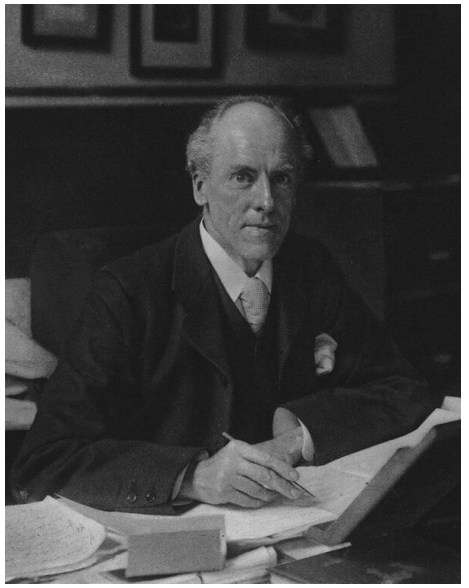
主 成 分 分 析

Principal Component Analysis

邢琛聪
51184506047

主成分分析-PCA

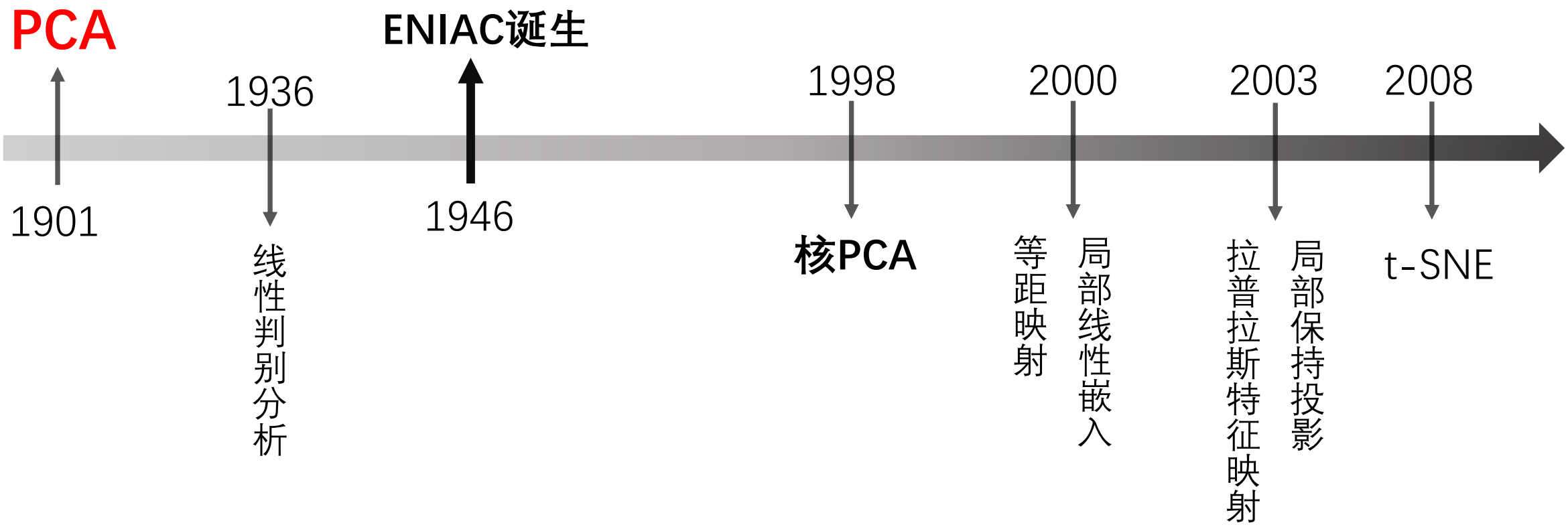
由卡尔·皮尔逊于1901年发明，是一种分析、简化数据集的技术。PCA作为最重要的降维方法之一，在数据压缩消除冗余和数据噪音消除等领域都有广泛的应用。



卡尔·皮尔逊（Karl Pearson）
1857年－1936年
英国数学家和自由思想家

PCA的数学定义是：一个正交化线性变换，把数据变换到一个新的坐标系统中，使得这一数据的任何投影的第一大方差在第一个坐标（称为第一主成分）上，第二大方差在第二个坐标（第二主成分）上，依次类推

数据降维-百年发展史

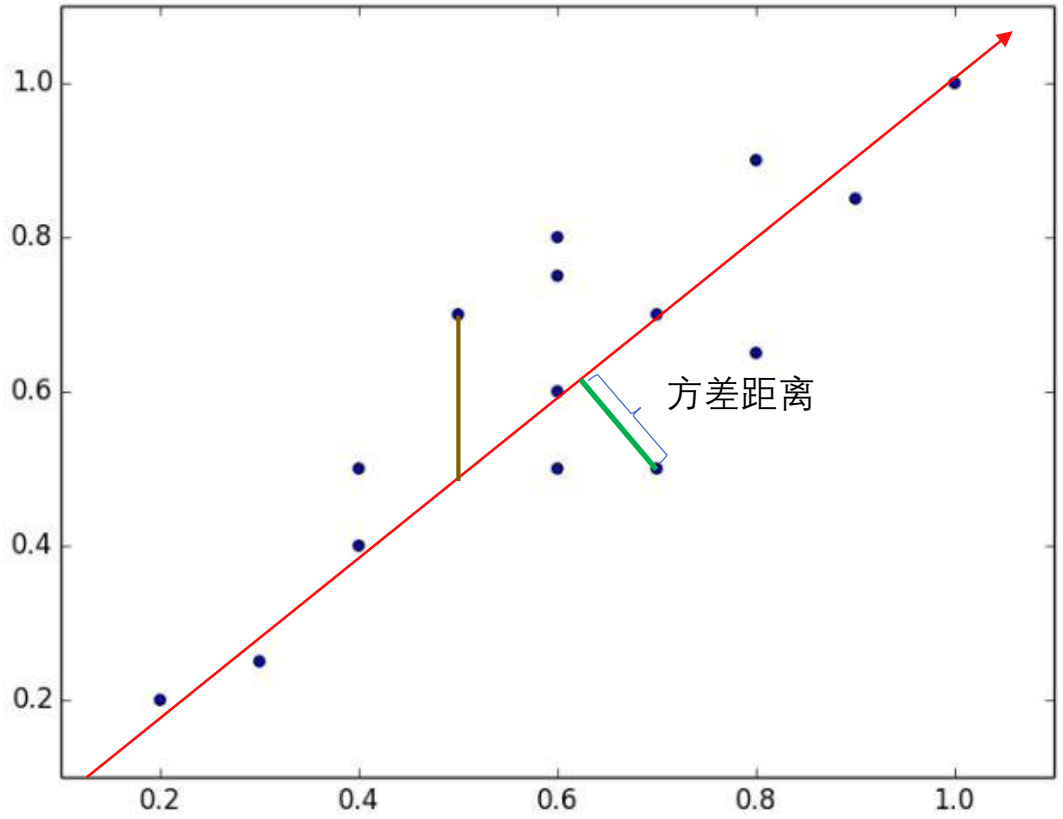

PCA
1901
1936
线性判别分析
ENIAC诞生
1946
1998
核PCA
2000
等距映射
局部线性嵌入
2003
拉普拉斯特征映射
局部保持投影
2008
t-SNE

主成分分析-理论定义

PCA有2种经常使用的定义，这两种定义会给出相同的算法。

最大方差形式：PCA被定义为数据在低维线性空间（主子空间）上的正交投影，使得投影数据的方差被最大化。(Hotelling,1933)

最小误差形式：PCA被定义为使得平均投影代价最小的线性投影。平均投影代价指数据点与其投影之间的平均平方距离。(Pearson,1901)

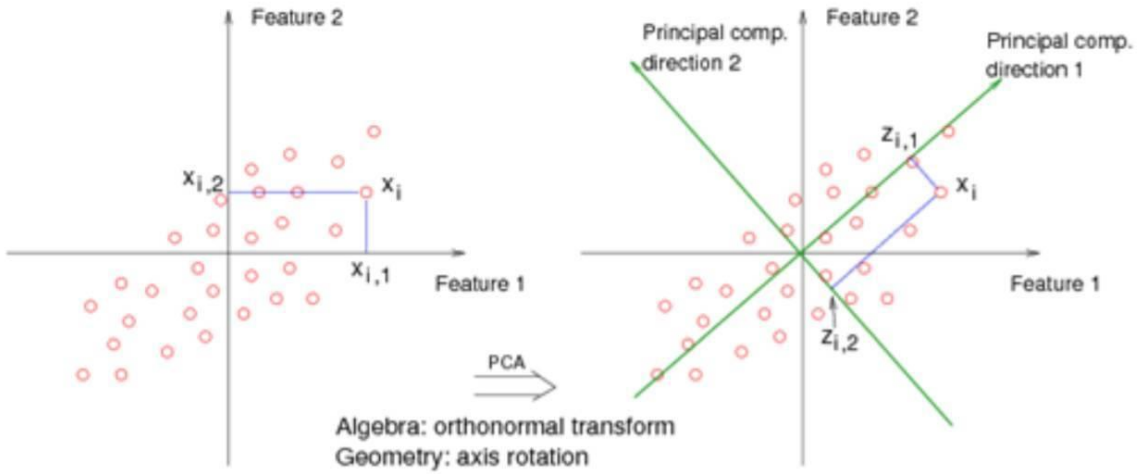


主成分分析-最大方差投影

寻找一个垂直的新的坐标系， 然后将原始数据投影过去

1.找这个坐标系的标准或者目标是什么?

2.为什么要垂直， 如果不是垂直的呢?



主成分分析-最大方差投影

投影：$\mathbf{w}^T\boldsymbol{x}$

方差：$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{w}^T\boldsymbol{x}_i)^2 = \mathbf{w}^T S\mathbf{w}$

$$S = \frac{1}{n}\sum_i \boldsymbol{x}_i\boldsymbol{x}_i^T$$

最大方差：

$$\begin{aligned} \max_{\mathbf{w}} \quad & \mathbf{w}^T S\mathbf{w} \\ s.t. \quad & \|\mathbf{w}\| = 1 \end{aligned}$$

拉格朗日乘数法：

$$L = \mathbf{w}^T S\mathbf{w} + \lambda(1 - \mathbf{w}^T\mathbf{w})$$

$$\frac{\partial L}{\partial \mathbf{w}} = 2S\mathbf{w} - 2\lambda\mathbf{w}$$

$$S\mathbf{w} = \lambda\mathbf{w}$$

方差：

$$\mathbf{w}^T S\mathbf{w} = \mathbf{w}^T\lambda\mathbf{w} = \lambda$$

理解，PCA 中，矩阵特征值和特征向量的由来！

主成分分析-最小重建误差

重建:

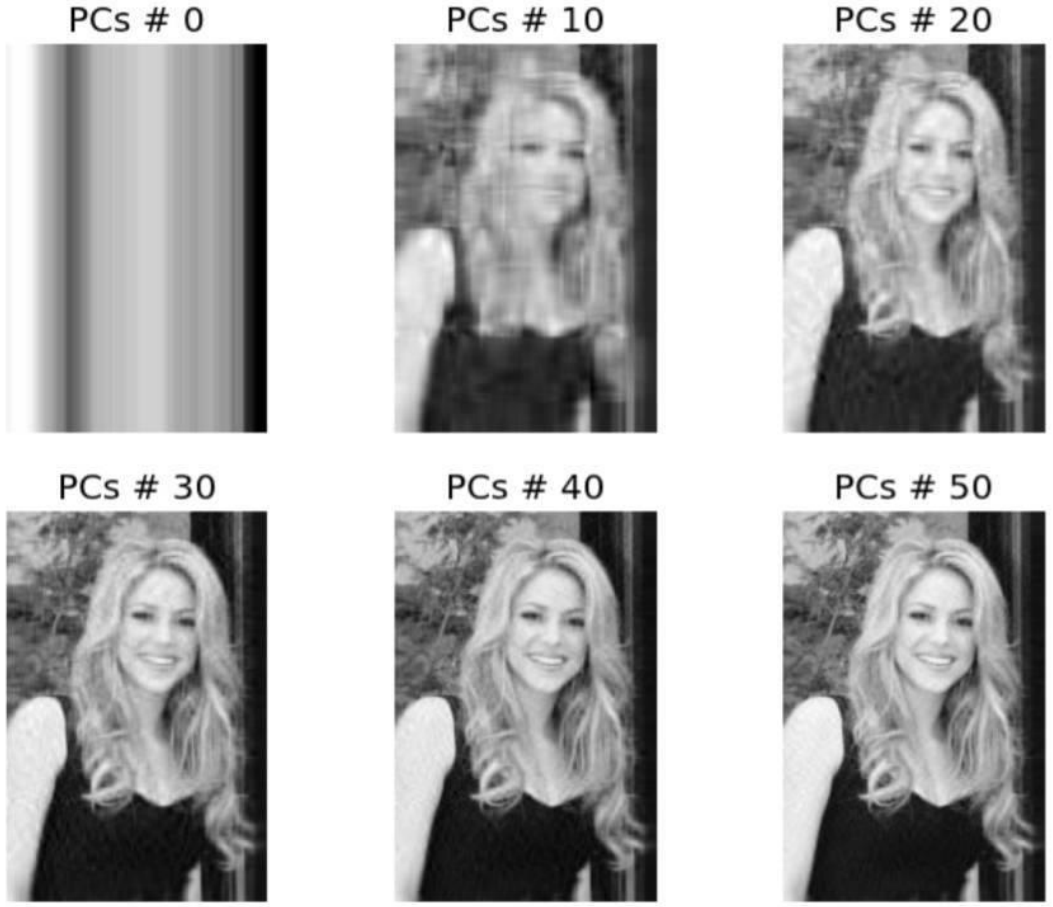
对于降维后的数据，将其恢复到原始数据空间内

正交基：$\mathbf{u}_1, \cdots, \mathbf{u}_D$

原始数据：$\boldsymbol{x}_i = \sum_{j=1}^{D} \alpha_{ij} \mathbf{u}_j$

基坐标：$\alpha_{ij} = \mathbf{u}_j^T \boldsymbol{x}_i$

降维重建：$\hat{\boldsymbol{x}}_i = \sum_{j=1}^{d} \alpha_{ij} \mathbf{u}_j$



主成分分析-最小重建误差

正交基： $\mathbf{u}_1, \cdots, \mathbf{u}_D$

原始数据： $\boldsymbol{x}_i = \sum_{j=1}^{D} \alpha_{ij} \mathbf{u}_j$

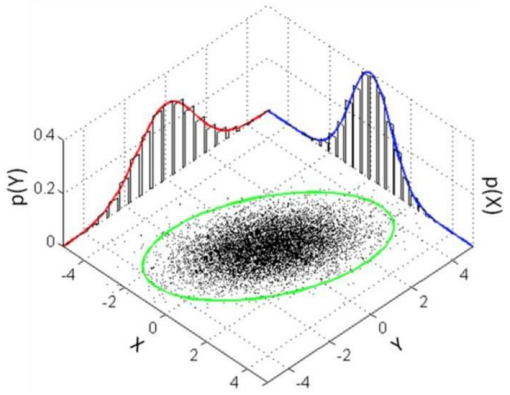
基坐标： $\alpha_{ij} = \mathbf{u}_j^T \boldsymbol{x}_i$

降维重建： $\hat{\boldsymbol{x}}_i = \sum_{j=1}^{d} \alpha_{ij} \mathbf{u}_j$

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i\|^2 &= \frac{1}{n}\sum_{i=1}^{n}\|\sum_{j=1}^{D}\alpha_{ij}\mathbf{u}_j - \sum_{j=1}^{d}\alpha_{ij}\mathbf{u}_j\|^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\|\sum_{j=d+1}^{D}\alpha_{ij}\mathbf{u}_j\|^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=d+1}^{D}\alpha_{ij}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=d+1}^{D}\mathbf{u}_j^T\boldsymbol{x}_i\boldsymbol{x}_i^T\mathbf{u}_j \\
&= \sum_{j=d+1}^{D}\mathbf{u}_j^T S\mathbf{u}_j \quad \text{等价方差最小}
\end{aligned}
$$

主成分分析-高斯先验误差

从最小重建误差， 我们可以求解最小二乘法， 从最小二乘法， 我们可以得到高斯先验误差。



假设 $X_1,...,X_n \quad Y_1,...,Y_n$ 满足如下：

$$y_i = \alpha + \beta x_i + \epsilon_i \Rightarrow \epsilon_i = y_i - (\alpha + \beta x_i)$$ 其中误差满足正态分布： $\varepsilon_i \sim N(0,\sigma^2)$

那么根据MLE， 我们得到：

$$L(\alpha,\beta,\sigma^2|\mathbf{y},\mathbf{x}) = \frac{1}{\sqrt{(2\pi\sigma^2)^n}}\exp -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-(\alpha+\beta x_i))^2$$

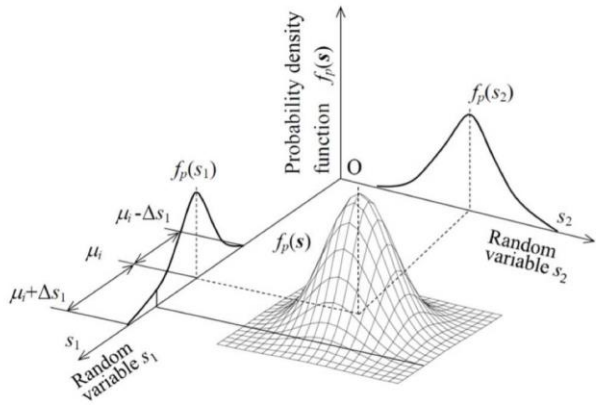
$$\frac{1}{\sqrt{2\pi\sigma^2}}\exp-\frac{\epsilon_1^2}{2\sigma^2}\cdot\frac{1}{\sqrt{2\pi\sigma^2}}\exp-\frac{\epsilon_2^2}{2\sigma^2}\cdots\frac{1}{\sqrt{2\pi\sigma^2}}\exp-\frac{\epsilon_n^2}{2\sigma^2}=\frac{1}{\sqrt{(2\pi\sigma^2)^n}}\exp-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\epsilon_i^2$$

求最大值 $\ln L(\alpha,\beta,\sigma^2|\mathbf{y},\mathbf{x}) = -\frac{n}{2}(\ln 2\pi + \ln\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-(\alpha+\beta x_i))^2$ 求最小值

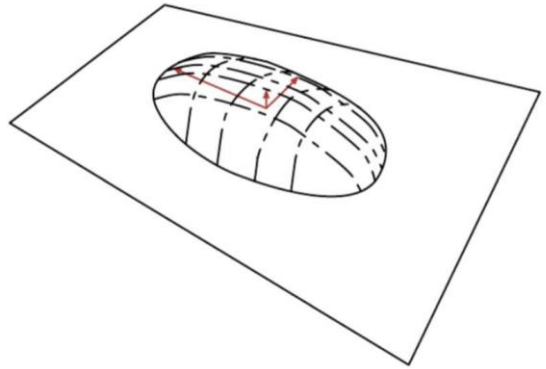
$$SS(\alpha.\beta) = \sum_{i=1}^{n}(y_i-(\alpha+\beta x_i))^2$$ ← 这个刚好是LSE的表达式

主成分分析-线性流形对齐

如果我们把高斯先验的认识， 到到数据联合分布， 但是如果把数据概率值看成是空间。那么我们可以直接到达一个新的空间认知



$$\frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{x}_i-\hat{\boldsymbol{x}}_i\|^2$$



主成分分析-步骤分解

A. 对所有样本中心化(标准化)

平衡各个特征尺度，去除量纲影响

$$X' = X - \frac{1}{m}\sum_{i=1}^{m} x_i$$

$$x_j = \frac{x_j - \frac{1}{m}\sum_{i=1}^{m} x_i}{S}$$

PCA步骤流程

对数据 $X = \{x_1, x_2, \ldots x_m\}$降维

1.对输入样本中心化

2.计算协方差矩阵

3.求解特征值与特征向量

4.按大小排列特征值，取前K个对应的特征向量组成矩阵P

5.对X进行基变换$Y = R^T X$

降维结果：Y

主成分分析-步骤分解

B. 计算协方差矩阵

由于对数据集进行了中心化处理，协方差计算无需再考虑均值

$$\Sigma = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)})(x^{(i)})^T = \frac{1}{m}XX^T$$

PCA步骤流程

对数据 $X = \{x_1, x_2, \dots x_m\}$降维

1.对输入样本中心化

2.计算协方差矩阵

3.求解特征值与特征向量

4.按大小排列特征值，取前K个对应的特征向量组成矩阵P

5.对X进行基变换$Y = R^T X$

降维结果：Y

主成分分析-步骤分解

C. 求解特征值与特征向量

常用的特征值求解方法 $\begin{cases} \varphi(M) = \det(\lambda I - M) = 0 \\ (\lambda I - M)X = \mathbf{0} \end{cases}$

Python可以使用numpy库实现：

numpy.linalg.eig

PCA步骤流程

对数据 $X = \{x_1, x_2, \dots x_m\}$降维

1.对输入样本中心化

2.计算协方差矩阵

3.求解特征值与特征向量

4.按大小排列特征值，取前K个对应的特征向量组成矩阵P

5.对X进行基变换$Y = R^T X$

降维结果：Y

主成分分析-步骤分解

D. 特征值排序，构成约简矩阵

1）对于求解出的特征值，按大小排序。

2）将排序后的前K个特征值对应的特征向量组成约简矩阵R

$$\mathrm{R} = (u^{(1)}, u^{(2)}, \ldots, u^{(k)})$$

PCA步骤流程

对数据 $X = \{x_1, x_2, \ldots x_m\}$降维

1.对输入样本中心化

2.计算协方差矩阵

3.求解特征值与特征向量

4.按大小排列特征值，取前K个对应的特征向量组成矩阵R

5.对X进行基变换$Y = R^T X$

降维结果：Y

主成分分析-步骤分解

E. 计算新的特征向量

使用约简矩阵R对去中心化的数据降维

$$Y = R^T X$$

附加. 数据还原（重建）

降维后的数据也可以还原到原始空间中:

$$X_{approx} = RY$$

PCA步骤流程

对数据 $X = \{x_1, x_2, \dots x_m\}$降维

1.对输入样本中心化

2.计算协方差矩阵

3.求解特征值与特征向量

4.按大小排列特征值，取前K个对应的特征向量组成矩阵R

5.对X进行基变换$Y = R^T X$

降维结果：Y

主成分分析-K值选取

PCA算法需要人为指定降维后的维度数k，如果k值选取太大，则性能提升不大，如果k值太小，则会丢失过多信息。可以使用以下方法评估k值选取

投影均方误差

$$\frac{1}{m}\sum_{i=1}^{m}\left\|x^{(i)}-x_{approx}^{(i)}\right\|^2$$

数据总变差

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

评估方法

$$\frac{\frac{1}{m}\sum_{i=1}^{m}\left\|x^{(i)}-x_{approx}^{(i)}\right\|^2}{\frac{1}{m}\sum_{i=1}^{m}\|x^{(i)}\|^2}\leq\varepsilon$$

← ε常取0.01，0.001，0.0001

主成分分析-综合评价

优势

PCA不要求数据呈正态分布，主成分就是按数据离散程度最大的方向对基组进行旋转

PCA通过对原始变量进行综合与简化，可以客观地确定各个指标的权重，避免主观判断的随意性

劣势

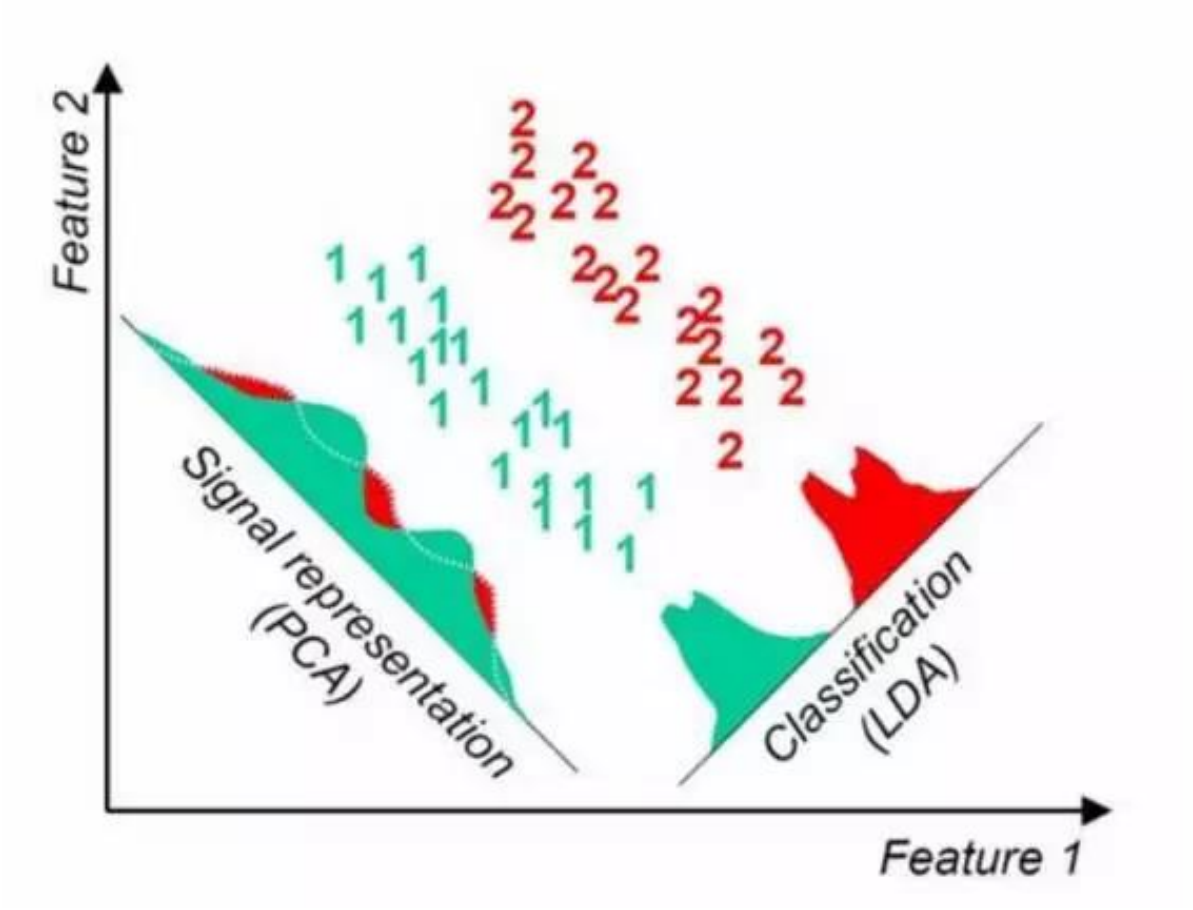
PCA类似于有损压缩，会导致数据信息丢失

PCA降低数据维度，避免维数过高导致的训练问题，但也会带来过拟合现象

PCA在计算协方差矩阵，特征值、特征向量时，比较耗时

数据降维方法比较

PCA	LDA
主成分分析	线性判别分析
线性方法	线性方法
无监督学习	有监督学习
保证样本在空间中保持原来变量信息	保证样本在空间中有最佳的可分离性
保留样本拥有最大方差	保证类别内距离越近越好，类别间距离越远越好



主成分分析-举例介绍

以鸢尾花卉数据集（Iris）为例，演示PCA算法过程：

数据集包含150条4维数据,每一条数据属于（setosa, versicolor, virginica）内的一种

这里使用PCA方法将其降至2维，并可视化降维后数据点

萼片长度	萼片宽度	花瓣长度	花瓣宽度	类别
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3	1.4	0.2	Iris-setosa
7.0	3.2	4.7	1.4	Iris-versicolor
6.4	3.2	4.5	1.5	Iris-versicolor
5.8	2.7	5.1	1.9	Iris-virginica
6.5	3.2	5.1	2.0	Iris-virginica
…	…	…	…	…

主成分分析-举例介绍

第一步：中心化操作

	萼片长度	萼片宽度	花瓣长度	花瓣宽度	类别
	5.1	3.5	1.4	0.2	Iris-setosa
	4.9	3	1.4	0.2	Iris-setosa
	7.0	3.2	4.7	1.4	Iris-versicolor
	6.4	3.2	4.5	1.5	Iris-versicolor
	5.8	2.7	5.1	1.9	Iris-virginica
	6.5	3.2	5.1	2.0	Iris-virginica
	...	...	...	...	...
平均值	**5.844**	**3.055**	**3.758**	**1.198**	

主成分分析-举例介绍

第一步：中心化操作

萼片长度	萼片宽度	花瓣长度	花瓣宽度	类别
0.742	-0.4453	2.357	0.998	Iris-setosa
0.9453	0.0547	2.357	0.998	Iris-setosa
-1.156	-0.1445	-0.9414	-0.2021	Iris-versicolor
-0.5547	-0.1445	-0.742	-0.3018	Iris-versicolor
-0.457	-0.2461	-2.242	-1.302	Iris-virginica
0.04297	0.3555	-1.344	-0.702	Iris-virginica
...	...	...	...	...

主成分分析-举例介绍

第二步：求协方差矩阵

[0.6854026845637584, -0.039272231543624164, 1.273489932885906, 0.5167785234899329]

[-0.039272231543624164, 0.1880243288590604, -0.3217281879194630 6, -0.1179739932885906 1]

[1.273489932885906, -0.3217281879194630 6, 3.1124161073825505, 1.29614093959731 54]

[0.5167785234899329, -0.1179739932885906 1, 1.29614093959731 54, 0.5826342281879194]

主成分分析-举例介绍

第三步：求解特征值、特征向量

协方差矩阵的特征值:
[4.22396988 0.24215651 0.07857844 0.02377251]

协方差矩阵的特征向量:
[
[0.36158919 -0.65615687 -0.58012383 0.31963693]
[-0.08228975 -0.730109   0.59493085 -0.32592413]
[0.85655687 0.17550995 0.07085606 -0.48008959]
[0.35887601 0.0748016   0.55180889 0.74907922]
]

主成分分析-举例介绍

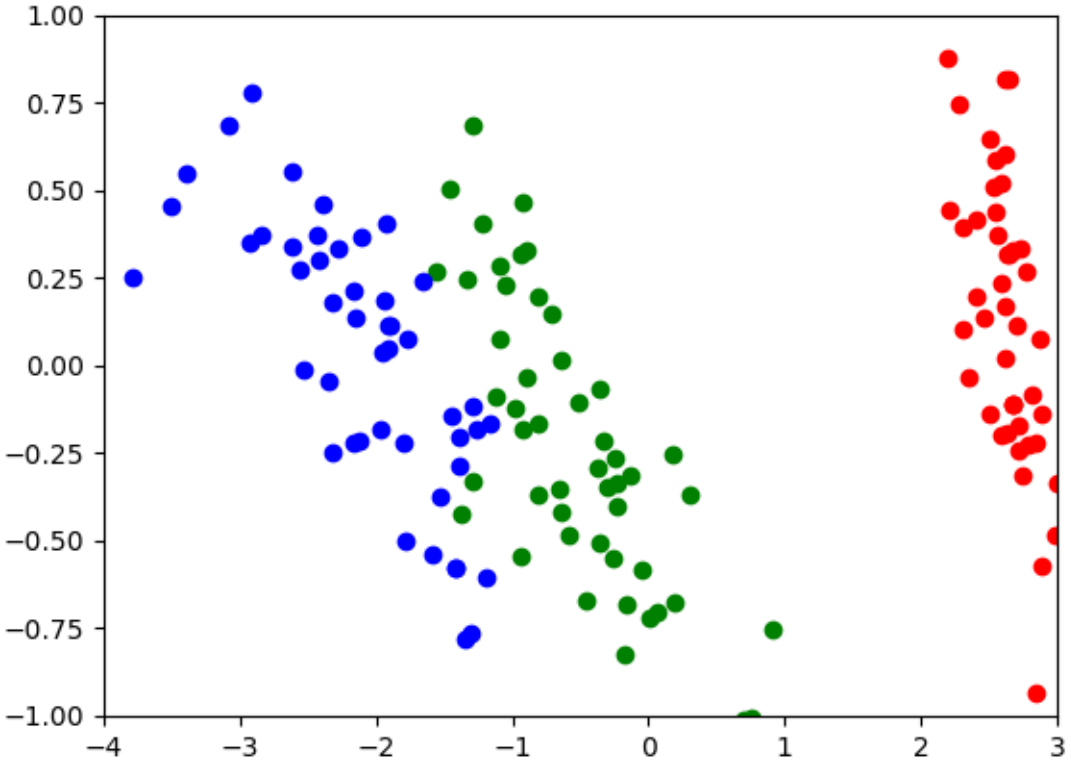
第四步：特征值排序，构成约简矩阵

协方差矩阵的特征值:
[4.22396988 0.24215651 0.07857844 0.02377251]

协方差矩阵的特征向量:
[
[0.36158919 -0.65615687 -0.58012383 0.31963693]
[-0.08228975 -0.730109   0.59493085 -0.32592413]
[0.85655687 0.17550995 0.07085606 -0.48008959]
[0.35887601 0.0748016   0.55180889 0.74907922]
]

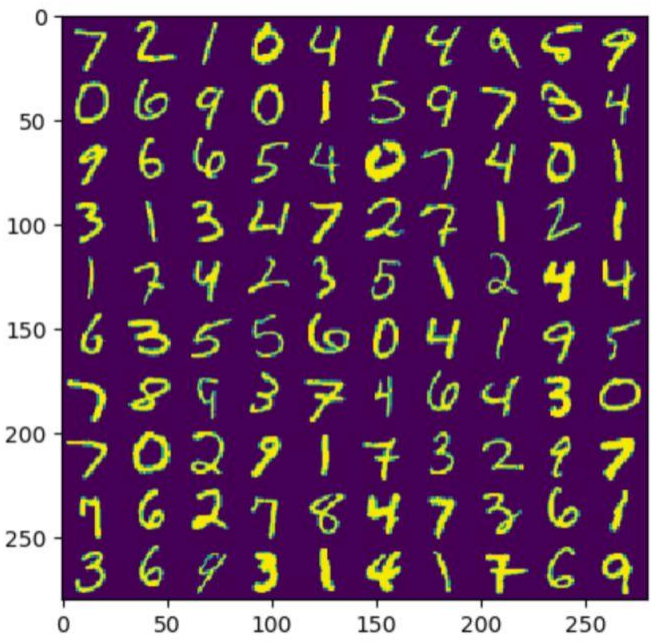
主成分分析-举例介绍

第五步：生成降维数据并可视化

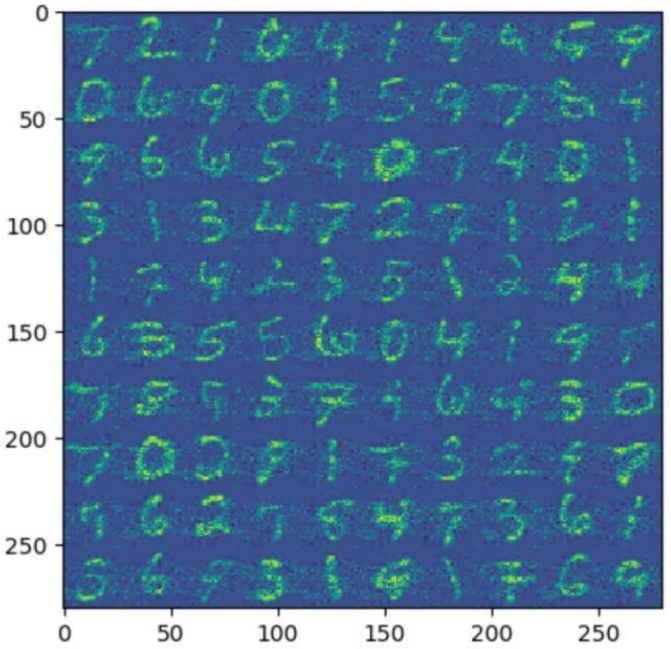


主成分分析-举例介绍

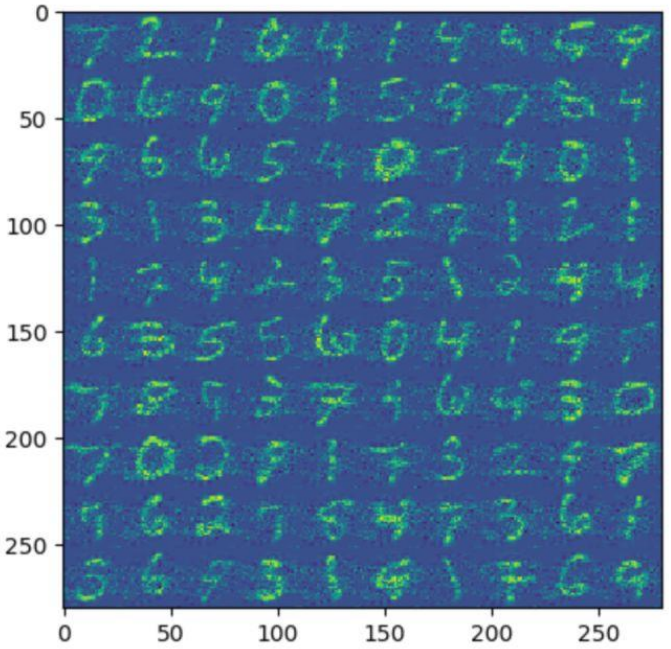
数据可视化



原始数据



前200个主成分



前300个主成分

主成分分析-作业

任务：

在MNIST数据集上使用PCA降维，并选择任意算法实现手写数字分类。将train-images-idx3-ubyte.gz中的前6000张图像和train-labels-idx1-ubyte.gz中的前6000个标签作为样本，对t10k-images-idx3-ubyte.gz中的前500张图像进行分类。

提交：

1、代码文件。对数据读入、PCA，分类器等部分标出位置。

2、说明文档。文档中必须包含2个分类正确率（使用PCA及不使用PCA）、PCA的k值选取，以及k值对应的ε。

主成分分析-作业

评分标准：

1、基础分 50分。每晚交12小时扣5分

2、代码 20分。数据读入、PCA、分类算法。

3、正确率 15分。给出2个准确率10分。高于85%，每增加1%，加1分

4、报告15分。基础分8分。给出K值选择过程并计算ε，加3分。

数学基础-协方差矩阵

协方差(Covariance)可以反映随机变量间的线性相关关系。

$$Cov(X,Y) = E\left((X-\mu_x)(Y-\mu_y)\right) = E(XY) - \mu_x\,\mu_y = \sum(x_i-\bar{x})(y_i-\bar{y})$$

对于矩阵M，协方差矩阵可以表示为：

$$M = \begin{pmatrix} x_1 & x_2 & \cdots & x_m \\ y_1 & y_2 & \cdots & y_m \end{pmatrix} \quad Cov(M) = \begin{pmatrix} \frac{1}{m}\sum_{i=1}^{m}(x_1-\mu_x)^2 & \frac{1}{m}\sum_{i=1}^{m}(x_1-\mu_x)(y_1-\mu_y) \\ \frac{1}{m}\sum_{i=1}^{m}(x_1-\mu_x)(y_1-\mu_y) & \frac{1}{m}\sum_{i=1}^{m}(y_1-\mu_y)^2 \end{pmatrix}$$

数学基础-特征值、特征向量

对于 $m \times m$ 的方阵M，特征向量方程为：

$$M\mu_i = \lambda_i \mu_i$$

λ_i：特征值

μ_i：特征向量

一般的求解方法：

$$\begin{cases} \varphi(M) = \det(\lambda I - M) = 0 \\ (\lambda I - M)X = \mathbf{0} \end{cases}$$

数学基础-特征值、特征向量

Jacobi 方法

对于实对称矩阵M，则必有正交矩阵W，使:

$$W^T M W = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \dots & 0 \\ 0 & 0 & \lambda_n \end{bmatrix} = \Lambda$$

其中Λ的对角线是M的n个特征值，W的第i列对应特征值λ_i的特征向量

定义旋转矩阵P:

$$\begin{bmatrix} 1 & & 0 & & 0 \\ & cos\varphi & & -sin\varphi & \\ \dots & 0 & & 0 & \dots \\ & sin\varphi & & cos\varphi & \\ 0 & & \dots & & 1 \end{bmatrix}$$

迭代计算 $P^T M P$,使得主对角线外的元素趋向于0

数学基础-奇异值分解

对于$m \times n$的矩阵M，其SVD分解形式为：

除了主对角线上的元素以外全为0，主对角线上的每个元素都称为奇异值σ

$$A = U\Sigma V^T$$

$m \times m$　$m \times n$　$m \times n$

U, V都是酉矩阵，即满足 $U^T U = I, V^T V = I$

$$A^T = V\Sigma U^T$$

$$A^T A = V\Sigma U^T \; U\Sigma V^T$$

$$A^T A = V\Sigma^2 V^T$$

$$\sigma^2 = \lambda$$

$$(AA^T)\mu_i = \lambda_i \mu_i$$

$m \times m$

求出的m个特征值和特征向量u，将特征向量张成 $m \times m$ 的矩阵U，U也被称为A的左奇异向量

$$(A^T A)\mu_i = \lambda_i \mu_i$$

$n \times n$

求出的n个特征值和特征向量v，将特征向量张成 $n \times n$ 的矩阵V，V也被称为A的右奇异向量

SVD性质： 可以用最大的k个的奇异值和对应的左右奇异向量来近似描述矩阵

$$A_{m\times n} = U_{m\times m}\Sigma_{m\times n}V^T_{n\times n} \approx U_{m\times k}\Sigma_{k\times k}V^T_{k\times n}$$