主成分分析(PCA)

01

PART ONE

PCA的计算过程

主成分分析

- 主成分分析（Principal Components Analysis, PCA）是由Hotelling于1933年首先提出，亦被称为Karhunen-Loéve变换(KLT)。
- 动机：多个变量之间往往存在着一定程度的相关性，可以通过线性组合的方式，从其中提取信息。
- 思想：将n维特征映射到k维上（k<n），这k维是全新的正交特征。这k维特征称为主元，是重新构造出来的k维特征，而不是简单地从n维特征中去除其余n-k维特征。
- 主成分分析：将原始的D维数据投影到低维空间，并尽可能保留更多信息：
  - 投影后的方差最大
  - 最小化重构平方误差

去中心化

Data =

x	y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

第一步：
去中心化

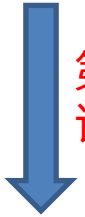
x的均值是1.81，y的均值是1.91

DataAdjust =

x	y
.69	.49
-1.31	-1.21
.39	.99
.09	.29
1.29	1.09
.49	.79
.19	-.31
-.81	-.81
-.31	-.31
-.71	-1.01

计算协方差

$$C = \begin{pmatrix} cov(x,x) & cov(x,y) & cov(x,z) \\ cov(y,x) & cov(y,y) & cov(y,z) \\ cov(z,x) & cov(z,y) & cov(z,z) \end{pmatrix}$$



第二步：
计算协方差

$$cov = \begin{pmatrix} .616555556 & .615444444 \\ .615444444 & .716555556 \end{pmatrix}$$

计算特征值和特征向量

定义1: 设 A 是 n 阶方阵，若存在数 λ 和非零向量 x，

使得 $Ax = \lambda x \ (x \neq 0)$

则称 λ 是 A 的一个特征值，

x 为 A 的对应于特征值 λ 的特征向量。

第三步：
计算特征值和特征向量

$$eigenvalues = \begin{pmatrix} .0490833989 \\ 1.28402771 \end{pmatrix}$$

$$eigenvectors = \begin{pmatrix} -.735178656 & -.677873399 \\ .677873399 & -.735178656 \end{pmatrix}$$

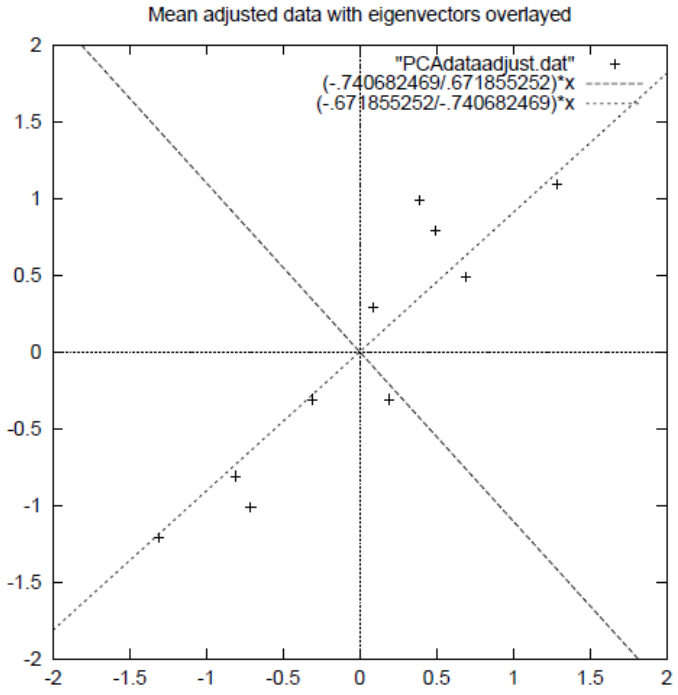
选取维度

第四步：
将特征值按照从大到小的顺序排序，选择其中最大的k个，然后将其对应的k个特征向量分别作为列向量组成特征向量矩阵

这里特征值只有两个，我们选择其中最大的那个，这里是1.28，对应的特征向量是 $[-0.74 \quad 0.68]^T$

$$A * u = \lambda * u = \begin{bmatrix} 1.28 & 0 \\ 0 & 0.05 \end{bmatrix} \begin{bmatrix} -0.68 & -0.74 \\ -0.74 & 0.68 \end{bmatrix}$$

计算降维后的数据

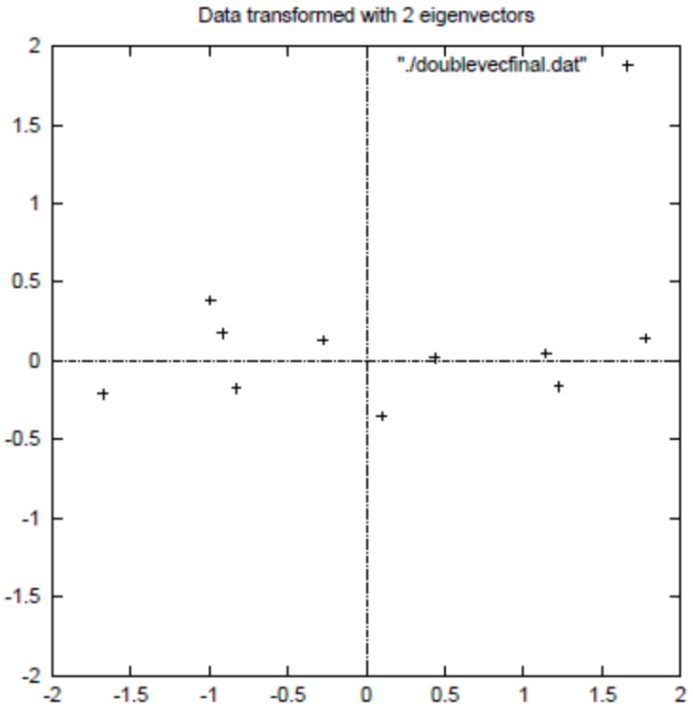


$$A * u_1^T$$

Transformed Data (Single eigenvector)

x
-.827970186
1.77758033
-.992197494
-.274210416
-1.67580142
-.912949103
.0991094375
1.14457216
.438046137
1.22382056

重构序列



Transformed Data=

x	y
-.827970186	-.175115307
1.77758033	.142857227
-.992197494	.384374989
-.274210416	.130417207
-1.67580142	-.209498461
-.912949103	.175282444
.0991094375	-.349824698
1.14457216	.0464172582
.438046137	.0177646297
1.22382056	-.162675287

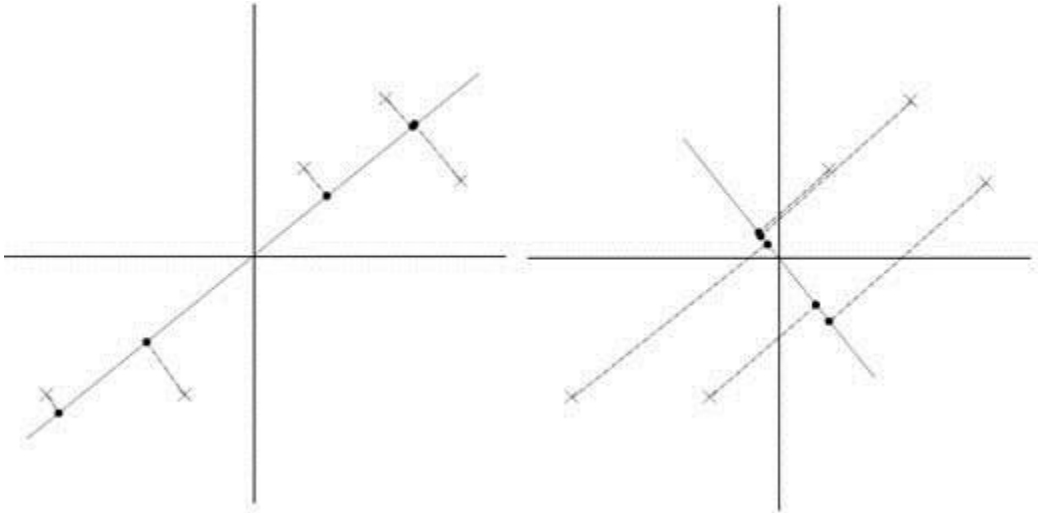
02

PART TWO

最大方差理论

最大方差理论

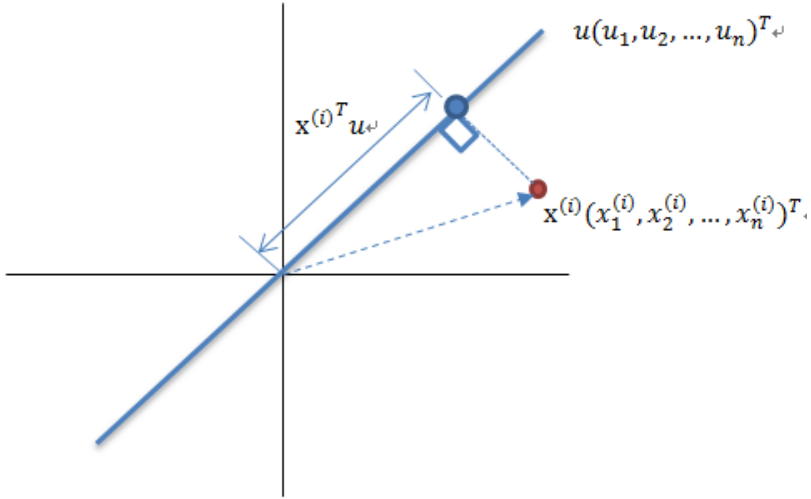
在信号处理中认为信号具有较大的方差，噪声有较小的方差，信噪比就是信号与噪声的方差比，越大越好。



假设我们选择两条不同的直线做投影，那么左右两条中哪个好呢？根据我们之前的方差最大化理论，左边的好，因为投影后的样本点之间方差最大。

投影的概念

- 红色点表示原样本点$x^{(i)}$，u是蓝色直线的斜率也是直线的方向向量，而且是单位向量，直线上的蓝色点表示原样本点$x^{(i)}$在u上的投影。



我们要求的是最佳的u，使得投影后的样本点方差最大。

- 容易知道投影点离原点的距离是$x^{(i)T}u$，由于这些原始样本点的每一维特征均值都为0，因此投影到u上的样本点的均值仍然是0。

计算

协方差矩阵

由于投影后均值为0，因此方差为：

$$\frac{1}{m}\sum_{i=1}^{m}\left(x^{(i)T}u\right)^2 = \frac{1}{m}\sum_{i=1}^{m}u^T x^{(i)}x^{(i)T}u = u^T\left(\frac{1}{m}\sum_{i=1}^{m}x^{(i)}x^{(i)T}\right)u$$

λ ∑

代替一下：

$$\lambda = u^T \sum u$$

由于u是单位向量，即$u^T u = 1$

$$u\lambda = \lambda u = uu^T \sum u = \sum u \Rightarrow \sum u = \lambda u$$

λ就是∑的特征值，u是特征向量。

计算样本

最佳的投影直线是特征值λ最大时对应的特征向量，其次是λ第二大对应的特征向量，依次类推。

因此，我们只需要对协方差矩阵进行特征值分解，得到的前k大特征值对应的特征向量就是最佳的k维新特征，而且这k维新特征是正交的。得到前k个u以后，样例$x^{(i)}$通过以下变换可以得到新的样本。

$$y^{(i)} = \begin{bmatrix} u_1^T x^{(i)} \\ u_2^T x^{(i)} \\ \vdots \\ u_k^T x^{(i)} \end{bmatrix} \in \mathbb{R}^k.$$

通过选取最大的k个u，使得方差较小的特征（如噪声）被丢弃。这是其中一种对PCA的解释

PCA算法

- 给定数据$\{x_1, x_2, \cdots, x_n\}$， 计算协方差矩阵∑

$$\sum = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})(X_i - \bar{X})^T$$

- PCA 的基向量= ∑的特征向量

- 大的特征值→ 更重要的特征向量

03

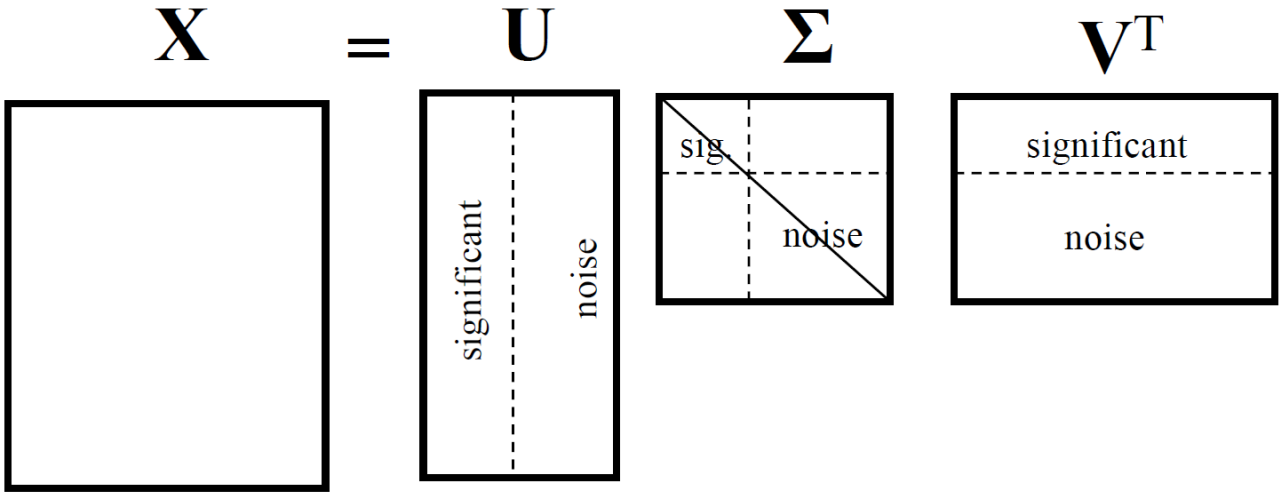
PART THREE

SVD分解

PCA:SVD分解

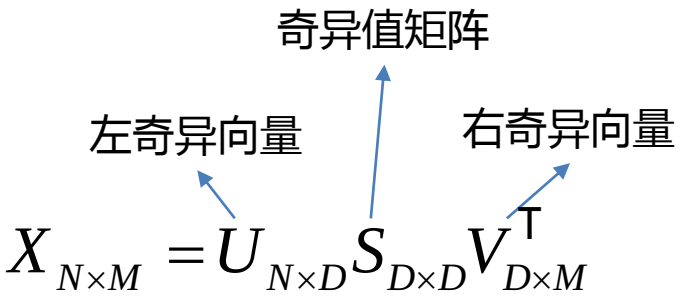
也可以对中心化后的数据矩阵X进行SVD分解实现PCA

$$X_{N\times M} = U_{N\times D} S_{D\times D} V^{\mathsf{T}}_{D\times M}$$



SVD分解

- X的SVD分解：



- 将矩阵X的转置$X^T X$，将会得到一个方阵，对这个方阵求特征值（特征值只对方阵有意义）：

$$(X^T X)V_j = \lambda_j V_j$$

这里得到的v，就是上面的右奇异向量。

- 此外，还可以得到：

$$\sigma_j = \sqrt{\lambda_j} \qquad U_j = \frac{1}{\sigma_j} X V_j$$

选择主成分的数目

- 第k个主成分对方差的**贡献率**为：$\lambda_k / \sum_{j=1}^{D} \lambda_j$
  - 前k个主成分贡献率的累计值称为累计贡献率。

- 主成分数目通常有两种方式：
  - 直接确定主成分数目
  - 根据主成分的累计贡献率确定主成分数目，如累计贡献率大于85%

04

PART FOUR

Scikit learn中的PCA实现

Scikit learn 中PCA的实现

- from sklearn.decomposition import PCA

  – PCA是一种线性降维技术，采用SVD对数据进行处理，保留最重要的前K个奇异值向量，用较低维度空间对原数据集进行映射；
  – PCA 类的实现采用scipy.linalg 来实现SVD ，只作用于密集矩阵，并且不能扩展到高维数据。对于n 维的n 个数据，时间复杂度是O(n^3)。

- PCA的构造函数

  sklearn.decomposition.PCA(n_components=None, copy=True, whiten=False, svd_solver='auto', tol=0.0, iterated_power='auto', random_state=None)

PCA参数

- n_components：int 或者string
  - 缺省: None，所有成分被保留
  - int: 要保留的主成分个数K
  - String:如n_components=’ mle’ ，将自动选取特征个数K，使得满足所要求的方差百分比
- copy：True或者False，表示在运行算法时是否将原始训练数据复制一份
  - 缺省时默认为True
  - True: 将保持原始数据不变，False 则直接在原始数据上进行计算
- whiten：白化，是否使得每个特征具有相同的方差
  - 缺省: False

PCA参数

- n_components：int 或者string
  - 缺省: None，所有成分被保留
  - int: 要保留的主成分个数K
  - String:如n_components=’ mle’ ，将自动选取特征个数K，使得满足所要求的方差百分比

05

PART FIVE

案例分析

案列分析

MNIST手写数字识别降维