



温州大學
WENZHOU UNIVERSITY

机器学习-第十一章 降维

黄海广 副教授

2021年06月

本章目录

2

01 降维概述

02 SVD(奇异值分解)

03 PCA(主成分分析)

1.降维概述

3

01 降维概述

02 SVD(奇异值分解)

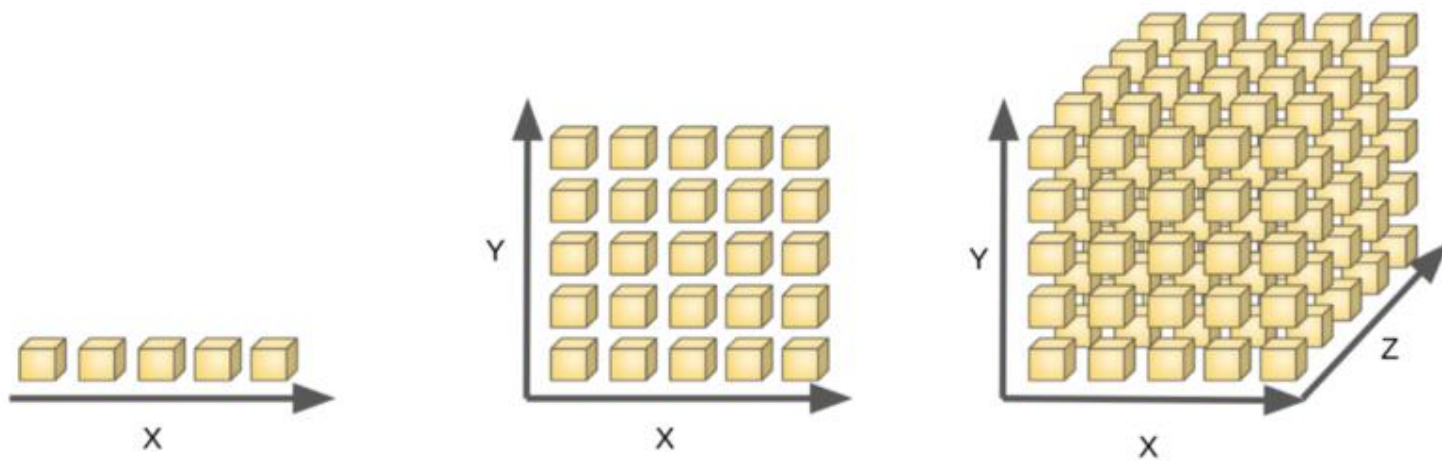
03 PCA(主成分分析)

1.降维概述

4

维数灾难(Curse of Dimensionality): 通常是指在涉及到向量的计算的问题中, 随着维数的增加, 计算量呈指数倍增长的一种现象。

在很多机器学习问题中, 训练集中的每条数据经常伴随着上千、甚至上万个特征。要处理这所有的特征的话, 不仅会让训练非常缓慢, 还会极大增加搜寻良好解决方案的困难。这个问题就是我们常说的维数灾难。



1.降维概述

5

维数灾难

维数灾难涉及数字分析、抽样、组合、机器学习、数据挖掘和数据库等诸多领域。在机器学习的建模过程中，通常指的是随着特征数量的增多，计算量会变得很大，如特征达到上亿维的话，在进行计算的时候是算不出来的。有的时候，维度太大也会导致机器学习性能的下降，并不是特征维度越大越好，模型的性能会**随着特征的增加先上升后下降**。

1.降维概述

6

什么是降维？

降维(Dimensionality Reduction)是将训练数据中的样本(实例)从高维空间转换到低维空间，该过程与信息论中有损压缩概念密切相关。同时要明白的，**不存在完全无损的降维。**

有很多种算法可以完成对原始数据的降维，在这些方法中，降维是通过对原始数据的线性变换实现的。

1.降维概述

7

为什么要降维

- 高维数据增加了运算的难度
- 高维使得学习算法的泛化能力变弱（例如，在最近邻分类器中，样本复杂度随着维度成指数增长），维度越高，算法的搜索难度和成本就越大。
- 降维能够增加数据的可读性，利于发掘数据的有意义的结构

1.降维概述

8

降维的主要作用

- 1.减少冗余特征，降低数据维度
- 2.数据可视化

1.降维概述

9

减少冗余特征

假设我们有两个特征：

x_1 :长度用厘米表示的身高； x_2 ：是用英寸表示的身高。

这两个分开的特征 x_1 和 x_2 ，实际上表示的内容相同，这样其实可以减少数据到一维，只有一个特征表示身高就够了。

很多特征具有**线性关系**，具有线性关系的特征很多都是冗余的特征，去掉冗余特征对机器学习的计算结果不会有影响。

1.降维概述

10

数据可视化

t-distributed Stochastic Neighbor Embedding(t-SNE)

t-SNE (TSNE) 将数据点之间的相似度转换为概率。原始空间中的相似度由高斯联合概率表示，嵌入空间的相似度由“学生t分布”表示。

虽然Isomap, LLE和variants等数据降维和可视化方法，更适合展开单个连续的低维的manifold。但如果要准确的可视化样本间的相似度关系，如下图所示的S曲线（不同颜色的图像表示不同类别的数据），t-SNE表现更好。因为**t-SNE主要是关注数据的局部结构。**

1.降维概述

11

降维的优缺点

降维的优点：

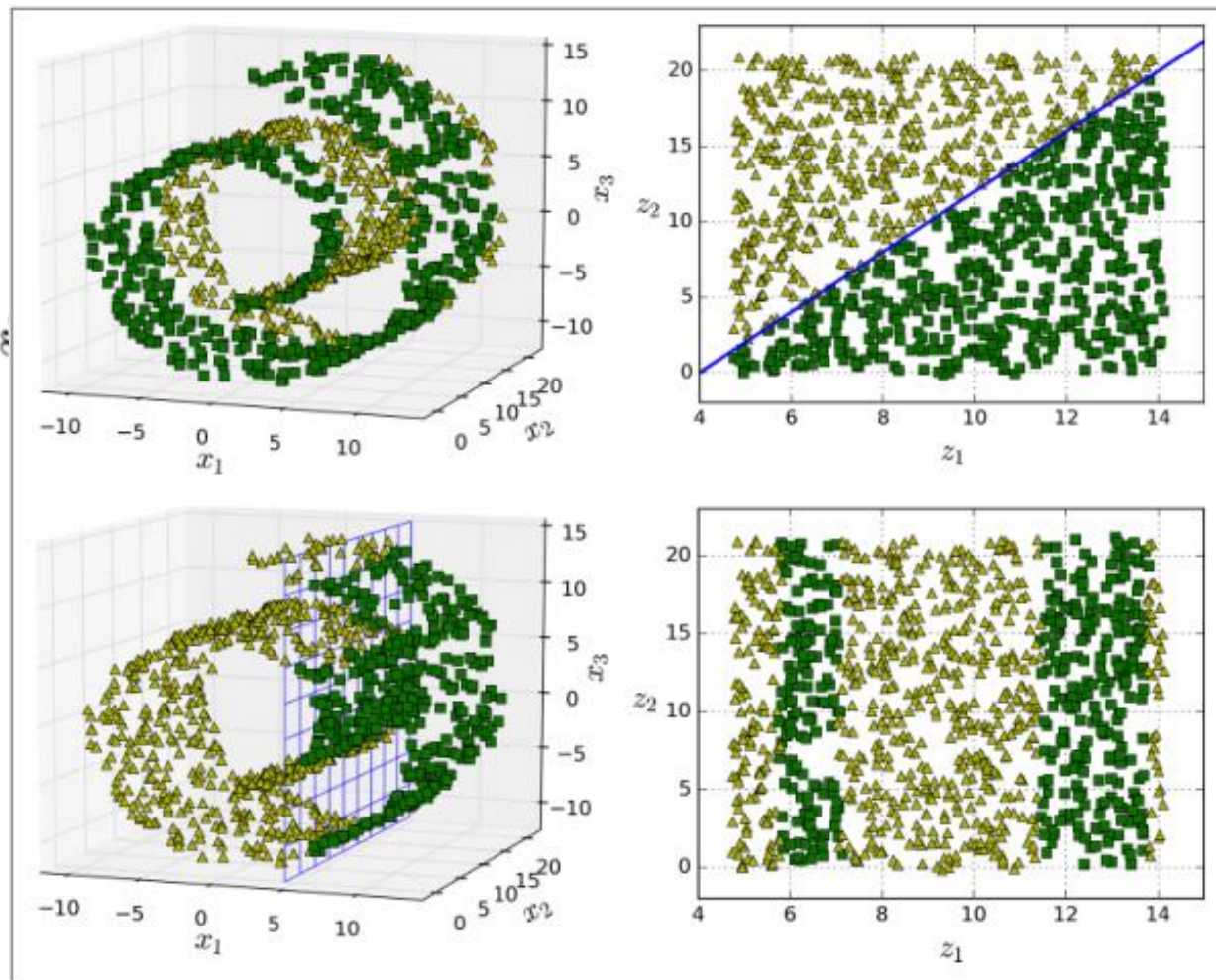
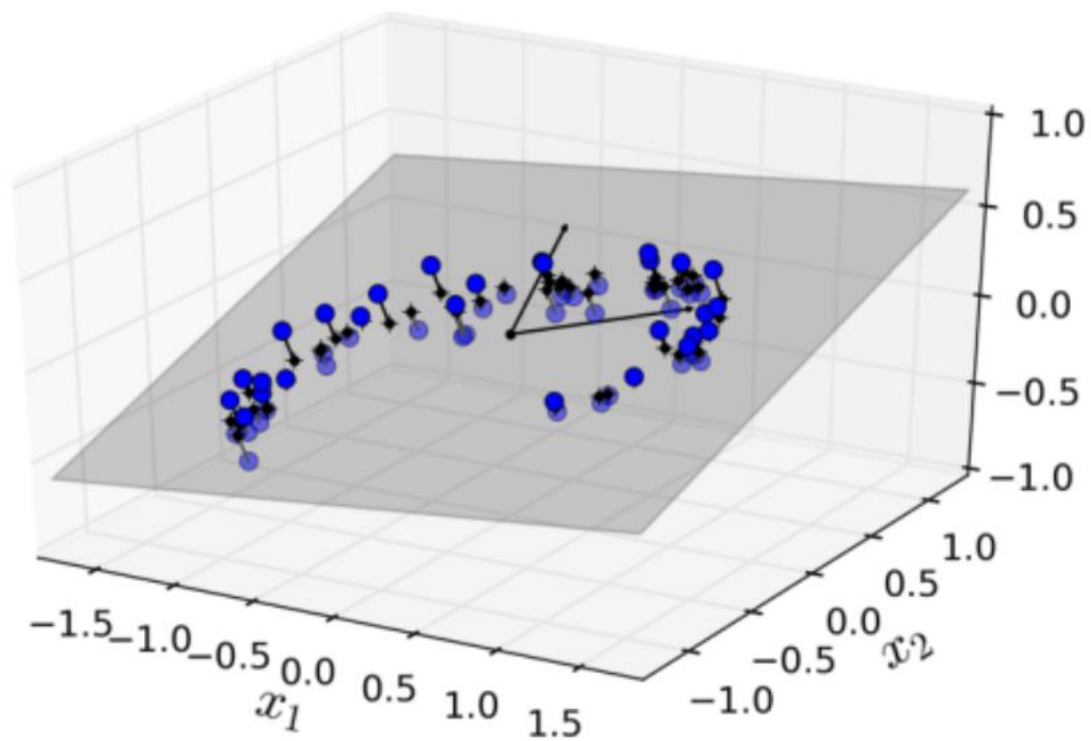
- 通过减少特征的维数，数据集存储所需的空间也相应减少，减少了特征维数所需的计算训练时间；
- 数据集特征的降维有助于快速可视化数据；
- 通过处理多重共线性消除冗余特征。

降维的缺点：

- 由于降维可能会丢失一些数据；
- 在主成分分析(PCA)降维技术中，有时需要考虑多少主成分是难以确定的，往往使用经验法则

1.降维概述

12



2.SVD(奇异值分解)

13

01 降维概述

02 SVD(奇异值分解)

03 PCA(主成分分析)

2.SVD(奇异值分解)

14

奇异值分解 (Singular Value Decomposition, 以下简称 SVD)是在机器学习领域广泛应用的算法, 它不光可以用于降维算法中的特征分解, 还可以用于推荐系统, 以及自然语言处理等领域。是很多机器学习算法的基石。

SVD可以将一个矩阵 A 分解为三个矩阵的乘积:

一个正交矩阵 U (orthogonal matrix),

一个对角矩阵 Σ (diagonal matrix),

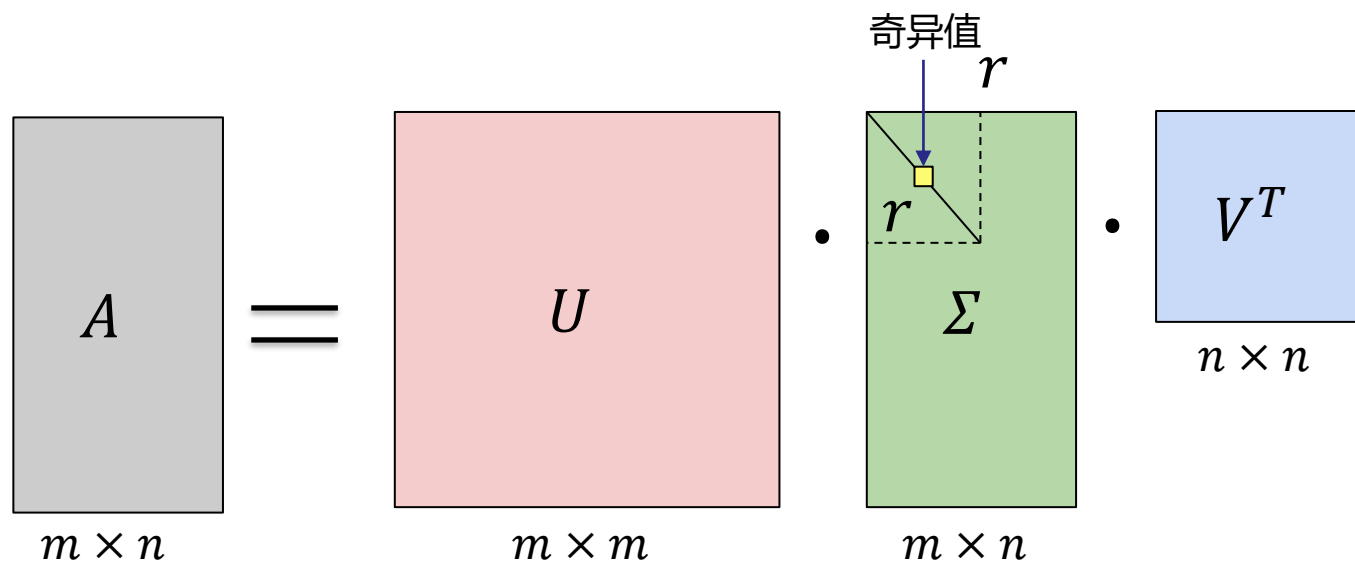
一个正交矩阵 V 的转置。

2.SVD(奇异值分解)

15

假设矩阵 A 是一个 $m \times n$ 的矩阵，通过SVD是对矩阵进行分解，
那么我们定义矩阵 A 的 SVD 为：

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



2.SVD(奇异值分解)

16

符号定义

$$A = U\Sigma V^T = u_1\sigma_1v_1^T + \cdots + u_r\sigma_rv_r^T$$

其中 U 是一个 $m \times m$ 的矩阵, 每个特征向量 u_i 叫做 A 的左奇异向量。

Σ 是一个 $m \times n$ 的矩阵, 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 。

V 是一个 $n \times n$ 的矩阵, 每个特征向量 v_i 叫做 A 的右奇异向量。

U 和 V 都是酉矩阵, 即满足: $U^T U = I, V^T V = I$ 。

r 为矩阵 A 的秩(rank)。

2.SVD(奇异值分解)

17

SVD求解 U 矩阵求解

方阵 AA^T 为 $m \times m$ 的一个方阵, 那么我们就可以进行特征分解, 得到的特征值和特征向量满足下式:

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

可以得到矩阵 AA^T 的 m 个特征值和对应的 m 个特征向量 u 了。

2.SVD(奇异值分解)

18

SVD求解 U 矩阵求解

将 AA^T 的所有特征向量组成一个 $m \times m$ 的矩阵 U ，就是我们 SVD 公式里面的 U 矩阵了。

一般我们将 U 中的每个特征向量叫做 A 的**左奇异向量**。

注意: $AA^T = (U\Sigma V^T)(U\Sigma V^T)^T = U(\Sigma\Sigma^T)U^T$

上式证明使用了 $V^TV = I, \Sigma^T = \Sigma$ 。可以看出的 AA^T 特征向量组成的矩阵就是我们 SVD 中的 U 矩阵。

2.SVD(奇异值分解)

19

V 矩阵求解

如果我们将 A 的转置和 A 做矩阵乘法, 那么会得到 $n \times n$ 的一个方阵 $A^T A$ 。既然 $A^T A$ 是方阵, 那么我们就可以进行特征分解, 得到的特征值和特征向量满足下式:

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

2.SVD(奇异值分解)

20

2.V矩阵求解

这样我们就可以得到矩阵 $A^T A$ 的 n 个特征值和对应的 n 个特征向量 v 了。
将 $A^T A$ 的所有特征向量组成一个 $n \times n$ 的矩阵 V ，就是我们 SVD 公式里面的 V 矩阵了。一般我们将 V 中的每个特征向量叫做 A 的右奇异向量。

注意：由于 $A^T A = (U \Sigma V^T)^T (U \Sigma V^T) = V (\Sigma^T \Sigma) V^T$

上式证明使用了 $U^T U = I, \Sigma^T = \Sigma$ 。可以看出 $A^T A$ 的特征向量组成的矩阵就是我们 SVD 中的 V 矩阵。

2.SVD(奇异值分解)

21

Σ 矩阵求解

进一步我们还可以看出我们的特征值矩阵等于奇异值矩阵的平方，也就

是说特征值和奇异值满足如下关系： $\sigma_i = \sqrt{\lambda_i}$

这样也就是说，我们可以不用 $\sigma_i = \frac{Av_i}{u_i}$ 来计算奇异值，也可以通过求出

$A^T A$ 的特征值取平方根来求奇异值。

2.SVD(奇异值分解)

22

3. Σ 矩阵求解

由于奇异值矩阵 Σ 除了对角线上是奇异值，而其他位置都是 0，那我们只需要求出每个奇异值 σ 就可以了。

我们注意到：

$$A = U\Sigma V^T, \text{ 则: } AV = U\Sigma V^T V$$

$$\text{由于: } V^T V = I, \text{ 则: } AV = U\Sigma$$

$$\text{得到: } Av_i = \sigma_i u_i, \sigma_i = Av_i / u_i$$

这样我们可以求出我们的每个奇异值，进而求出奇异值矩阵 Σ 。

2.SVD(奇异值分解)

23

SVD计算案例

设矩阵 A 定义为: $A = \begin{bmatrix} 3 & 0 \\ 4 & 5 \end{bmatrix}$ 则 A 的秩 $r = 2$ 。

$$A^T A = \begin{bmatrix} 3 & 4 \\ 0 & 5 \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 4 & 5 \end{bmatrix} = \begin{bmatrix} 25 & 20 \\ 20 & 25 \end{bmatrix}$$

$$A A^T = \begin{bmatrix} 3 & 0 \\ 4 & 5 \end{bmatrix} \begin{bmatrix} 3 & 4 \\ 0 & 5 \end{bmatrix} = \begin{bmatrix} 9 & 12 \\ 12 & 41 \end{bmatrix}$$

两者都有相同的迹, 都是50。

2.SVD(奇异值分解)

24

SVD计算案例

$$A = \begin{bmatrix} 3 & 0 \\ 4 & 5 \end{bmatrix}$$

进而求出 $A^T A$ 的特征值和特征向量:

$$\begin{vmatrix} 25 - \lambda & 20 \\ 20 & 25 - \lambda \end{vmatrix} = (25 - \lambda)^2 - 400 = (\lambda - 45)(\lambda - 5) = 0$$

求解得到特征值: $\lambda_1 = 45, \lambda_2 = 5$

由 $\sigma_i = \sqrt{\lambda_i}$, 可以得到奇异值为: $\sigma_1 = \sqrt{45}, \sigma_2 = \sqrt{5}$

2.SVD(奇异值分解)

25

接着求出 AA^T 的特征值和特征向量：

同理求得： $\lambda_1 = 45, \lambda_2 = 5$

$$v_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ 1 \\ \frac{1}{\sqrt{2}} \end{bmatrix}, \quad v_2 = \begin{bmatrix} -\frac{1}{\sqrt{2}} \\ 1 \\ \frac{1}{\sqrt{2}} \end{bmatrix}$$

利用 $Av_i = \sigma_i u_i, i = 1, 2$, 求奇异值：

$$Av_1 = \begin{bmatrix} \frac{3}{\sqrt{2}} \\ 9 \\ \frac{1}{\sqrt{2}} \end{bmatrix} = \sqrt{45} \begin{bmatrix} \frac{1}{\sqrt{10}} \\ 3 \\ \frac{1}{\sqrt{10}} \end{bmatrix} = \sigma_1 u_1, \quad Av_2 = \begin{bmatrix} -\frac{3}{\sqrt{2}} \\ 1 \\ \frac{1}{\sqrt{2}} \end{bmatrix} = \sqrt{5} \begin{bmatrix} -\frac{3}{\sqrt{10}} \\ 1 \\ \frac{1}{\sqrt{10}} \end{bmatrix} = \sigma_2 u_2$$

2.SVD(奇异值分解)

26

最终得到 A 的奇异值分解为：

$$U = \begin{bmatrix} \frac{1}{\sqrt{10}} & -\frac{3}{\sqrt{10}} \\ \frac{3}{\sqrt{10}} & \frac{1}{\sqrt{10}} \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \sqrt{45} & 0 \\ 0 & \sqrt{5} \end{bmatrix}, \quad V = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$$
$$A = U\Sigma V^T = \begin{bmatrix} \frac{1}{\sqrt{10}} & -\frac{3}{\sqrt{10}} \\ \frac{3}{\sqrt{10}} & \frac{1}{\sqrt{10}} \end{bmatrix} \begin{bmatrix} \sqrt{45} & 0 \\ 0 & \sqrt{5} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}^T = \begin{bmatrix} 3 & 0 \\ 4 & 5 \end{bmatrix}$$

同理：

$$A = U\Sigma V^T = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + \cdots + \sigma_r u_r v_r^T$$

2.SVD(奇异值分解)

27

SVD分解可以将一个矩阵进行分解，对角矩阵对角线上的特征值递减存放，而且奇异值的减少特别的快，在很多情况下，前 10%甚至 1%的奇异值的和就占了全部的奇异值之和的 99%以上的比例。

也就是说，对于奇异值，它跟我们特征分解中的特征值类似，我们也可以用最大的 k 个的奇异值和对应的左右奇异向量来近似描述矩阵。

也就是说：

$$A_{m \times n} = U_{m \times m} \Sigma_{m \times n} V_{n \times n}^T \approx U_{m \times k} \Sigma_{k \times k} V_{k \times n}^T$$

2.SVD(奇异值分解)

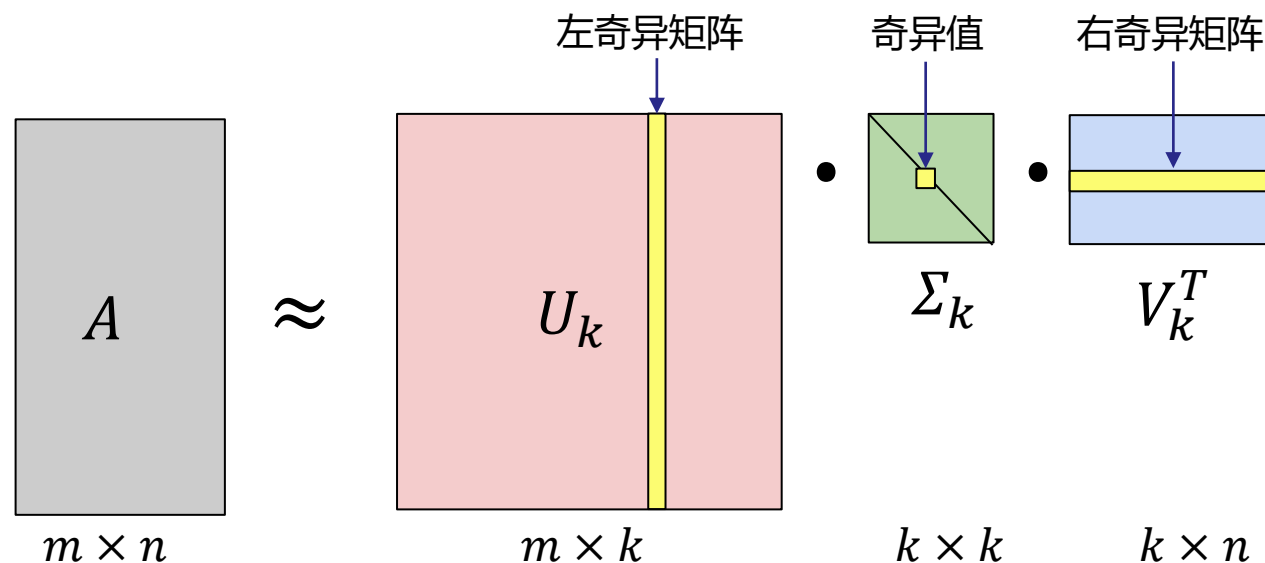
28

其中 k 要比 n 小很多, 也就是一个大的矩阵 A 可以用三个小的矩阵

$U_{m \times k}, \Sigma_{k \times k}, V_{k \times n}^T$ 来表示。

$$A_{m \times n} = U_{m \times m} \Sigma_{m \times n} V_{n \times n}^T \approx U_{m \times k} \Sigma_{k \times k} V_{k \times n}^T$$

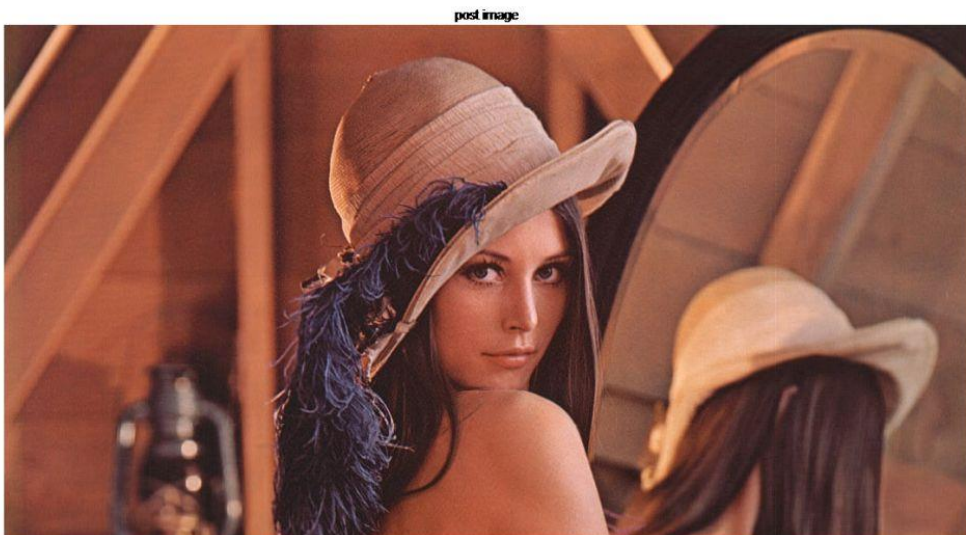
如图所示, 现在我们的矩阵 A 只需要黄色的部分的三个小矩阵就可以近似描述了。



2.SVD(奇异值分解)

29

SVD案例

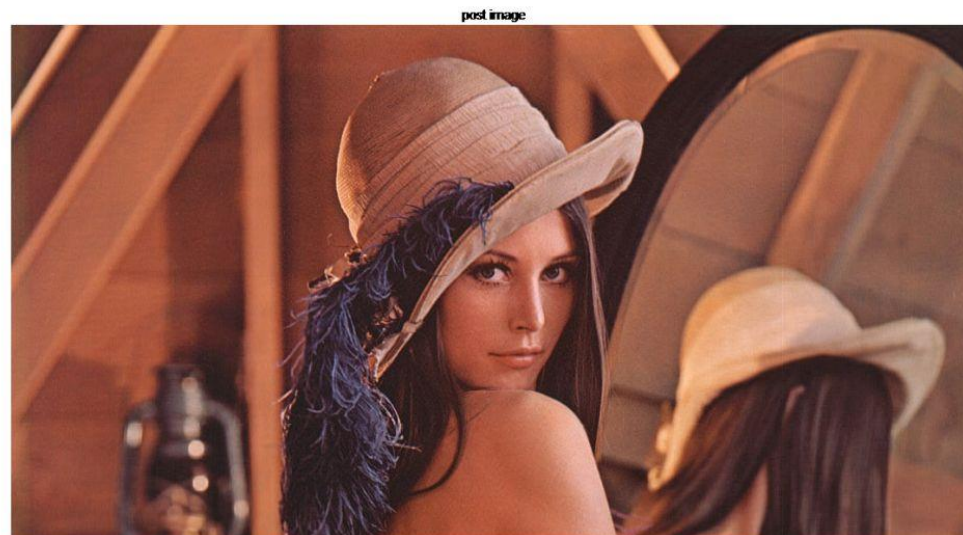


原始图像

$$m = 575, \quad n = 1081, \quad k = 150$$

$$\text{原始维度} A = 575 \times 1081 \times 3 = 1864725$$

$$\text{则原始图像经过压缩后的维度: } 3 \times (575 \times 150 + 150 \times 150 + 1081 \times 150) = 812700$$



处理后的图像

设 $k = 150$, 则经过SVD分解后的矩阵及维度:

$$U_{m \times k} = 575 \times 150, \quad \Sigma_{k \times k} = 150 \times 150, \quad V_{k \times n}^T = 1081 \times 150$$

3.PCA(主成分分析)

30

01 降维概述

02 SVD(奇异值分解)

03 PCA(主成分分析)

3.PCA(主成分分析)

31

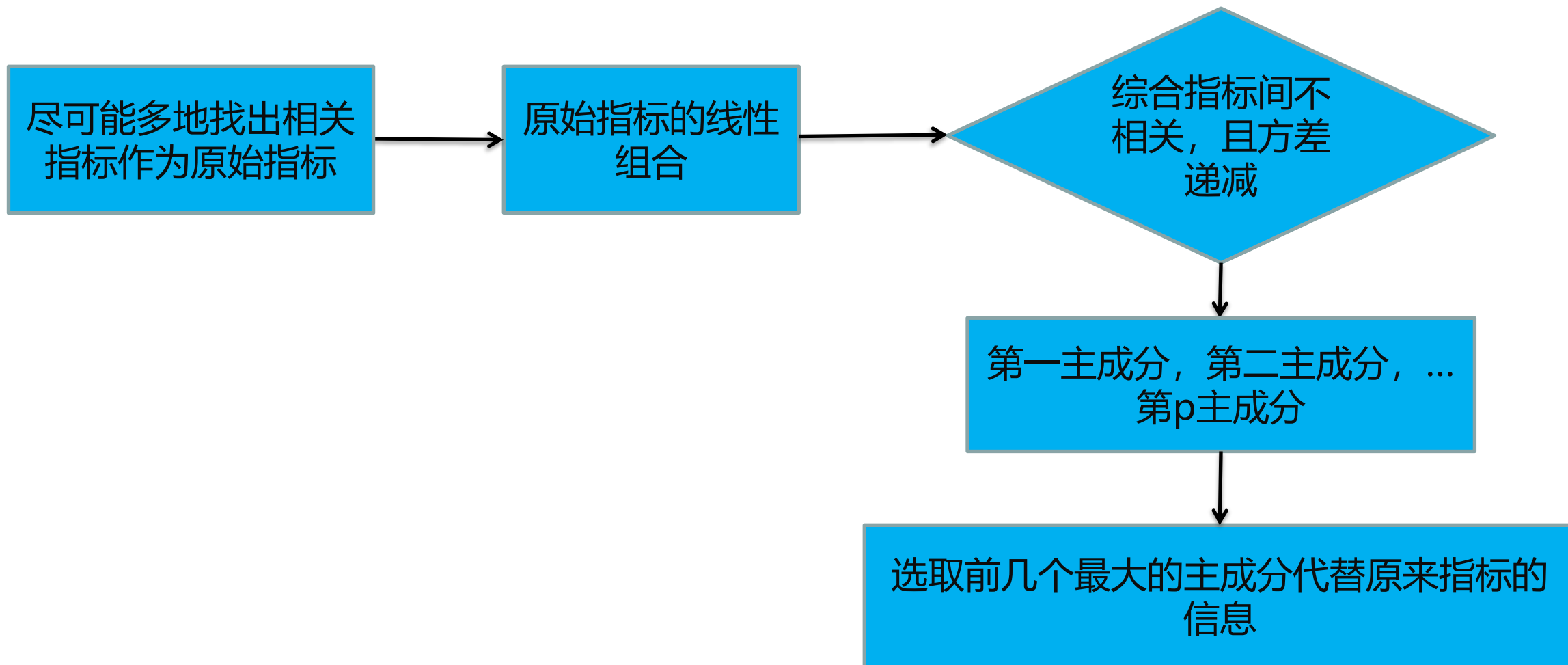
主成分分析 (Principal Component Analysis,PCA) 是一种降维方法, 通过将一个大的特征集转换成一个较小的特征集, 这个特征集**仍然包含了原始数据中的大部分信息**, 从而降低了原始数据的维数。

减少一个数据集的特征数量自然是以牺牲准确性为代价的, 但降维的诀窍是用一点准确性换取简单性。因为更小的数据集更容易探索和可视化, 并且对于机器学习算法来说, 分析数据会更快、更容易, 而不需要处理额外的特征。

3.PCA(主成分分析)

32

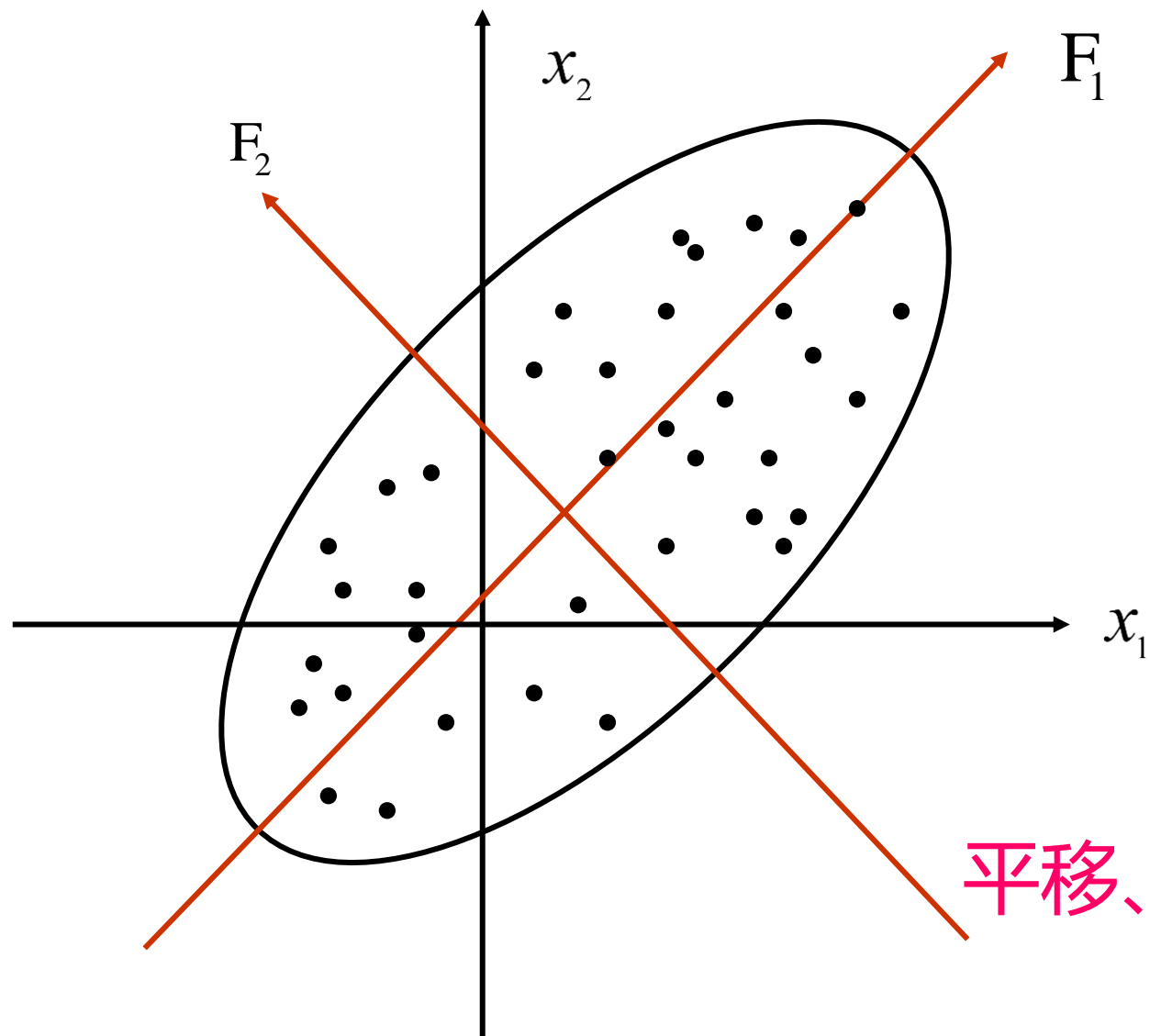
主成分分析流程图：



3.PCA(主成分分析)

33

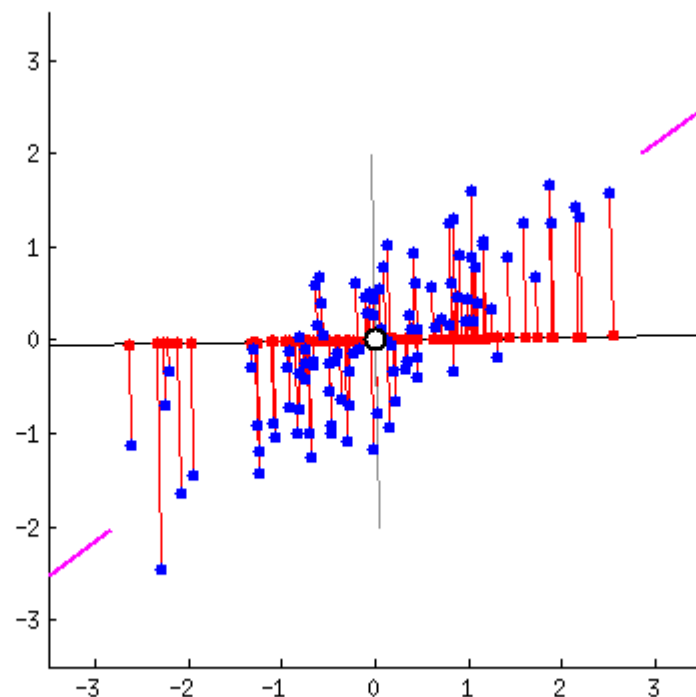
主成分分析的几何解释



平移、旋转坐标轴

3.PCA(主成分分析)

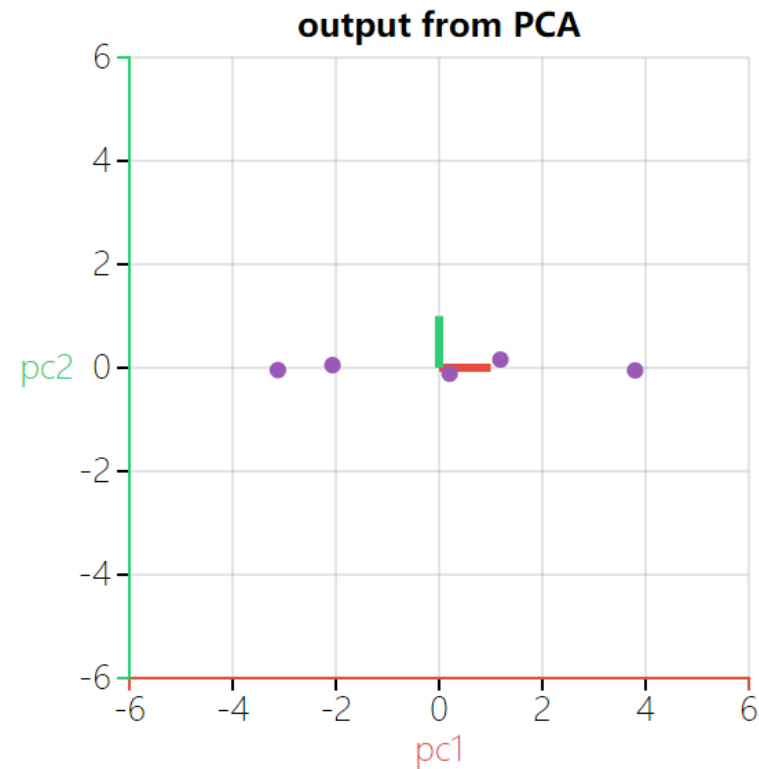
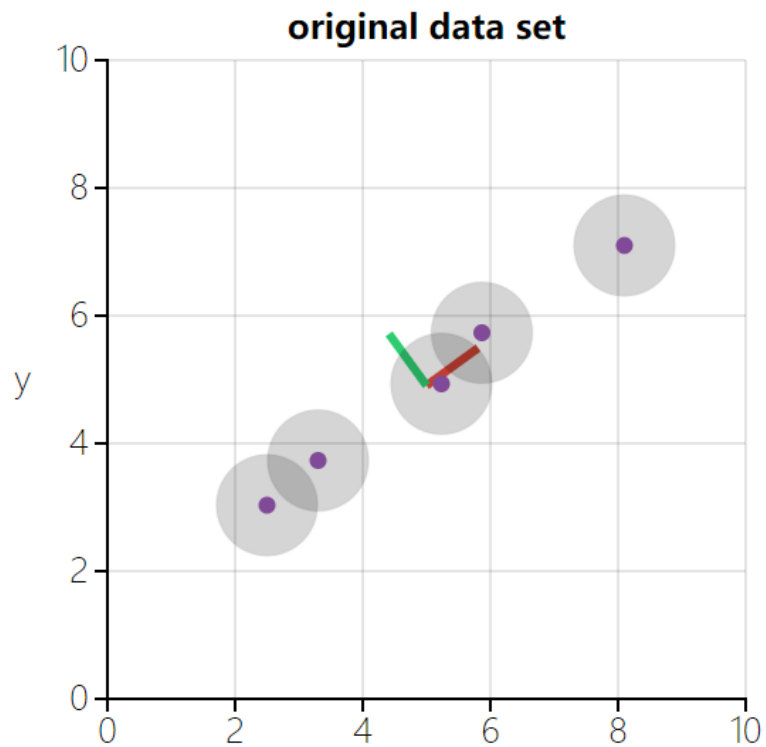
34



PCA的思想很简单——减少数据集的特征数量，同时尽可能地保留信息。

3.PCA(主成分分析)

35



通过平移、旋转坐标轴，找到主成分pc1和pc2

3.PCA(主成分分析)

36

PCA识别在训练集中占最大方差量的轴。

在图1中，它是实线。 它还找到与第一个轴正交的第二个轴，它考虑了剩余方差的最大量。在这个2D示例中，没有选择：它是虚线。如果它是一个更高维的数据集，PCA还会找到与前两个轴正交的第三个轴，以及第四个，第五个等等 - 与数据集中的维数一样多的轴。

定义第 i 轴的单位向量称为第 i 个主成分 (PC) 。

- 在图1中，第一个 **PC**为 c_1 ，第二个 **PC** 为 c_2 。
- 在图2中，前两个 **PC**由平面中的正交箭头表示，第三个 **PC**与平面正交（向上或向下）。

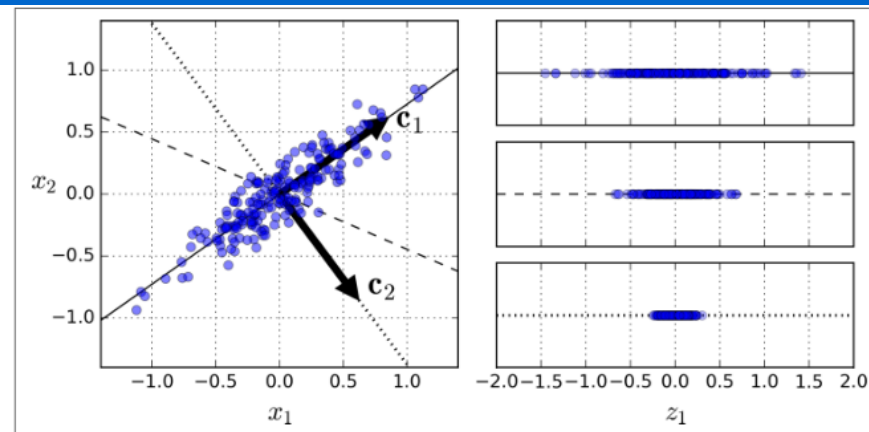


图1 选择要投影到的子空间

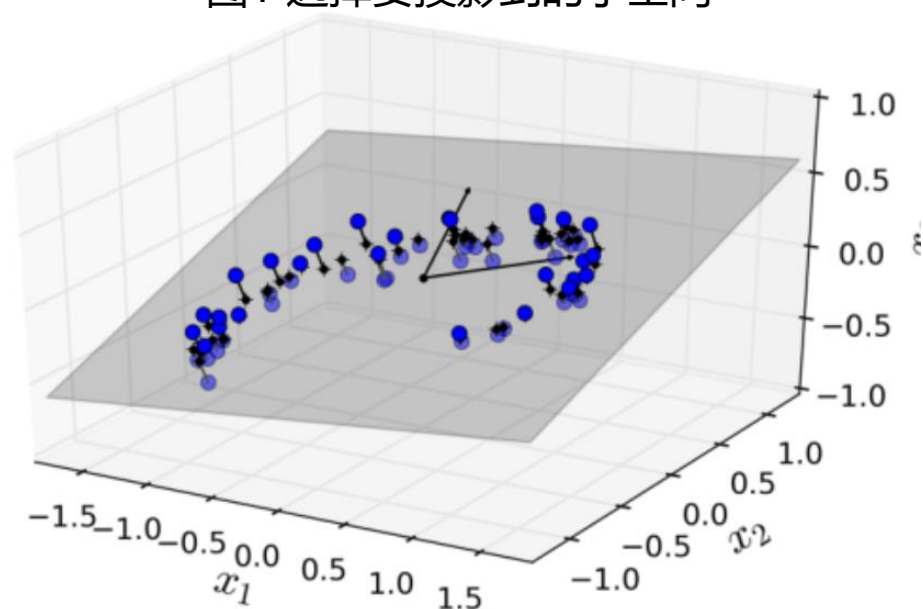


图2

3.PCA(主成分分析)

37

如何得到这些包含最大差异性的主成分方向呢？

通过计算数据矩阵的协方差矩阵

然后得到协方差矩阵的特征值特征向量

选择特征值最大(即方差最大)的k个特征所对应的特征向量组成的矩阵。

这样就可以将数据矩阵转换到新的空间当中，实现数据特征的降维。

3.PCA(主成分分析)

38

PCA的算法两种实现方法

- (1) 基于SVD分解协方差矩阵实现PCA算法
- (2) 基于特征值分解协方差矩阵实现PCA算法

3.PCA(主成分分析)

39

(1)基于SVD分解协方差矩阵实现PCA算法

PCA 减少 n 维到 k 维:

设有 m 条 n 维数据, 将原始数据按列组成 n 行 m 列矩阵 X

第一步是均值归一化。我们需要计算出所有特征的均值, 然后令 $x_j = x_j - \mu_j$ 。 (μ_j 为均值)。如果特征是在不同的数量级上, 我们还需要将其除以标准差 σ^2 。

第二步是计算**协方差矩阵** (**covariance matrix**) Σ :

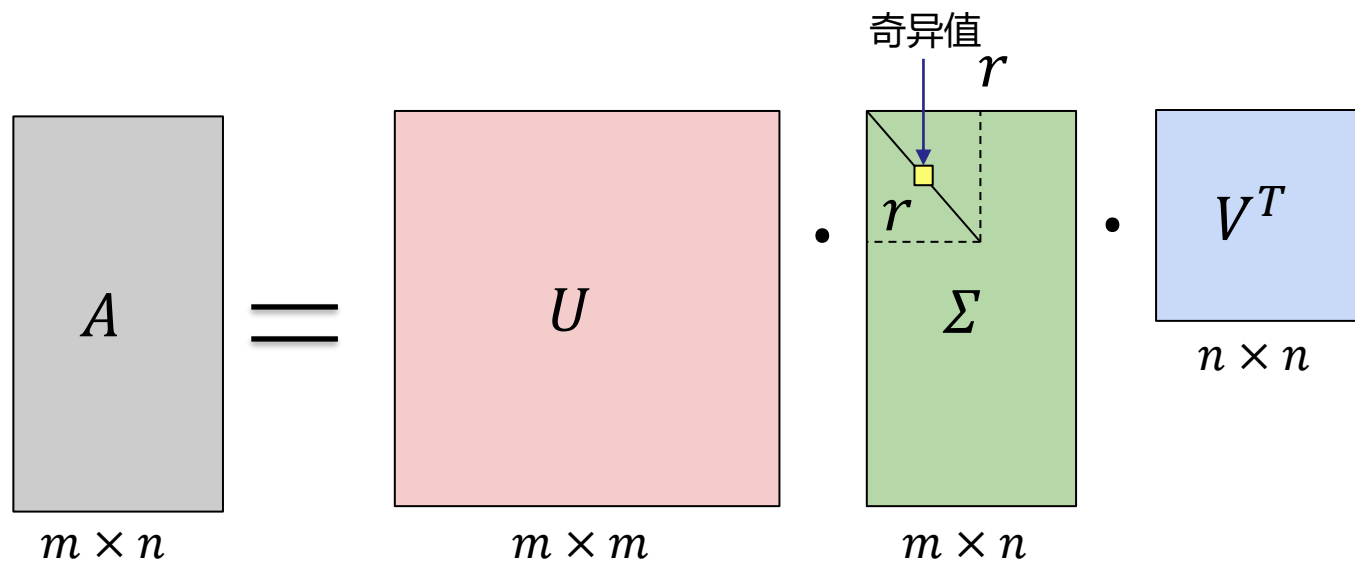
$$\Sigma = \frac{1}{m} \sum_{i=1}^n (x^{(i)}) (x^{(i)})^T$$

第三步是计算协方差矩阵 Σ 的**特征向量** (**eigenvectors**) ,可以利用奇异值分解(SVD)来求解。

3.PCA(主成分分析)

40

奇异值分解 (SVD) 的标准矩阵分解技术可以将训练集矩阵 A 分解为三个矩阵 $U \cdot \Sigma \cdot V^T$ 的点积, 其中 V^T 包含我们正在寻找的所有主成分。



3.PCA(主成分分析)

41

(2) 基于特征值分解协方差矩阵实现PCA算法

背景知识

1) 特征值与特征向量

如果一个向量 v 是矩阵 A 的特征向量，将一定可以表示成下面的形式：

$$Av = \lambda v$$

其中， λ 是特征向量 A 对应的特征值，一个矩阵的一组特征向量是一组正交向量。

3.PCA(主成分分析)

42

(2) 基于特征值分解协方差矩阵实现PCA算法

2) 特征值分解矩阵

对于矩阵 A ，有一组特征向量 v ，将这组向量进行正交化单位化，就能得到一组正交单位向量。特征值分解，就是将矩阵 A 分解为如下式：

$$A = P\Sigma P^{-1}$$

其中， P 是矩阵 A 的特征向量组成的矩阵， Σ 则是一个对角阵，对角线上的元素就是特征值。

备注：对于正交矩阵 P ，有 $P^{-1} = P^T$

3.PCA(主成分分析)

43

(2) 基于特征值分解协方差矩阵实现PCA算法

设有 m 条 n 维数据，将原始数据按列组成 n 行 m 列矩阵 X

1) 均值归一化。我们需要计算出所有特征的均值，然后令 $x_j = x_j - \mu_j$ 。 (μ_j 为均值)。

如果特征是在不同的数量级上，我们还需要将其除以标准差 σ^2 。

2) 计算协方差矩阵 Σ 。 $\Sigma = \frac{1}{m} \sum_{i=1}^n (x^{(i)})(x^{(i)})^T$

3) 用特征值分解方法计算协方差矩阵 Σ 的特征值和特征向量。

4) 对特征值从大到小排序，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行向量组成特征向量矩阵 P 。

5) 将数据转换到 k 个特征向量构建的新空间中，即 $Y = PX$ 。

3.PCA(主成分分析)

44

PCA的算方案例

$$X = \begin{pmatrix} -1 & -1 & 0 & 2 & 0 \\ -2 & 0 & 0 & 1 & 1 \end{pmatrix}$$

以这个为例，我们用PCA的方法将这组二维数据降到一维

因为这个矩阵的每行已经是零均值，所以我们可以直接求协方差矩阵：

$$\Sigma = \frac{1}{5} \begin{pmatrix} -1 & -1 & 0 & 2 & 0 \\ -2 & 0 & 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} -1 & -2 \\ -1 & 0 \\ 0 & 0 \\ 2 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} \frac{6}{5} & \frac{4}{5} \\ \frac{4}{5} & \frac{6}{5} \end{pmatrix}$$

3.PCA(主成分分析)

45

然后求 Σ 的特征值和特征向量:

$$|A - \lambda E| = \begin{vmatrix} \frac{6}{5} - \lambda & \frac{4}{5} \\ \frac{4}{5} & \frac{6}{5} - \lambda \end{vmatrix} = (\frac{6}{5} - \lambda)^2 - \frac{16}{25} = (\lambda - 2)(\lambda - 2/5) = 0$$

求解得到特征值: $\lambda_1 = 2, \lambda_2 = 2/5$

其对应的特征向量分别是: $\Sigma_1 \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \Sigma_2 \begin{pmatrix} -1 \\ 1 \end{pmatrix}$

3.PCA(主成分分析)

46

由于对应的特征向量分别是一个通解， Σ_1 和 Σ_2 可取任意实数。那么标准化后的特征向量为：

$$\begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}, \begin{pmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$$

因此我们的矩阵 P 是：

$$P = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}$$

3.PCA(主成分分析)

47

可以验证协方差矩阵 Σ 的对角化

$$P\Sigma P^T = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 6/5 & 4/5 \\ 4/5 & 6/5 \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 2/5 \end{pmatrix}$$

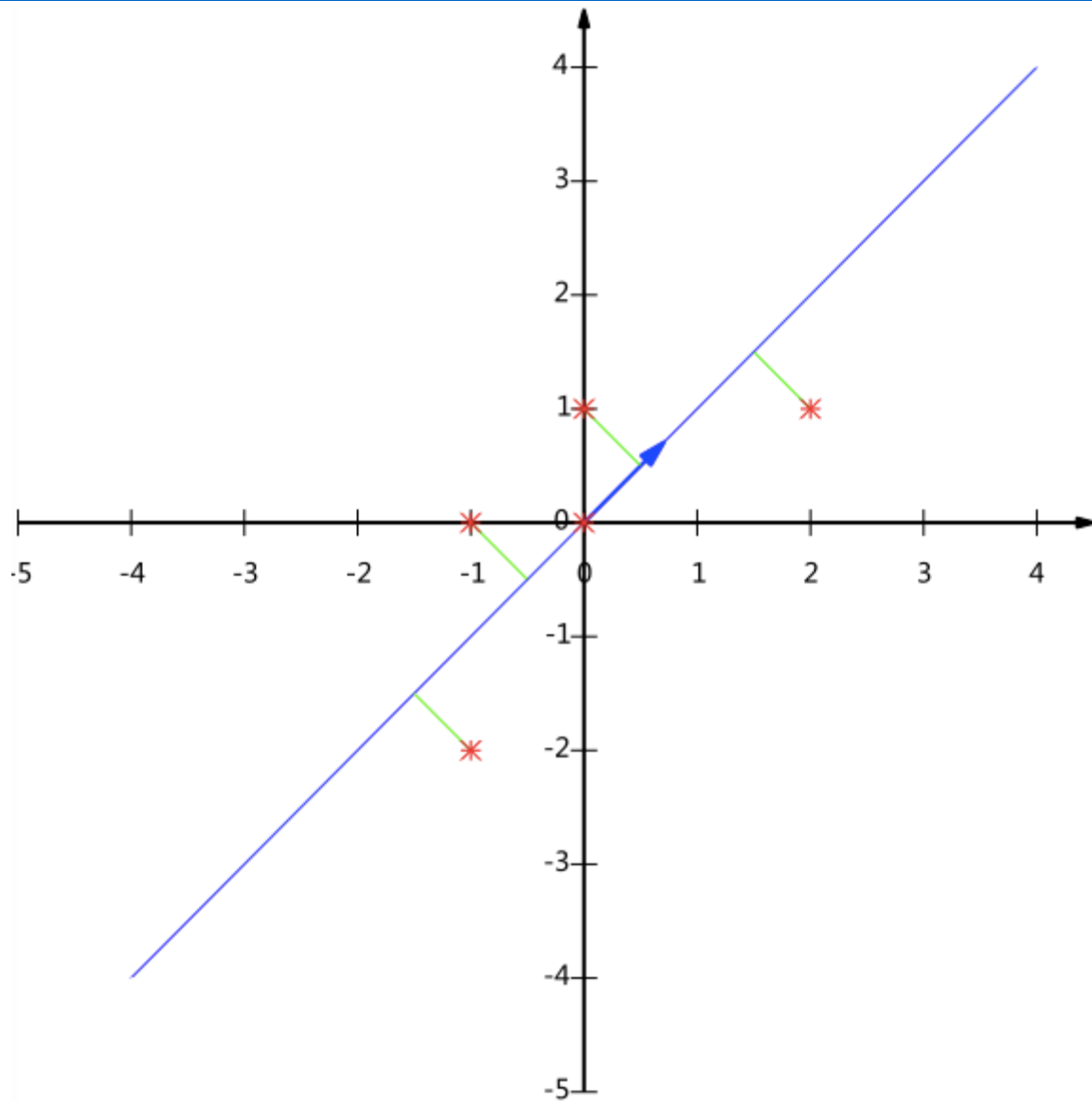
最后我们用 P 的第一行乘以数据矩阵，就得到了降维后的数据表示：

$$Y = (1/\sqrt{2} \quad 1/\sqrt{2}) \begin{pmatrix} -1 & -1 & 0 & 2 & 0 \\ -2 & 0 & 0 & 1 & 1 \end{pmatrix} = (-3/\sqrt{2} \quad -1/\sqrt{2} \quad 0 \quad 3/\sqrt{2} \quad -1/\sqrt{2})$$

3.PCA(主成分分析)

48

降维后的投影结果如下图：



3.PCA(主成分分析)

49

PCA算法优点

1. 仅仅需要以方差衡量信息量, 不受数据集以外的因素影响
2. 各主成分之间正交, 可消除原始数据成分间的相互影响的因素
3. 计算方法简单, 主要运算时特征值分解, 易于实现
4. 它是无监督学习, 完全无参数限制的

PCA算法缺点

1. 主成分各个特征维度的含义具有一定的模糊性, 不如原始样本特征的解释性强
2. 方差小的非主成分也可能含有对样本差异的重要信息, 因降维丢弃可能对后续数据处理有影响

1. 《统计学习方法》，清华大学出版社，李航著，2019年出版
2. 《机器学习》，清华大学出版社，周志华著，2016年出版
3. Christopher M. Bishop, Pattern Recognition and Machine Learning, Springer-Verlag, 2006
4. <http://blog.codinglabs.org/articles/pca-tutorial.html>

谢 谢!

