中国石油大学
CHINA UNIVERSITY OF PETROLEUM

《油气人工智能基础及应用》

1.2 特征提取算法

董少群
dshaoqun@163.com
理学院数学系

CUP
厚积薄发 开物成务

特征提取

中国石油大学
CHINA UNIVERSITY OF PETROLEUM

Feature extraction refers to the process of transforming raw data into numerical features that can be processed while preserving the information in the original data set. It yields better results than applying machine learning directly to the raw data.

将原始数据转换为可处理的数值特征，同时保留原始数据集中的信息。

Feature extraction starts from an initial set of measured data and builds derived values (features) intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps, and in some cases leading to better human interpretations. Feature extraction is related to dimensionality reduction.

特征提取从一组初始测量数据开始，并构建旨在提供信息和非冗余的派生值（特征），从而促进后续的学习和泛化步骤

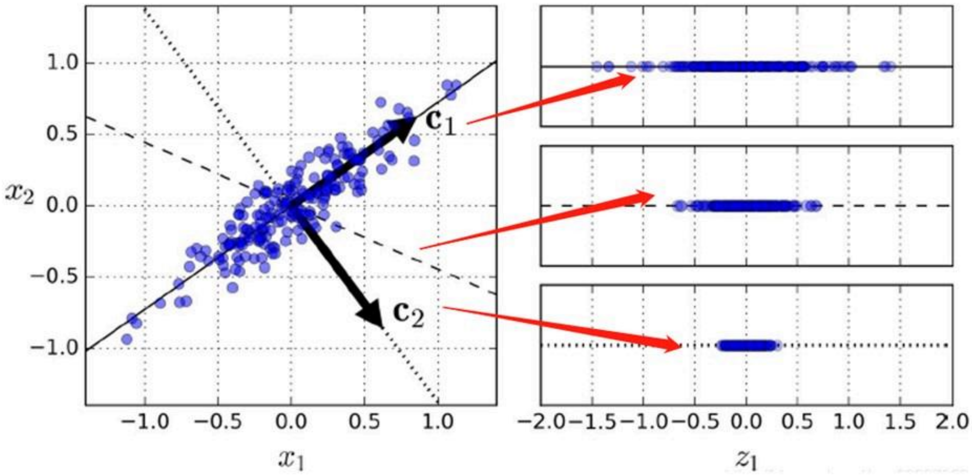
The aim of feature extraction is to find the most compacted and informative set of features (distinct patterns) to enhance the efficiency of the classifier.

特征提取的目的是找到最紧凑和信息最丰富的特征集（不同的模式），以提高分类器的效率。

2

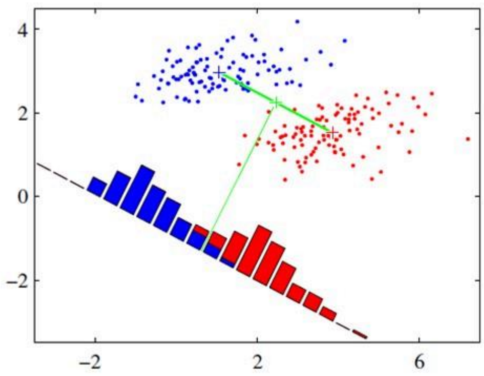
特征提取

将原始数据转换为可处理的数值特征，同时保留原始数据集中的信息。

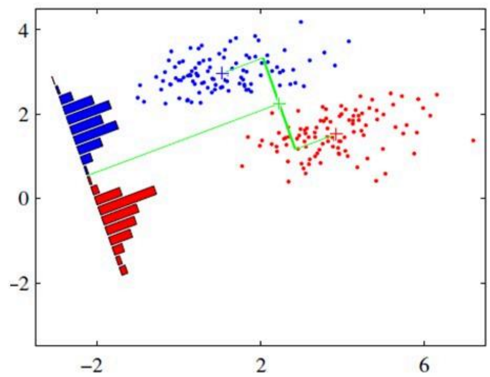


特征提取

将原始数据转换为可处理的数值特征，同时保留原始数据集中的信息。



线性 无监督



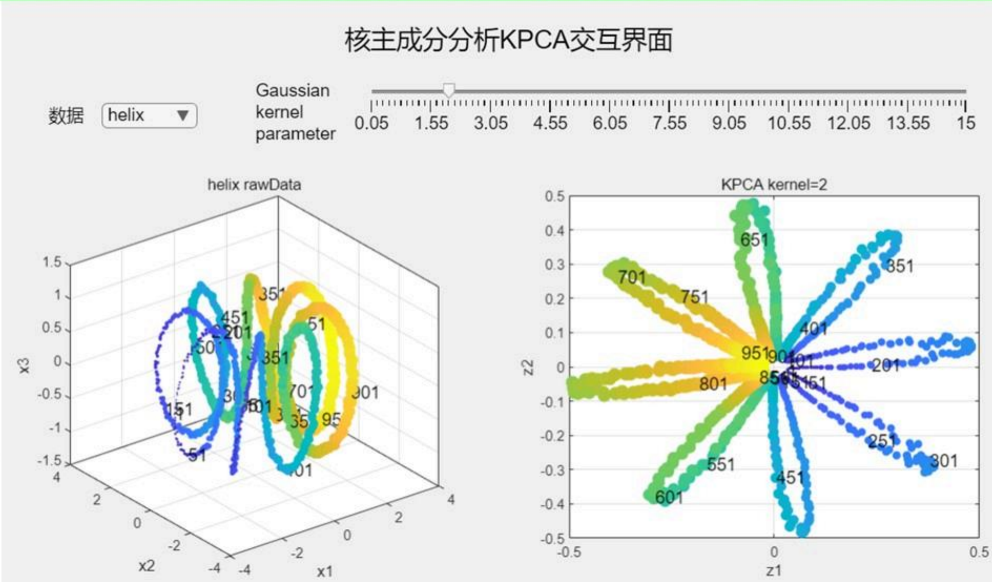
线性 有监督

特征提取的目的是找到最紧凑和信息最丰富的特征集（不同的模式），以提高分类器的效率。

特征提取

中国石油大学 CHINA UNIVERSITY OF PETROLEUM

将原始数据转换为可处理的数值特征，同时保留原始数据集中的信息。



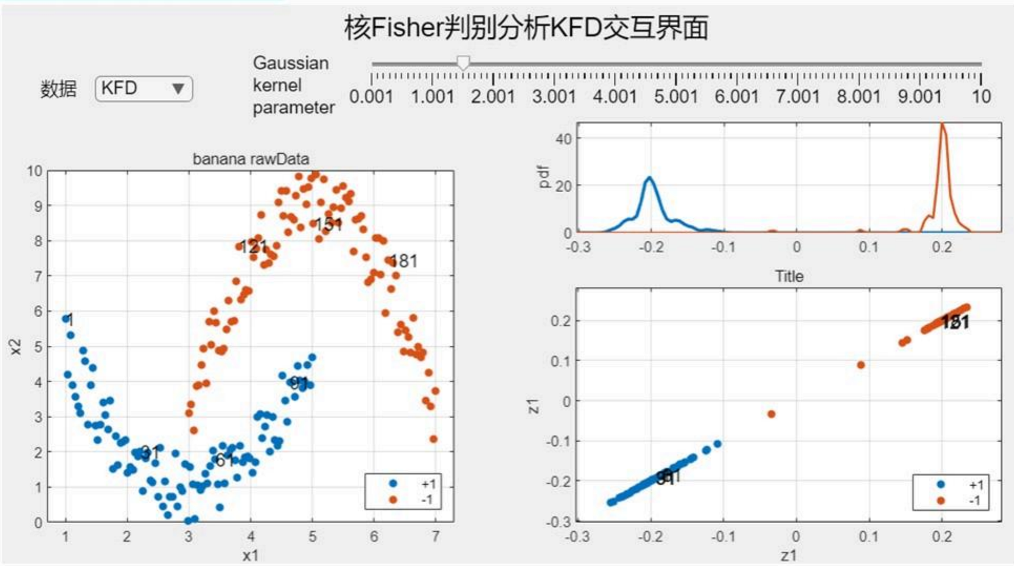
非线性
无监督

5

特征提取

中国石油大学 CHINA UNIVERSITY OF PETROLEUM

特征提取的目的是找到最紧凑和信息最丰富的特征集（不同的模式），以提高分类器的效率。



非线性
有监督

6

一、方法概述

- 特征提取
  - 无监督
    - 线性方法 — 主成分分析（PCA）
    - 核方法 — 核主成分分析（KPCA）
    - 基于流形学习
      - 拉普拉斯特征映射（LE）
      - 局部线性嵌入（LLE）
      - 等容特征映射（ISOMAP）
  - 有监督
    - 线性方法 — 线性判别分析法（LDA）
    - 核方法 — 核Fisher判别分析（KFD）
    - 基于流形学习 — 监督局部线性嵌入（SLLE）

7

二、主成分分析（Principal Component Analysis， PCA）

- 由英国统计学家Pearson在1901年提出
- 后来由Hotelling在1933年加以发展提出的一种多变量的统计方法。
- 它是一种使用最广泛的数据降维算法。其主要思想是将n维特征映射到k维上，这k维是全新的正交特征也被称为主成分，是在原有n维特征的基础上重新构造出来的k维特征。

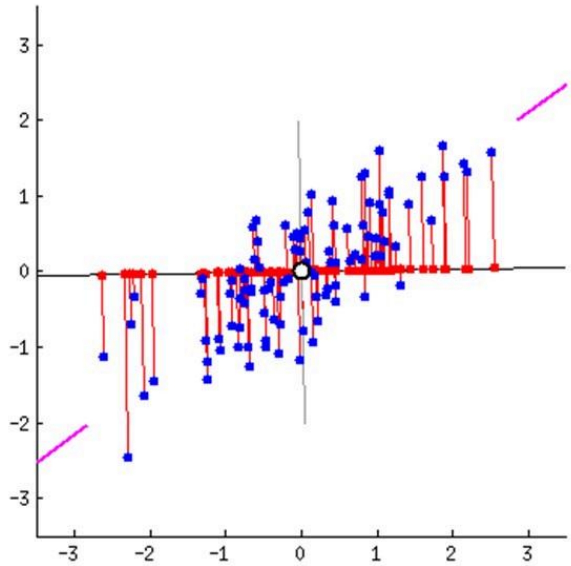
优点：（1）使得数据集更易使用；
（2）降低算法的计算开销；
（3）使得结果容易理解。
缺点：（1）特征值分解有一些局限性，比如变换的矩阵必须是方阵；
（2）在非高斯分布情况下，PCA得出的主元可能并不是最优的。

9

二、主成分分析（PCA）

中国石油大学 CHINA UNIVERSITY OF PETROLEUM

- 在PCA中，数据从原来的坐标系转换到新的坐标系
- 在新的坐标系中要求数据之间的协方差为0（即不同维度之间的数据线性不相关），而方差尽可能的大。
- 因此，第一坐标轴是原始数据中方差最大的方向，第二个坐标轴是与第一个新坐标轴正交且方差次大的方向，重复该过程，直到需要的降维数。



10

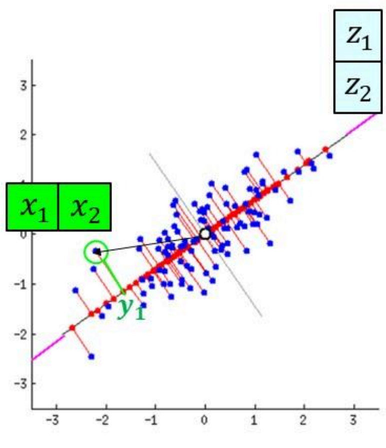
二、主成分分析（PCA）

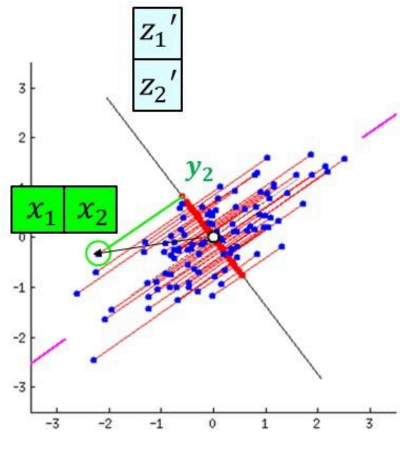
中国石油大学 CHINA UNIVERSITY OF PETROLEUM

$$\boldsymbol{x}\ \begin{bmatrix} x_1 & x_2 \end{bmatrix} \times \overset{\boldsymbol{v}_1}{\begin{bmatrix} z_1 \\ z_2 \end{bmatrix}} = \begin{bmatrix} y_1 \end{bmatrix}$$

$$\boldsymbol{x}\ \begin{bmatrix} x_1 & x_2 \end{bmatrix} \times \overset{\boldsymbol{v}_2}{\begin{bmatrix} z_1' \\ z_2' \end{bmatrix}} = \begin{bmatrix} y_2 \end{bmatrix}$$

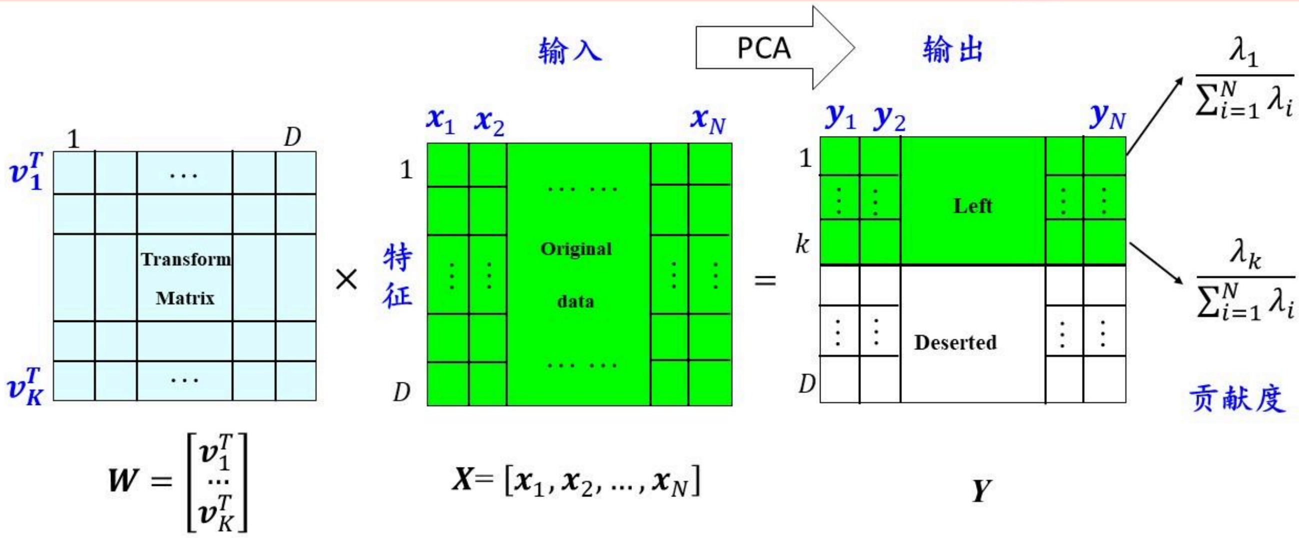
$$\boldsymbol{x}\ \begin{bmatrix} x_1 & x_2 \end{bmatrix} \times \overset{\boldsymbol{v}_1\ \ \boldsymbol{v}_2}{\begin{bmatrix} z_1 & z_1' \\ z_2 & z_2' \end{bmatrix}} = \begin{bmatrix} y_1 & y_2 \end{bmatrix}$$





11

二、主成分分析（PCA）



二、主成分分析（PCA）

求方差最大转化为下面的优化模型

$$\begin{aligned}&max\ \ \sigma^2 = \boldsymbol{v}^T\boldsymbol{C}\boldsymbol{v}\\ &s.t.\ \ \boldsymbol{v}^T\boldsymbol{v}=1\end{aligned}$$

再根据拉格朗日乘子法，我们得到相应的拉格朗日函数：

$$F(\boldsymbol{v}) = -\boldsymbol{v}^T\boldsymbol{C}\boldsymbol{v} - \lambda(1-\boldsymbol{v}^T\boldsymbol{v})$$

求偏导之后得到的方程如下

$$\frac{\partial}{\partial \boldsymbol{v}}F(\boldsymbol{v}) = 0 \Rightarrow 2\boldsymbol{C}\boldsymbol{v} - 2\lambda\boldsymbol{v} = 0 \Rightarrow \boldsymbol{C}\boldsymbol{v} = \lambda\boldsymbol{v}$$

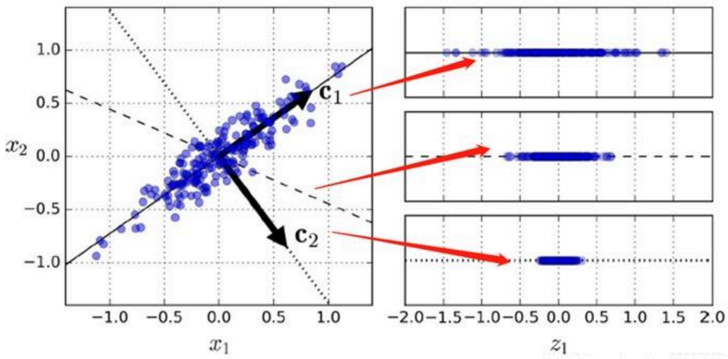
求出协方差矩阵的特征值及对应的特征向量

二、主成分分析（PCA）

$$\boldsymbol{v} = \underbrace{argmax}_{\boldsymbol{v}}\, \sigma^2$$

$$\sigma^2 = \boldsymbol{v}^T \boldsymbol{C} \boldsymbol{v} = \boldsymbol{v}^T \lambda \boldsymbol{v} = \lambda \boldsymbol{v}^T \boldsymbol{v} = \lambda$$

其中$\boldsymbol{C} = \frac{1}{N}\boldsymbol{X}\boldsymbol{X}^T$



特征值求出来，然后按大小排个序，选出最大的几个特征值，并求出对应的特征向量，最后用这几个特征向量来完成数据集在其上的投影，这样就完成了特征的筛选。

14

二、主成分分析（PCA）

$K\times N \qquad N\times K$

$$\boldsymbol{C} = \frac{1}{N}\boldsymbol{X}\boldsymbol{X}^T = \frac{1}{N}[\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N]\begin{bmatrix}\boldsymbol{x}_1^T \\ \cdots \\ \boldsymbol{x}_N^T\end{bmatrix}$$

$$= \frac{1}{N}\begin{bmatrix}\sum_{i=1}^N \boldsymbol{x}_i(1)\boldsymbol{x}_i(1) & \cdots & \sum_{i=1}^N \boldsymbol{x}_i(1)\boldsymbol{x}_i(D) \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^N \boldsymbol{x}_i(D)\boldsymbol{x}_i(1) & \cdots & \sum_{i=1}^N \boldsymbol{x}_i(D)\boldsymbol{x}_i(D)\end{bmatrix}$$

15

二、主成分分析（PCA）

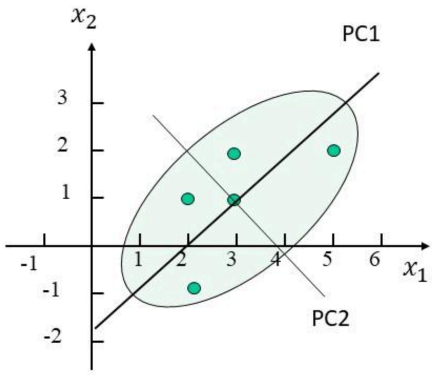
$$\boldsymbol{X}_{2*5} = \begin{bmatrix} 2 & 2 & 3 & 5 & 3 \\ -1 & 1 & 1 & 2 & 2 \end{bmatrix}^T$$

1.均值化： $\boldsymbol{X} = \boldsymbol{X}_{2*5} - \overline{\boldsymbol{X}}_{2*5} = \begin{bmatrix} -1 & -1 & 0 & 2 & 0 \\ -2 & 0 & 0 & 1 & 1 \end{bmatrix}^T$

2.协方差矩阵： $\boldsymbol{C} = \dfrac{1}{5-1}\boldsymbol{X}\boldsymbol{X}^T = \begin{bmatrix} 3/2 & 1 \\ 1 & 3/2 \end{bmatrix}$

3.特征值为： $\lambda_1 = 5/2 \qquad \lambda_2 = 1/2$

对应的特征向量为：$\boldsymbol{w}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ $\boldsymbol{w}_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$



Python: numpy.linalg的eig ( )函数
Matlab: eig ( )函数

16

二、主成分分析（PCA）

中国石油大学 CHINA UNIVERSITY OF PETROLEUM

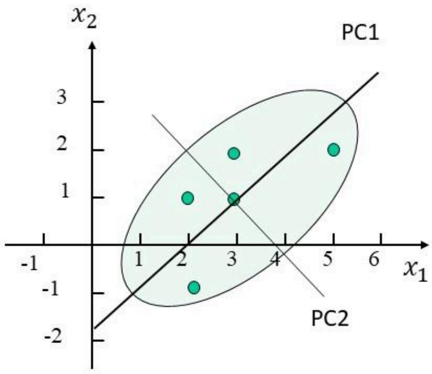
特征向量单位化 $\boldsymbol{v}_1 = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$ $\boldsymbol{v}_2 = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix}$

4.此时，PCA转换矩阵： $\boldsymbol{W} = [\boldsymbol{v}_1 \quad \boldsymbol{v}_2]$

5.按照特征值大小排序，计算$\boldsymbol{v}_1$、$\boldsymbol{v}_2$**贡献度 83.3%、16.7%**

$\boldsymbol{v}_1$、$\boldsymbol{v}_2$**的累积贡献度 83.3%、100%**

$\boldsymbol{W} = [\boldsymbol{v}_1]$



6.这5个样本降维后特征$\boldsymbol{Y}$为：

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{W} = \begin{bmatrix} -1 & -1 & 0 & 2 & 0 \\ -2 & 0 & 0 & 1 & 1 \end{bmatrix}^T \left[1/\sqrt{2}, 1/\sqrt{2}\right]^T$$
$$= \left[-3/\sqrt{2} \quad -1/\sqrt{2} \quad 0 \quad 3/\sqrt{2} \quad 1/\sqrt{2}\right]$$

17

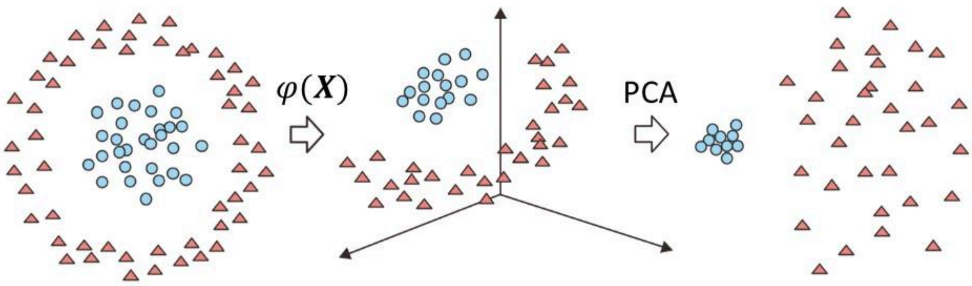
三、核主成分分析（Kernel PCA, KPCA）

- KPCA是一种降低数据维度的有效技术，由马克斯普朗克智能系统研究所的Scholkopf教授于1998年提出。
- 它是一种非线性数据处理方法，其核心思想是通过一个非线性映射把原始空间的数据投影到高维特征空间，然后在高维特征空间中进行基于PCA的数据处理。

优点：
（1）能够处理线性不可分数据；
（2）支持多种核函数 (linear, gaussian, polynomial, sigmoid, laplacian)。
缺点：
核的不同选择影响效果。

三、核主成分分析（KPCA）

中国石油大学
CHINA UNIVERSITY OF PETROLEUM

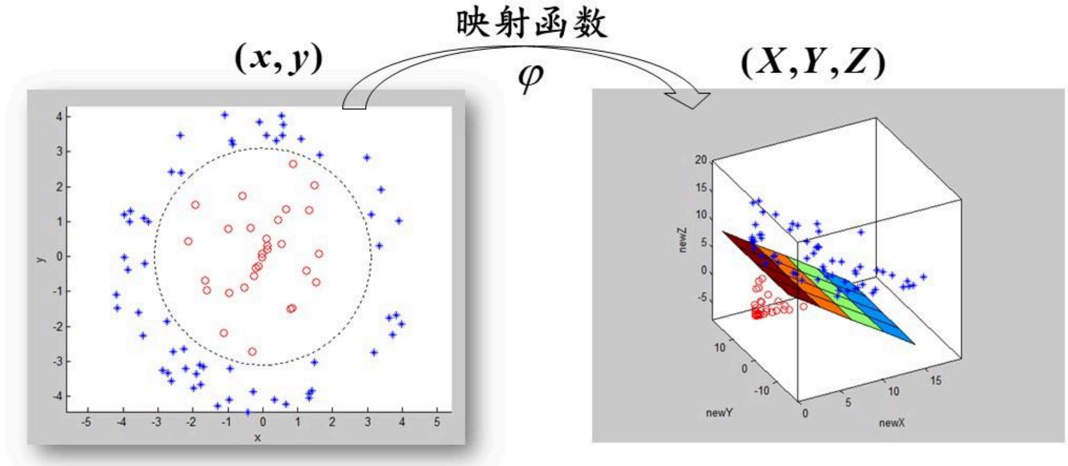


- ✓ 根据Cover定理，适当的非线性映射到高维非线性特征空间可以使这个问题线性可分。
- ✓ 提高维数的关键在于确定非线性映射函数$\varphi(X)$。
- ✓ 不幸的是，要获得一个合适的显式非线性映射函数是相当困难的。
- ✓ 然而，为了解决显式映射函数的确定问题，核方法利用核函数直接计算高维特征空间中所有对数据样本之间的内乘积，避免了非线性特征映射的显式计算，称为“核技巧”。即实现向量的内积变换：

三、核主成分分析（KPCA）

核方法巧妙地解决了求取映射函数的问题，不直接显示的求取，而是通过**核函数**隐式的表示

$$\kappa\left(X_i, X_j\right) = \varphi^T\left(X_i\right)\varphi\left(X_j\right)$$



三、核主成分分析（KPCA）

数据降维:核主成分分析(Kernel PCA)原理解析

https://zhuanlan.zhihu.com/p/59775730

三、核主成分分析（KPCA）

中国石油大学 CHINA UNIVERSITY OF PETROLEUM

以往在矩阵X中，我们用一行表示一个样本点。而现在我们需要使用映射，由于向量通常是指列向量x_i，所以在KPCA相关的文献中，常常用 的每一列来表示一个样本，即$\boldsymbol{X}=[\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N]$

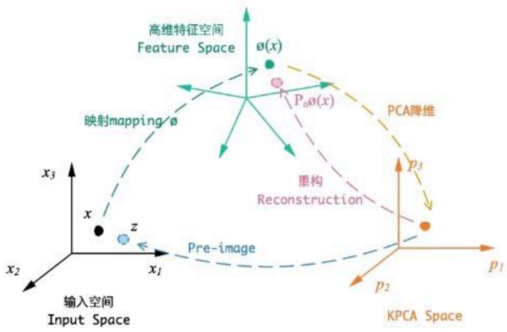
X_i为K维列向量，X中共有N个样本

$K\times N$的矩阵X所在的空间称为输入空间(Input space)

现在用一个非线性映射$\boldsymbol{\phi}$将X中的向量x_i 映射到高维空间(记为D维):

$$\boldsymbol{\phi}(\boldsymbol{x}): R^K \longrightarrow R^D,\ K \ll D$$

这个高维空间称为特征空间(Feature space)，记为$\mathcal{F}$。



25

三、核主成分分析（KPCA）

中国石油大学 CHINA UNIVERSITY OF PETROLEUM

$\boldsymbol{\lambda} \neq \boldsymbol{0}$时，特征空间中的主成分的方向

$$\boldsymbol{p} = \frac{1}{\lambda}\sum_{i=1}^{N} \boldsymbol{\phi}(\boldsymbol{x}_i)\boldsymbol{\phi}(\boldsymbol{x}_i)^T\boldsymbol{p} = \frac{1}{\lambda}\sum_{i=1}^{N} \boldsymbol{\phi}(\boldsymbol{x}_i)[\boldsymbol{\phi}(\boldsymbol{x}_i)^T\boldsymbol{p}]$$

$$\underbrace{\boldsymbol{p}}_{D\times 1} = \sum_{i=1}^{N} \alpha_i \boldsymbol{\phi}(\boldsymbol{x}_i) = \underbrace{\boldsymbol{\phi}(\boldsymbol{X})}_{D\times N}\underbrace{\boldsymbol{\alpha}}_{N\times 1}$$

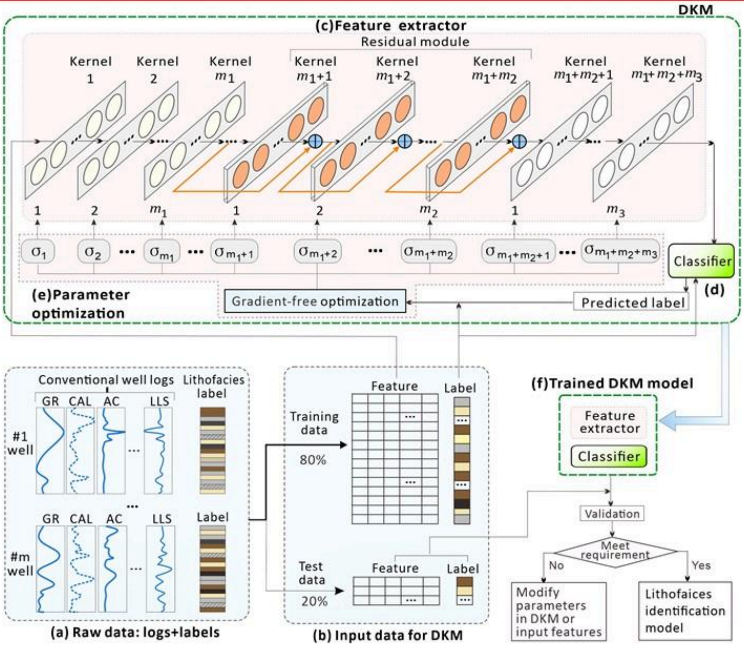
$\boldsymbol{\alpha}=[\alpha_1, \alpha_2, \ldots, \alpha_N]^T$

主成分方向$\boldsymbol{p}$可以表示为所有$\boldsymbol{\phi}(\boldsymbol{x}_i)$的线性组合

29

三、核主成分分析（KPCA）

中国石油大学 CHINA UNIVERSITY OF PETROLEUM



Shaoqun Dong, Zhaohui Zhong, Xuehui Cui, et al. A deep kernel method for lithofacies identification using conventional well logs[J]. Petroleum Science,2023

http://www.letpub.com.cn/index.php?journalid=6576&page=journalapp&view=detail

38

五、线性判别分析法（LDA）

中国石油大学 CHINA UNIVERSITY OF PETROLEUM

- 线性判别分析法（**Linear Discriminant Analysis，LDA**）1936年由英国统计学家和遗传学家Fisher提出。
- 其算法的思想是将数据投影到低维空间之后，使得同一类数据尽可能的紧凑，不同类的数据尽可能分散。
- LDA算法是一种有监督的机器学习算法。

优点：

（1）在降维过程中可以使用类别的先验知识经验。

（2）在样本分类信息依赖均值而不是方差时，比PCA之类的算法较优。

缺点：

（1）不适合对非高斯分布样本进行降维，PCA也有这个问题。

（2）降维最多降到类别数k-1的维数。

（3）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。

40

六、核Fisher判别分析（KFD）

中国石油大学 CHINA UNIVERSITY OF PETROLEUM

- □ 核Fisher判别分析（Kernel Fisher Discriminant Analysis）1999年由Fraunhofer FIRST大学的Mika提出，是基于Fisher线性判别提出的一种非线性分类方法。
- □ KFD的主要思路是通过核方法在高维空间进行Fisher判别分析，实现类间离散度最大的同时类内离散度最小，从而达到降维的目的。

优点:

（1）在降维过程中可以使用类别的先验知识经验。

（2）在样本分类信息依赖均值而不是方差时，比PCA之类的算法较优。

缺点:

（1）不适合对非高斯分布样本进行降维，PCA也有这个问题。

（2）降维最多降到类别数k-1的维数。

（3）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。

45

六、核Fisher判别分析（KFD）

中国石油大学 CHINA UNIVERSITY OF PETROLEUM

w^{φ}可以表示为： $w^{\varphi}=\sum_{i=1}^{n} a_i \varphi(x_i)$

KFD的目标可转换为：$J(a)=\frac{a^T K_b a}{a^T K_w a}$

$K_b=\sum_{i=1}^{K}\frac{N_i}{N}(u_i-u_0)(u_i-u_0)^T$

$K_w=\sum_{i=1}^{K}\sum_{j=1}^{N_i}(\xi_{xj}-u_i)(\xi_{xj}-u_i)^T$

$\xi_x=[K(x_1,x),\ldots,K(x_N,x)]^T$

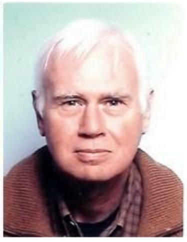
47

本节课结束！
谢谢！

七、监督局部线性嵌入（SLLE）

- 监督局部线性嵌入（Supervised Locally Linear Embedding，SLLE）是由荷兰代尔夫特理工大学的Dick（左1）和Robert（右1）于2002提出的有监督性的局部线性嵌入分析方法，是LLE的改进方法。
- SLLE算法在计算点与点之间的距离时与SLLE算法不同，其余步骤一致。

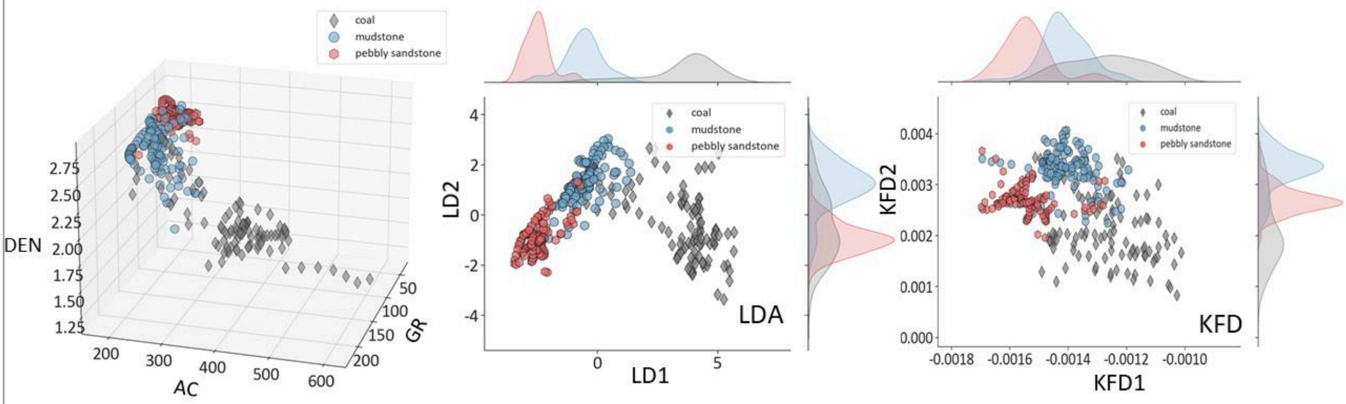

Dick


Robert

优点：
（1）SLLE考虑到了样本点的类别信息；
（2）增加了α经验参数，使得模型更可控。
缺点：
（1）对数据的流形分布特征有严格的要求；
（2）任然依赖于最近邻样本数的选择。

八、实例应用

实战—岩性数据

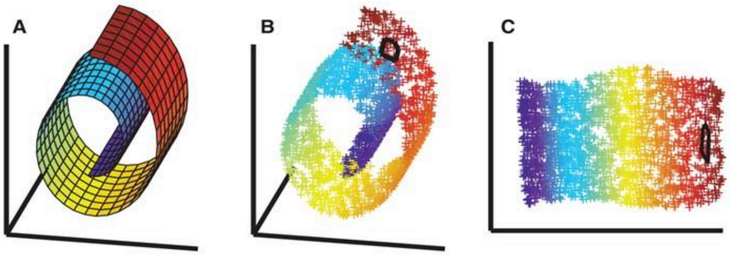


四、局部线性嵌入（LLE）

- □ 局部线性嵌入(Locally Linear Embedding，LLE) 是2000年由加拿大计算机科学家S.T.Roweis等人提出了一种针对非线性数据的无监督降维方法.
- □ 其属于流形学习的一种。和传统的PCA，LDA等关注样本方差的降维方法相比，LLE关注于降维时保持样本局部的线性特征。

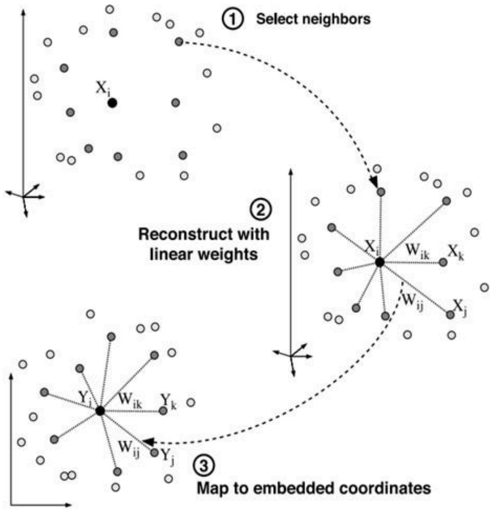
优点:
（1）该算法可以很好表达数据的内在流形结构，能够保留数据的本质特征，这样可以很好的保留原有数据特征；
（2）该算法本身参数的选择很少，故能更好的进行特征参数优化。
缺点:
（1）对数据的流形分布特征有严格的要求；
（2）算法对最近邻样本数的选择非常敏感，不同的最近邻数对最后的降维结果有很大影响。

四、局部线性嵌入（LLE）



(Sam T. Roweis et al., 2000, Science)

步骤：



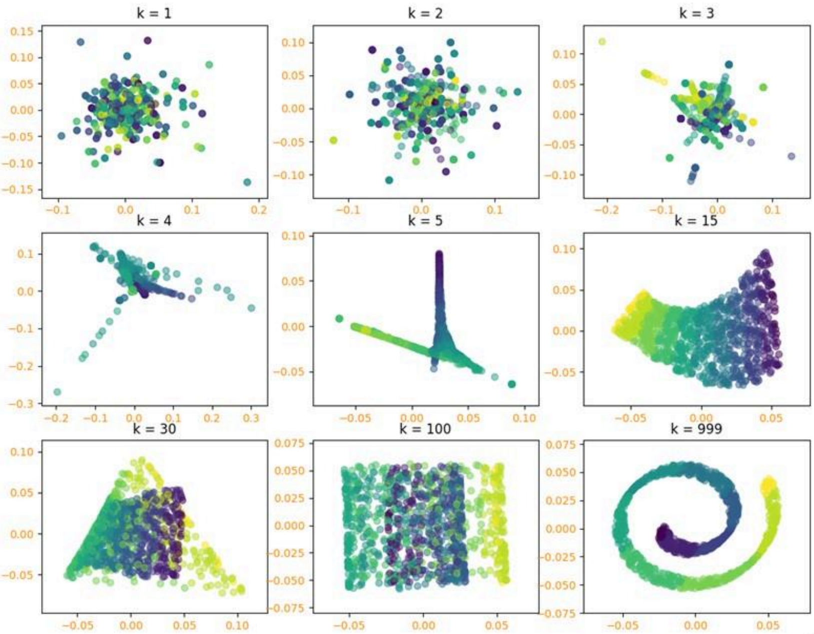
66

四、局部线性嵌入（LLE）

■ 当k取值较小（k=1、2、3、4）时，因为当近邻个数太少时，难以很好地反映数据的拓扑关系，LLE算法不能将数据很好地映射到低维空间。

■ 当k值取值适合，比如k=15，可见不同颜色的数据能很好地被分开并保持适合的间距。

■ 但若k取值太大，如k=100，999，不同颜色的数据开始相互重叠，说明选取的近邻个数太多则不能反映数据的流形信息。



71

五、线性判别分析法（LDA）

中国石油大学 CHINA UNIVERSITY OF PETROLEUM

数据集$\boldsymbol{D}=\{(\boldsymbol{x_1},\boldsymbol{y_1}),(\boldsymbol{x_2},\boldsymbol{y_2}),\ldots,(\boldsymbol{x_m},\boldsymbol{y_m})\}$，其中任意样本$\boldsymbol{x_i}$为n维向量，$\boldsymbol{y_i}\in\{\boldsymbol{0},\boldsymbol{1}\}$。$\boldsymbol{N_j}(\boldsymbol{j}=\boldsymbol{0},\boldsymbol{1})$为第$\boldsymbol{j}$类样本的个数，$\boldsymbol{X_j}(\boldsymbol{j}=\boldsymbol{0},\boldsymbol{1})$为第$\boldsymbol{j}$类样本的集合，$\boldsymbol{u_j}(\boldsymbol{j}=\boldsymbol{0},\boldsymbol{1})$为第$\boldsymbol{j}$类样本的均值向量，$\sum_j(\boldsymbol{j}=\boldsymbol{0},\boldsymbol{1})$为第样本的协方差矩阵。

最大化$\left\|\boldsymbol{w^T u_0}-\boldsymbol{w^T u_1}\right\|_2^2$，最小化$\boldsymbol{w^T}\sum_0\boldsymbol{w}+\boldsymbol{w^T}\sum_1\boldsymbol{w}$。因此优化目标为：

$$\underbrace{argmax}_{w}J(w)=\frac{\left\|\boldsymbol{w^T u_0}-\boldsymbol{w^T u_1}\right\|_2^2}{\boldsymbol{w^T}\sum_0\boldsymbol{w}+\boldsymbol{w^T}\sum_1\boldsymbol{w}}=\frac{\boldsymbol{w^T}(u_0-u_1)(u_0-u_1)^T\boldsymbol{w}}{\boldsymbol{w^T}(\sum_0+\sum_1)\boldsymbol{w}}=\frac{\boldsymbol{w^T S_b w}}{\boldsymbol{w^T S_w w}}$$

定义类内散度矩阵$\boldsymbol{S_w}$为：

$$\boldsymbol{S_w}=\sum_0+\sum_1=\sum_{x\in X_i}(\boldsymbol{u_0}-\boldsymbol{u_1})(\boldsymbol{u_0}-\boldsymbol{u_1})^T$$

同时定义类间散度矩阵$\boldsymbol{S_b}$为：$\boldsymbol{S_b}=(\boldsymbol{u_0}-\boldsymbol{u_1})(\boldsymbol{u_0}-\boldsymbol{u_1})^T$

其中 $\boldsymbol{u_j}$的表达式为：$\boldsymbol{u_j}=\frac{\boldsymbol{1}}{\boldsymbol{N_j}}\sum_{x\in X_j}\boldsymbol{x}(\boldsymbol{j}=\boldsymbol{0},\boldsymbol{1})$

$\sum_j$的表达式为：$\sum_j=\sum_{x\in X_j}(\boldsymbol{x}-\boldsymbol{u_j})\left(\boldsymbol{x}-\boldsymbol{u_j}\right)^T)(\boldsymbol{j}=\boldsymbol{0},\boldsymbol{1})$

74