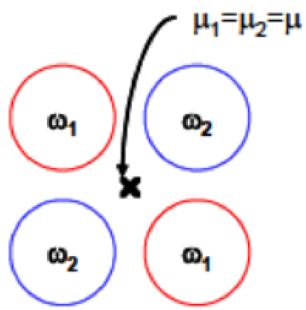


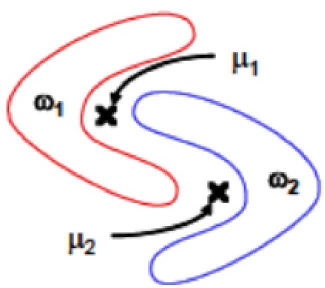
使用LDA将样本投影到平面上

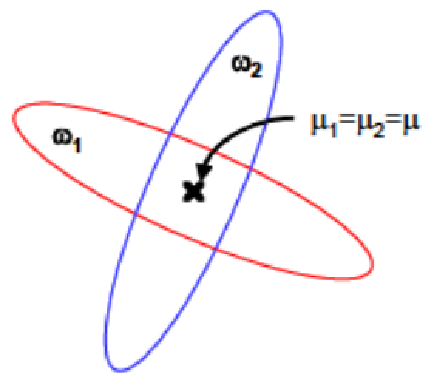


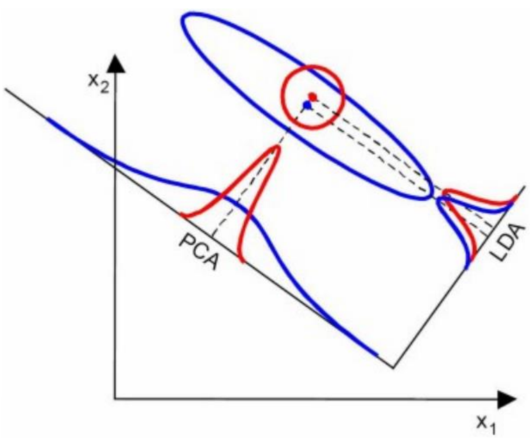
LDA特点

- 由LDA的计算公式看出，LDA是强依赖均值的。如果类别之间的均值相差不大或者需要方差等高阶矩来分类，效果一般。
- 若均值无法有效代表概率分布，LDA效果一般。
  - LDA适用于类别是高斯分布的分类。





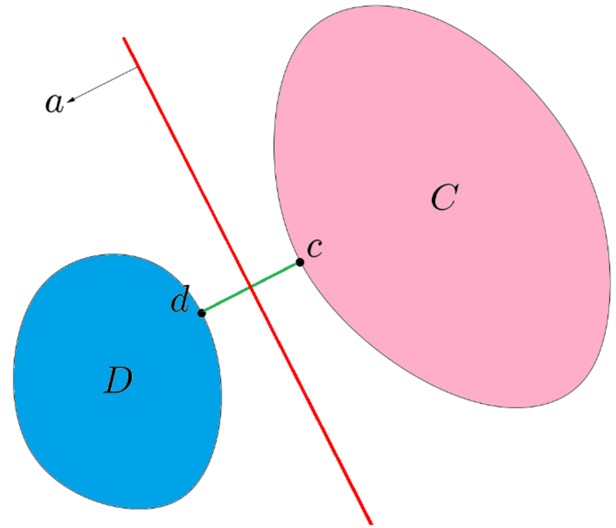




LDA与线性回归的关系

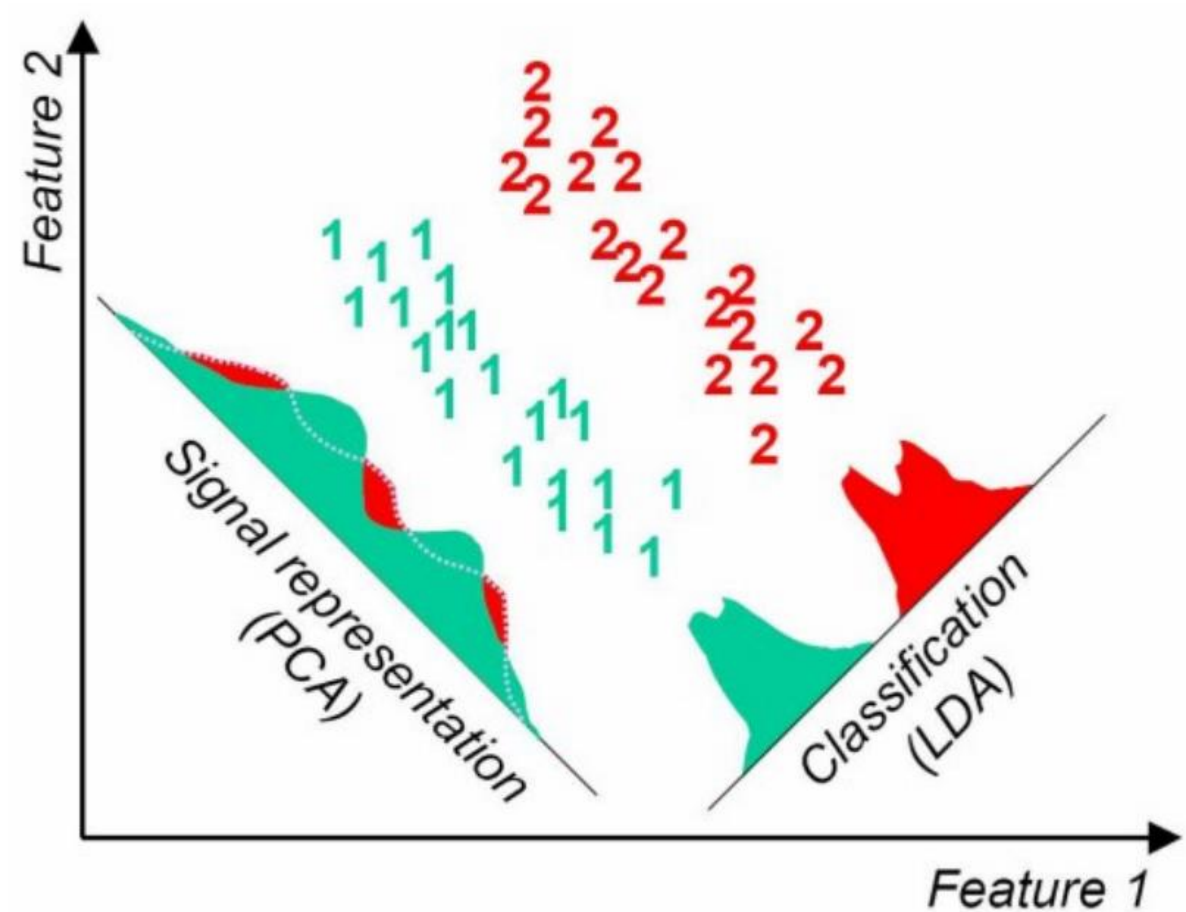
- 假定正样本$(\mathbf{x},1)^{(i)}$个数为N_1，负样本$(\mathbf{x},-1)^{(i)}$个数为N_2；将标记加权成正样本$(\mathbf{x},1/N_1)^{(i)}$，负样本$(\mathbf{x},-1/N_2)^{(i)}$，则使用线性回归得到的决策面方向与LDA相同。

$$\begin{cases}(x,1)^{(i)} \Rightarrow \left(x,\dfrac{1}{N_1}\right)^{(i)} \\ (x,-1)^{(i)} \Rightarrow \left(x,-\dfrac{1}{N_2}\right)^{(i)}\end{cases}$$



LDA与PCA

- LDA：
  - 分类性能最好的方向
- PCA：
  - 样本点投影具有最大方差的方向



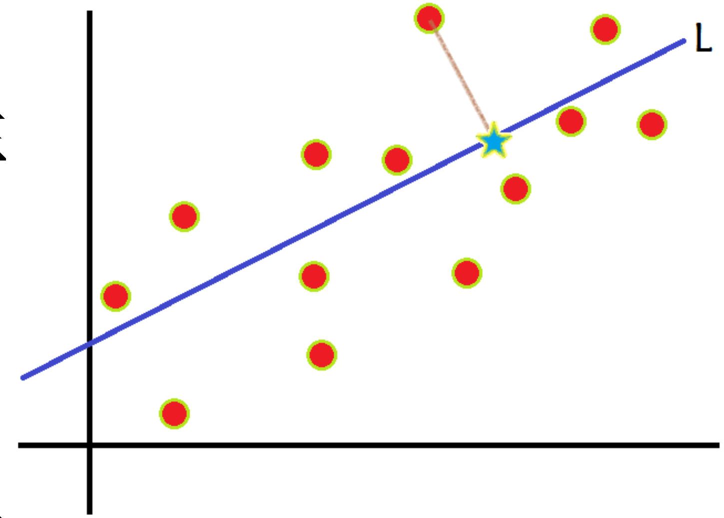
复习：实对称阵特征向量的正交性

- 实对阵矩阵的不同特征值对应的特征向量一定是正交的
- 证明：
- 令实对称矩阵为A，它的两个不同的特征值$\lambda_1\lambda_2$对应的特征向量分别是$\mu_1\mu_2$
- 则有：$A\mu_1=\lambda_1\mu_1$， $A\mu_2=\lambda_2\mu_2$
- $(A\mu_1)^T=(\lambda_1\mu_1)^T$，从而：$\mu_1^TA=\lambda_1\mu_1^T$
- 所以：$\mu_1^TA\mu_2=\lambda_1\mu_1^T\mu_2$
- 同时，$\mu_1^TA\mu_2=\mu_1^T(A\mu_2)=\mu_1^T\lambda_2\mu_2=\lambda_2\mu_1^T\mu_2$
- 所以，$\lambda_1\mu_1^T\mu_2=\lambda_2\mu_1^T\mu_2$
- 故：$(\lambda_1-\lambda_2)\mu_1^T\mu_2=0$
- 而$\lambda_1\neq\lambda_2$，所以$\mu_1^T\mu_2=0$，即：μ_1，μ_2正交。

问题的提出

- 实际问题往往需要研究多个特征，而这些特征存在一定的相关性。
  – 数据量增加了问题的复杂性。
- 将多个特征综合为少数几个代表性特征：
  – 既能够代表原始特征的绝大多数信息，
  – 组合后的特征又互不相关，降低相关性。
  – 主成分
- 即主成分分析。

考察降维后的样



- 对于n个特征的m个样本，将每个样本写成行向量，得到矩阵A

$$A=\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m-1,1} & a_{m-1,2} & \cdots & a_{m-1,n} \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}=\begin{pmatrix} a_1^T \\ a_2^T \\ \vdots \\ a_{m-1}^T \\ a_m^T \end{pmatrix}$$

- 思路：寻找样本的主方向**u**：将m个样本值投影到某直线L上，得到m个位于直线L上的点，计算m个投影点的方差。认为方差最大的直线方向是主方向。
  - 假定样本是去均值化的；若没有去均值化，则计算m个样本的均值，将样本真实值减去均值。

计算投影样本点的方差

- 取投影直线L的延伸方向**u**，计算A×**u**的值

$$A\cdot u=\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m-1,1} & a_{m-1,2} & \cdots & a_{m-1,n} \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}\cdot u=\begin{pmatrix} a_1^T \\ a_2^T \\ \vdots \\ a_{m-1}^T \\ a_m^T \end{pmatrix}\cdot u=\begin{pmatrix} a_1^T\cdot u \\ a_2^T\cdot u \\ \vdots \\ a_{m-1}^T\cdot u \\ a_m^T\cdot u \end{pmatrix}$$

- 求向量A×**u**的方差

$$Var(A\cdot u)=(Au-E)^T(Au-E)=(Au)^T(Au)=u^TA^TAu$$

- 目标函数：

$$J(u)=\frac{1}{2}u^TA^TAu$$

目标函数

$$J(u)=\frac{1}{2}u^T A^T Au$$

- 由于u数乘得到的方向和u相同，因此，增加u是单位向量的约束，即：$\|u\|_2=1$
- 从而： $\|u\|_2=1 \Rightarrow u^T u=1$
- 建立Lagrange方程：

$$L(u)=\frac{1}{2}u^T A^T Au-\lambda\left(u^T u-1\right)$$

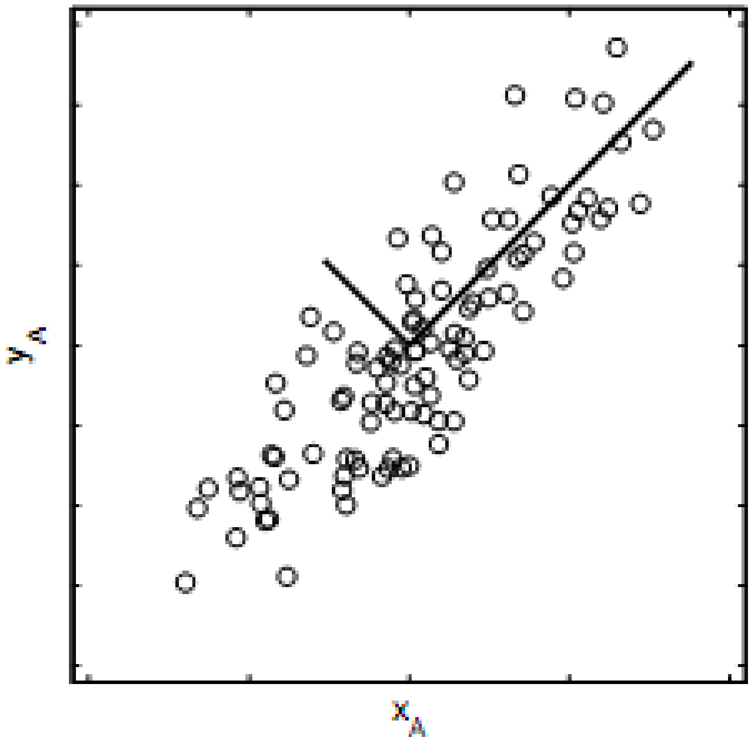
$$\frac{\partial L(u)}{\partial u}=A^T Au-\lambda u\overset{令}{=}0 \Rightarrow \left(A^T A\right)u=\lambda u$$

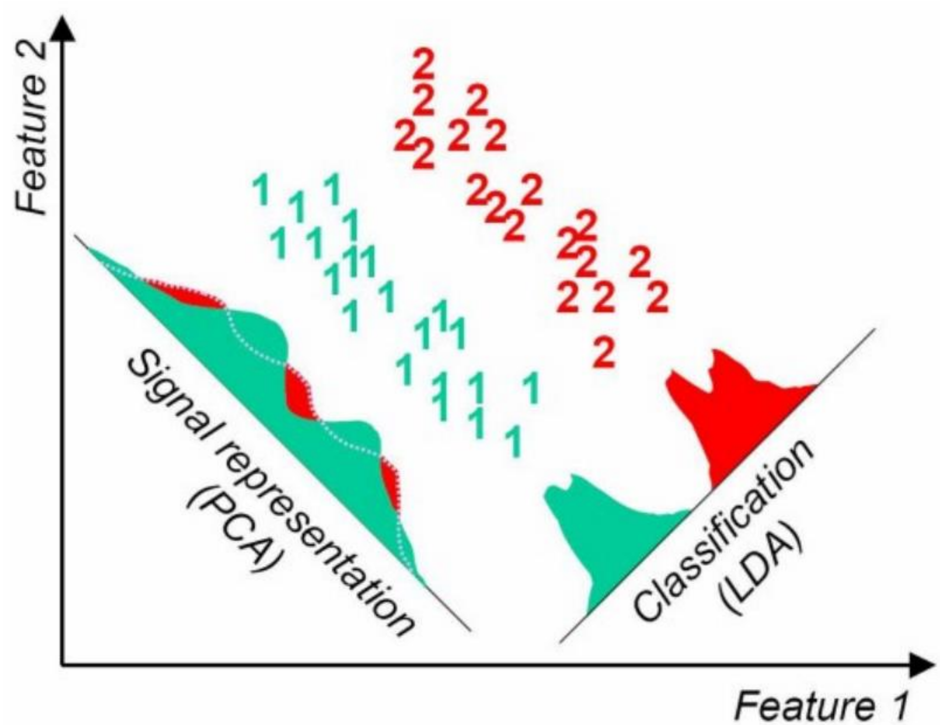
方差和特征值

$$A^T Au = \lambda u$$

- 若A中的样本都是去均值化的，则A^TA与A的协方差矩阵仅相差系数n-1
  - A^TA常常称为散列矩阵(scatter matrix)
- 根据上式，**u**是A^TA的一个特征向量，λ的值的大小为原始观测数据的特征在向量**u**的方向上投影值的方差。
- 以上即为主成分分析PCA的核心推导过程。

PCA的两个特征向量

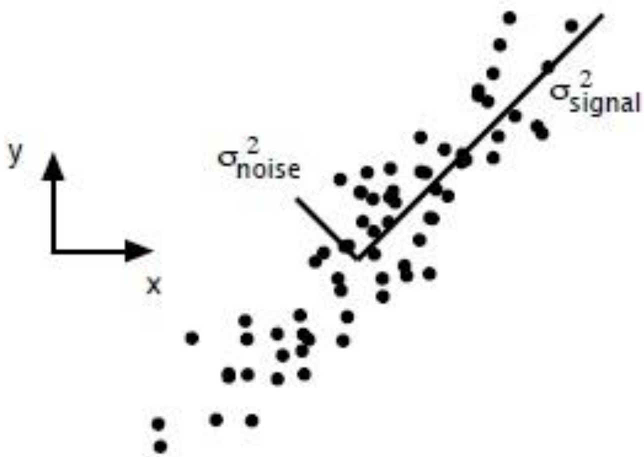

y_A
x_A

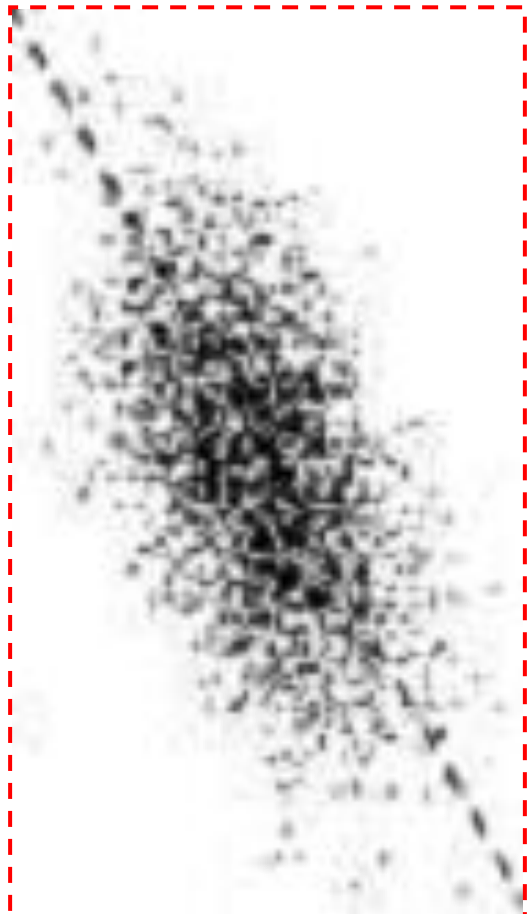

Feature 2
Feature 1
Signal representation
(PCA)
Classification
(LDA)

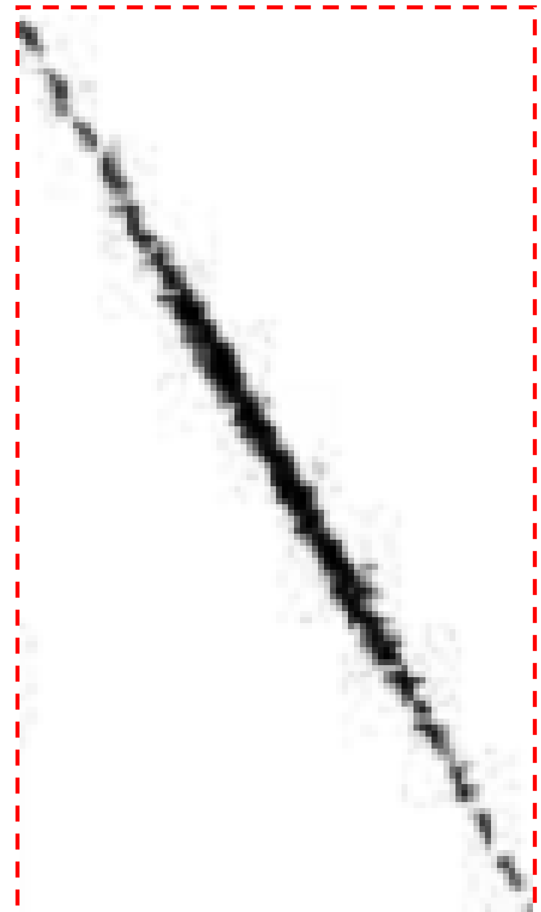
PCA的重要应用

- OBB树
  - Oriented Bounding Box
  - GIS中的空间索引
- 特征提取
- 数据压缩
  - 降维
  - 对原始观测数据A在λ值前k大的特征向量u上投影后，获得一个A(m×n)Q(n×k) 的序列，再加上特征向量矩阵Q，即将A原来的m×n个数据压缩到m×k＋k×n个数据。

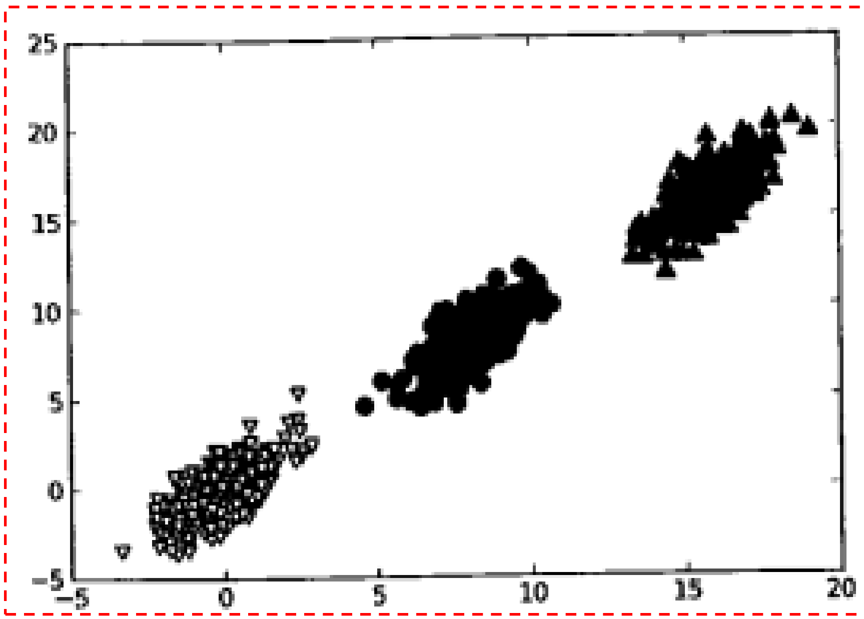
PCA的重要应用——去噪

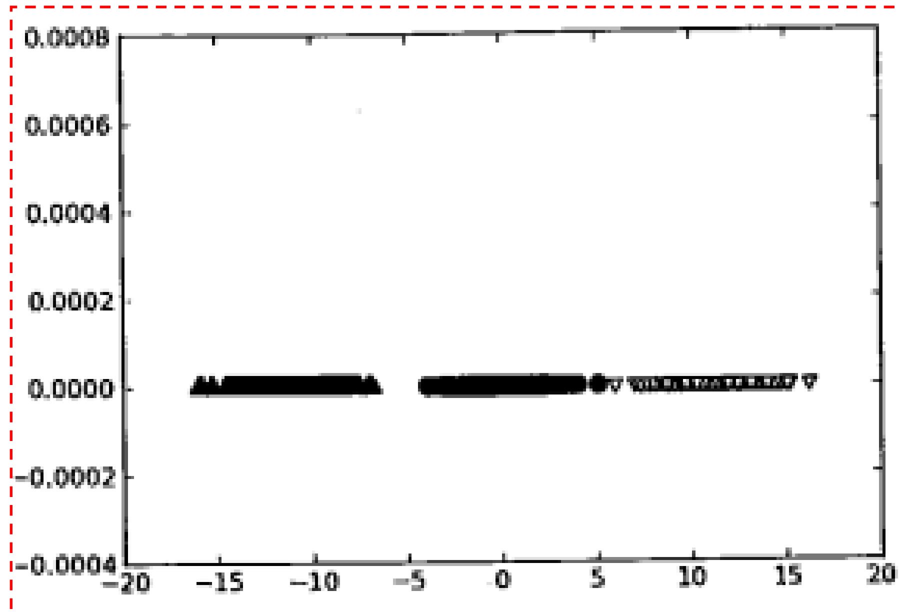






PCA的重要应用——降维


25
20
15
10
5
0
−5
−5
0
5
10
15
20


0.0008
0.0006
0.0004
0.0002
0.0000
−0.0002
−0.0004
−20
−15
−10
−5
0
5
10
15
20

PCA总结

- 实对称阵的特征值一定是实数，不同特征值对应的特征向量一定正交，重数为r的特征值一定有r个线性无关的特征向量；
- 样本矩阵的协方差矩阵必然一定是对称阵，协方差矩阵的元素即各个特征间相关性的度量；
  - 具体实践中考虑是否去均值化；
- 将协方差矩阵C的特征向量组成矩阵P，可以将C合同为对角矩阵D，对角阵D的对角元素即为A的特征值。
  - $P^TCP=D$
  - 协方差矩阵的特征向量，往往单位化，即特征向量的模为1，从而，P是标准正交阵：$P^TP=I$。
  - 即将特征空间线性加权，使得加权后的特征组合间是不相关的。选择若干最大的特征值对应的特征向量(即新的特征组合)，即完成了PCA的过程。