



北京交通大学
BEIJING JIAOTONG UNIVERSITY



1

机器学习

第九章 降维

鲍鹏
北京交通大学

01 **k**近邻学习

02 降维概述

03 低维嵌入

04 主成分分析

01 k近邻学习

02 降维概述

03 低维嵌入

04 主成分分析

01 k近邻学习

4

k 近邻 (k -Nearest Neighbor, k NN) 学习是一种常用的**监督学习方法**，其工作机制非常简单：

给定测试样本，基于某种**距离度量**找出训练集中与其最靠近的 k 个“邻居”的信息来进行预测。

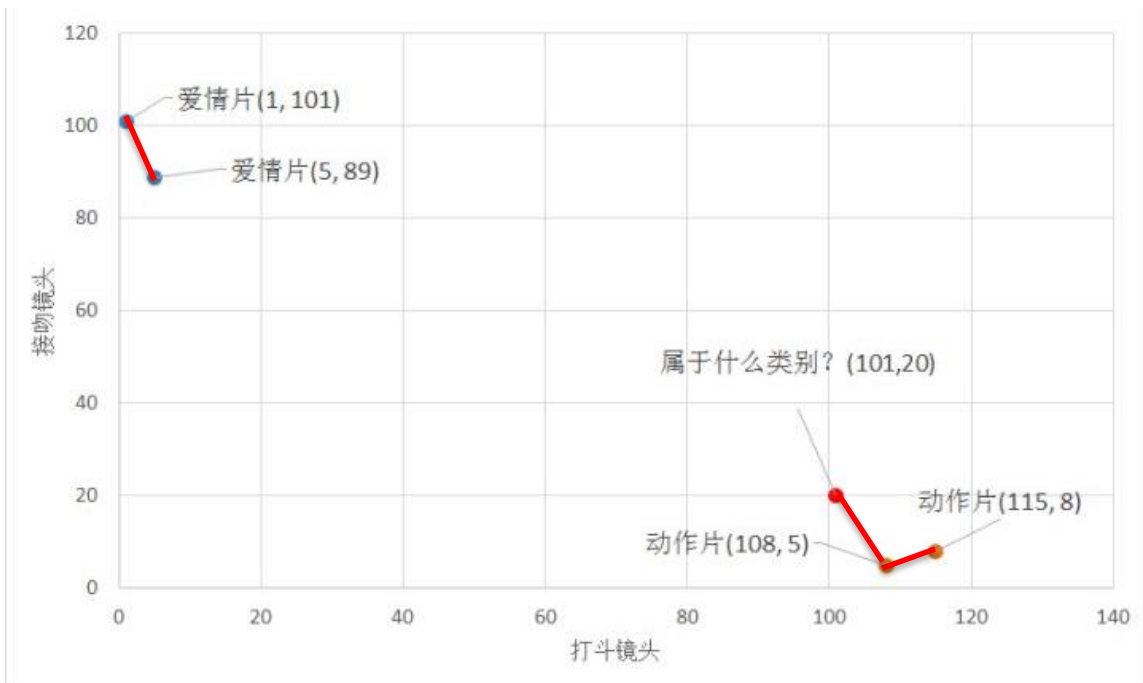
01 k近邻学习

5

距离度量：欧氏距离(Euclidean distance)

$$d(x, y) = \sqrt{\sum_i (x_i - y_i)^2}$$

电影分类

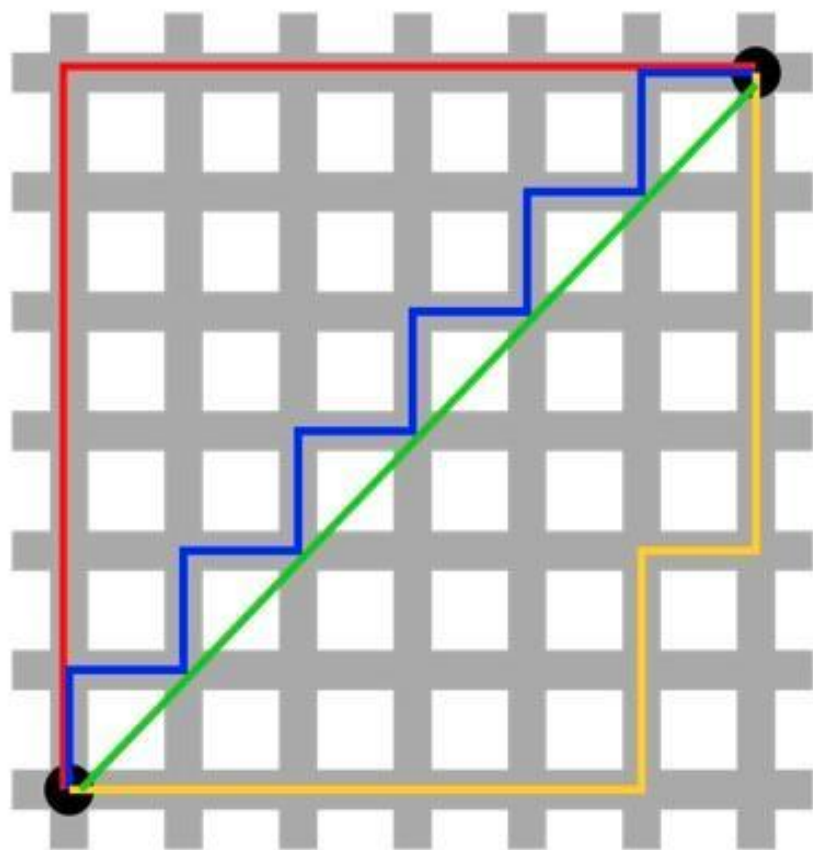


欧几里得度量 (Euclidean Metric) (也称欧氏距离) 是一个通常采用的距离定义，指在 m 维空间中两个点之间的真实距离，或者向量的自然长度 (即该点到原点的距离)。在二维和三维空间中的欧氏距离就是两点之间的实际距离。

01 k近邻学习

6

距离度量：曼哈顿距离(Manhattan distance)




$$d(x, y) = \sum_i |x_i - y_i|$$

想象你在城市道路里，要从一个十字路口开车到另外一个十字路口，驾驶距离是两点间的直线距离吗？显然不是，除非你能穿越大楼。实际驾驶距离就是这个“曼哈顿距离”。而这也是曼哈顿距离名称的来源，曼哈顿距离也称为城市街区距离(City Block distance)。

01 k近邻学习

7

距离度量：切比雪夫距离(Chebyshev distance)

	a	b	c	d	e	f	g	h	
8	5	4	3	2	2	2	2	2	8
7	5	4	3	2	1	1	1	2	7
6	5	4	3	2	1		1	2	6
5	5	4	3	2	1	1	1	2	5
4	5	4	3	2	2	2	2	2	4
3	5	4	3	3	3	3	3	3	3
2	5	4	4	4	4	4	4	4	2
1	5	5	5	5	5	5	5	5	1
	a	b	c	d	e	f	g	h	

二个点之间的距离定义是其各坐标数值差绝对值的最大值。

国际象棋棋盘上二个位置间的切比雪夫距离是指王要从一个位子移至另一个位子需要走的步数。由于王可以往斜前或斜后方向移动一格，因此可以较有效率的到达目的的格子。上图是棋盘上所有位置距f6位置的切比雪夫距离。

01 k近邻学习

8

距离度量：闵可夫斯基距离(Minkowski distance)

p 取1或2时的闵氏距离是最为常用的

$p = 2$ 即为欧氏距离，

$p = 1$ 时则为曼哈顿距离。

当 p 取无穷时的极限情况下，可以得到切比雪夫距离

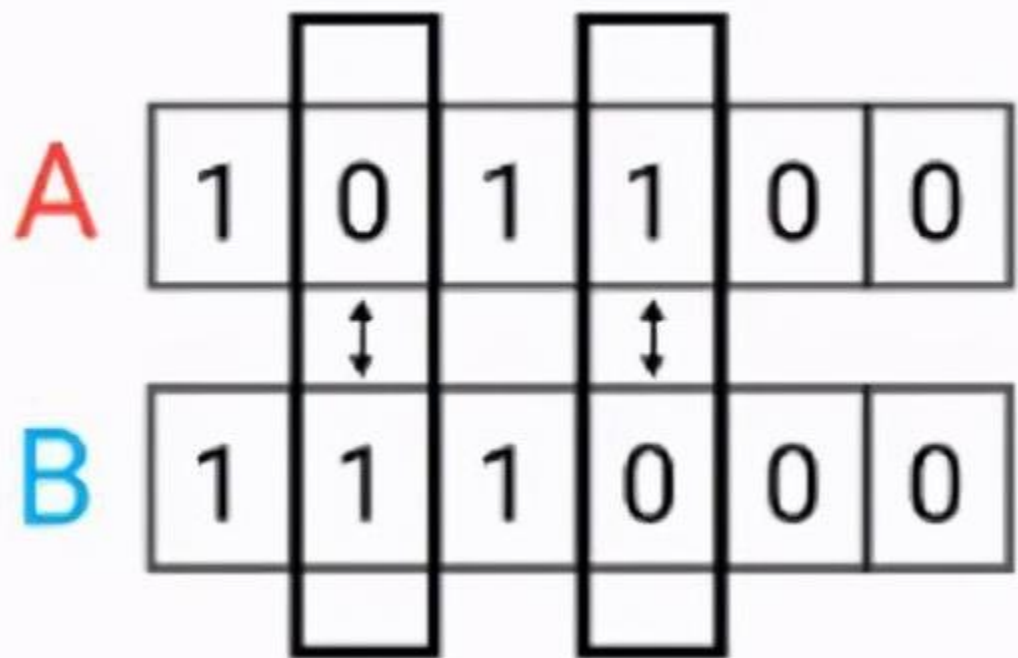
$$d(x, y) = \left(\sum_i |x_i - y_i|^p \right)^{\frac{1}{p}}$$

01 k近邻学习

9

距离度量：汉明距离(Hamming distance)

$$d(x, y) = \frac{1}{N} \sum_i 1_{x_i \neq y_i}$$



汉明距离是使用在数据传输差错控制编码里面的，汉明距离是一个概念，它表示两个（相同长度）字对应位不同的数量，我们以表示两个字之间的汉明距离。对两个字符串进行异或运算，并统计结果为1的个数，那么这个数就是汉明距离。

01 k近邻学习

10

k 近邻法是一种比较成熟也是最简单的机器学习算法，可以用于基本的分类与回归方法。

- 投票法：选择这 k 个样本中出现最多的类别标记作为预测结果。（**分类任务**）
- 平均法：将这 k 个样本的实值输出标记的平均值作为预测结果。（**回归任务**）

除此之外，还可基于距离远近进行加权平均或加权投票，距离越近的样本权重越大。

01 k近邻学习

11

算法流程如下：

- 1.计算测试对象到训练集中每个对象的距离；
- 2.按照距离的远近排序；
- 3.选取与当前测试对象最近的 k 的训练对象，作为该测试对象的邻居；
- 4.统计这 k 个邻居的类别频次；
5. k 个邻居里频次最高的类别，即为测试对象的类别；

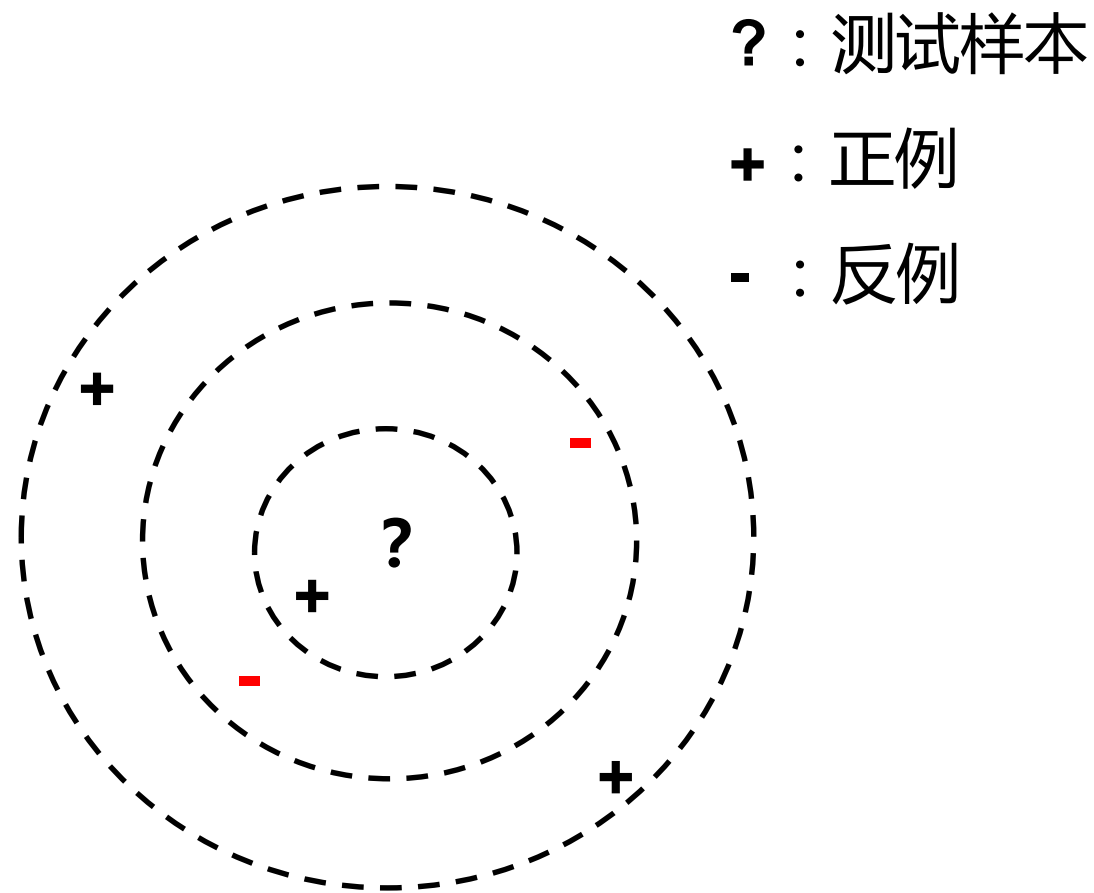
k 近邻法的三要素

- k 值选择。
- 距离度量。
- 决策规则。

01 k近邻学习

12

k 近邻分类器中的 k 是一个重要参数，当 k 取不同值时，分类结果会有显著不同。另一方面，若采用不同的距离计算方式，则找出的“近邻”可能有显著差别，从而也会导致分类结果有显著不同。



01 k近邻学习

13

分析1NN二分类错误率 $P(err)$

暂且假设距离计算是“恰当”的，即能够恰当地找出 k 个近邻，我们来对“最近邻分类器”（1NN，即 $k=1$ ）在二分类问题上的性能做一个简单的讨论。给定测试样本 \mathbf{x} ，若其最近邻样本为 \mathbf{z} ，则最近邻出错的概率就是 \mathbf{x} 与 \mathbf{z} 类别标记不同的概率，即

$$P(err) = 1 - \sum_{c \in \mathcal{Y}} P(c|\mathbf{x})P(c|\mathbf{z})$$

01 k近邻学习

14

分析1NN二分类错误率 $P(err)$

- 假设样本独立同分布，且对任意 \mathbf{x} 和任意小正整数 δ ，在 \mathbf{x} 附近 δ 距离范围内总能找到一个训练样本；换言之，对任意测试样本，总能在任意近的范围找到

$P(err) = 1 - \sum_{c \in y} P(c|\mathbf{x})P(c|\mathbf{z})$ 中的训练样本 \mathbf{z} 。

- 令 $c^* = \arg \max_{c \in y} P(c|\mathbf{x})$ 表示贝叶斯最优分类器的结果，有

$$\begin{aligned} P(err) &= 1 - \sum_{c \in y} P(c|\mathbf{x})P(c|\mathbf{z}) \approx 1 - \sum_{c \in y} P(c|\mathbf{x})^2 \\ &\leq 1 - P(c^*|\mathbf{x})^2 = (1 + P(c^*|\mathbf{x}))(1 - P(c^*|\mathbf{x})) \\ &\leq 2 \times (1 - P(c^*|\mathbf{x})) \end{aligned}$$

- 最近邻分类虽简单，但它的泛化错误率不超过贝叶斯最优分类器错误率的两倍！

01 k近邻学习

15

K近邻学习没有显式的训练过程，属于“懒惰学习”

- “懒惰学习” (lazy learning): 此类学习技术在训练阶段仅仅是把样本保存起来，训练时间开销为零，待收到测试样本后再进行处理。
- “急切学习” (eager learning): 在训练阶段就对样本进行学习处理的方法。

01 **k**近邻学习

02 降维概述

03 低维嵌入

04 主成分分析

02 降维概述

17

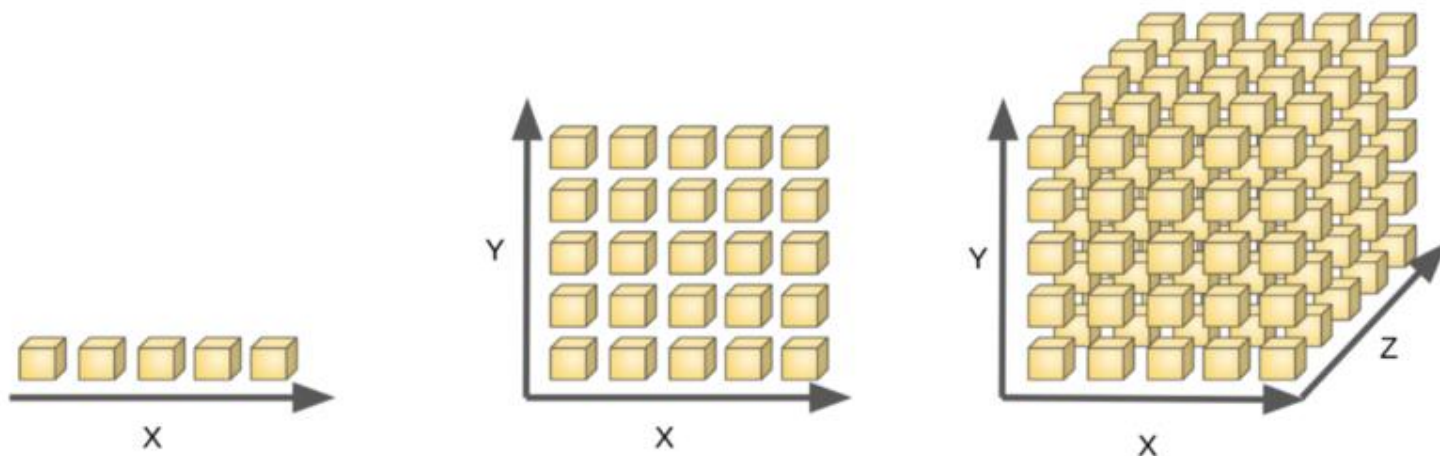
上述讨论基于一个重要的假设：任意测试样本 x 附近的任意小的 δ 距离范围内总能找到一个训练样本，即训练样本的采样密度足够大，或称为“密采样”。然而，这个假设在现实任务中通常很难满足：

- 若属性维数为1，当 $\delta=0.001$ ，仅考虑单个属性，则仅需1000个样本点平均分布在归一化后的属性取值范围内，即可使得任意测试样本在其附近0.001距离范围内总能找到一个训练样本，此时kNN分类器错误率不超过贝叶斯最优分类器错误率的两倍。若属性维数为20，若样本满足密采样条件，则至少需要 $(10^3)^{20}=10^{60}$ 个样本。
- 现实应用中属性维数经常成千上万，要满足密采样条件所需的样本数目是无法达到的天文数字。

02 降维概述

18

- 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。



02 降维概述

19

- 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**。

02 降维概述

20

什么是降维？

缓解维数灾难的一个重要途径是降维(dimension reduction)

- 即通过某种数学变换，将原始高维属性空间转变为一个低维“子空间”(subspace)，在这个子空间中样本密度大幅度提高，距离计算也变得更为容易。
- 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。

02 降维概述

21

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

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

02 降维概述

22

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。

01 **k**近邻学习

02 降维概述

03 低维嵌入

04 主成分分析

03 低纬嵌入

24

若要求原始空间中样本之间的距离在低维空间中得以保持，即得到“**多维缩放**”

(Multiple Dimensional Scaling, MDS) :

- 假定有 m 个样本，在原始空间中的距离矩阵为 $\mathbf{D} \in \mathbf{R}^{m \times m}$ ，其第 i 行 j 列的元素 $dist_{ij}$ 为样本 x_i 到 x_j 的距离。
- 目标是获得样本在 d' 维空间中的表示 $\mathbf{Z} \in \mathbf{R}^{d' \times m}$ ， $d' \leq d$ ，且任意两个样本在 d' 维空间中的欧氏距离等于原始空间中的距离，即

$$\|z_i - z_j\| = dist_{ij}$$

03 低纬嵌入

25

令 $\mathbf{B} = \mathbf{Z}^T \mathbf{Z} \in \mathbf{R}^{m \times m}$, 其中 \mathbf{B} 为降维后的内积矩阵 , $b_{ij} = z_i^T z_j$, 有

$$\begin{aligned} dist_{ij}^2 &= \|z_i\|^2 + \|z_j\|^2 - 2z_i^T z_j \\ &= b_{ii} + b_{jj} - 2b_{ij} \end{aligned}$$

为便于讨论 , 令降维后的样本 \mathbf{Z} 被中心化 , 即 $\sum_{i=1}^m z_i = 0$ 。显然 , 矩阵 \mathbf{B} 的行与列之和均为零 , 即 $\sum_{i=1}^m b_{ij} = \sum_{j=1}^m b_{ij} = 0$ 。

03 低纬嵌入

26

易知

$$\sum_{i=1}^m dist_{ij}^2 = tr(\mathbf{B}) + mb_{jj}$$

$$\sum_{j=1}^m dist_{ij}^2 = tr(\mathbf{B}) + mb_{ii}$$

$$\sum_{i=1}^m \sum_{j=1}^m dist_{ij}^2 = 2mtr(\mathbf{B})$$

其中 $tr(\cdot)$ 表示矩阵的迹(trace) ,

$$tr(\mathbf{B}) = \sum_{i=1}^m \|z_i\|^2 \text{ 。 令}$$

$$dist_{i\cdot}^2 = \frac{1}{m} \sum_{j=1}^m dist_{ij}^2$$

$$dist_{\cdot j}^2 = \frac{1}{m} \sum_{i=1}^m dist_{ij}^2$$

$$dist_{\cdot\cdot}^2 = \frac{1}{m^2} \sum_{i=1}^m \sum_{j=1}^m dist_{ij}^2$$

03 低纬嵌入

27

由此即可通过降维前后保持不变的距离矩阵 \mathbf{D} 求取内积矩阵 \mathbf{B}

$$b_{ij} = -\frac{1}{2}(\text{dist}_{ij}^2 - \text{dist}_{i.}^2 - \text{dist}_{.j}^2 + \text{dist}_{..}^2)$$

对矩阵 \mathbf{B} 做特征值分解(eigenvalue decomposition) $\mathbf{B} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$, 其中 $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_d)$ 为特征值构成的对角矩阵 , $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d$, \mathbf{V} 为特征向量矩阵 , 假定其中有 d^* 个非零特征值, 它们构成对角矩阵 $\mathbf{\Lambda}_* = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_{d^*})$, 令 \mathbf{V}_* 表示相应的特征向量矩阵 , 则 \mathbf{Z} 可表达为 $\mathbf{Z} = \mathbf{\Lambda}_*^{\frac{1}{2}}\mathbf{V}_*^T \in \mathbf{R}^{d^* \times m}$ 。

03 低维嵌入

28

在现实应用中为了有效降维，往往仅需降维后的距离与原始空间中的距离尽可能接近，而不必严格相等。此时可取 $d' \ll d$ 个最大特征值构成对角矩阵 $\tilde{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_{d'})$ ，令 $\tilde{\Lambda}$ 表示相应的特征向量矩阵，则 \mathbf{Z} 可表达为

$$\mathbf{Z} = \tilde{\Lambda}^{\frac{1}{2}} \tilde{\mathbf{V}}^T \in \mathbf{R}^{d' \times m}$$

03 低维嵌入

29

“多维缩放”（Multiple Dimensional Scaling, MDS）算法描述

输入：距离矩阵 $\mathbf{D} \in \mathbf{R}^{m \times m}$ ，其元素 $dist_{ij}$ 为样本 x_i 到 x_j 的距离；
低维空间维数 d' 。

过程：

- 1：计算 $dist_{i.}^2$ ， $dist_{.j}^2$ ， $dist_{ij}^2$ ；
- 2：计算矩阵 \mathbf{B} ；
- 3：对矩阵 \mathbf{B} 做特征值分解；
- 4：取 $\tilde{\Lambda}$ 为 d' 个最大特征值所构成的对角矩阵， $\tilde{\mathbf{V}}$ 为相应的特征向量矩阵。

输出：矩阵 $\tilde{\mathbf{V}}\tilde{\Lambda}^{\frac{1}{2}} \in \mathbf{R}^{m \times d'}$ ，每行是一个样本的低维坐标

03 低维嵌入

30

一般来说，欲获得低维子空间，最简单的是对原始高维空间进行线性变换。

线性降维方法：

给定 d 维空间的样本 $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m) \in \mathbb{R}^{d \times m}$ ，变换之后得到 $d' \leq d$ 维空间中的样本

$$\mathbf{Z} = \mathbf{W}^T \mathbf{X}$$

其中 $\mathbf{W} \in \mathbb{R}^{d \times d'}$ 是变换矩阵， $\mathbf{Z} \in \mathbb{R}^{d' \times m}$ 是样本在新空间中的表达。

03 低纬嵌入

31

变换矩阵 \mathbf{W} 可视为 d' 个 d 维属性向量。换言之， z_i 是原属性向量 x_i 在新坐标系 $\{w_1, w_2, \dots, w_{d'}\}$ 中的坐标向量。若 w_i 与 w_j ($i \neq j$) 正交，则新坐标系是一个正交坐标系，此时 \mathbf{W} 为正交变换。显然，新空间中的属性是原空间中的属性的线性组合。

基于线性变换来进行降维的方法称为线性降维方法，对低维子空间性质的不同要求可通过对 \mathbf{W} 施加不同的约束来实现。

01 **k**近邻学习

02 降维概述

03 低维嵌入

04 主成分分析

04 主成分分析

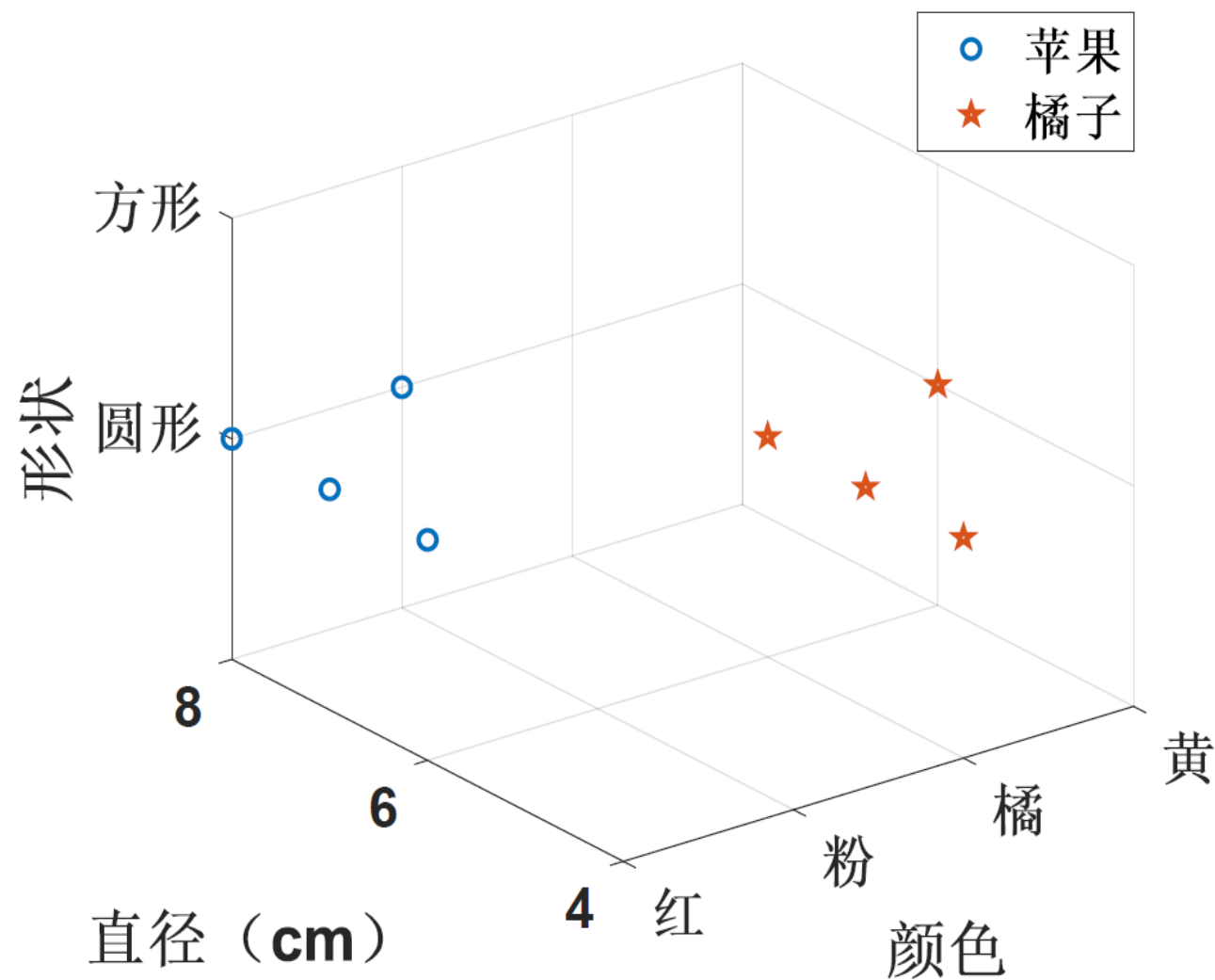
33

- **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。

04 主成分分析

34

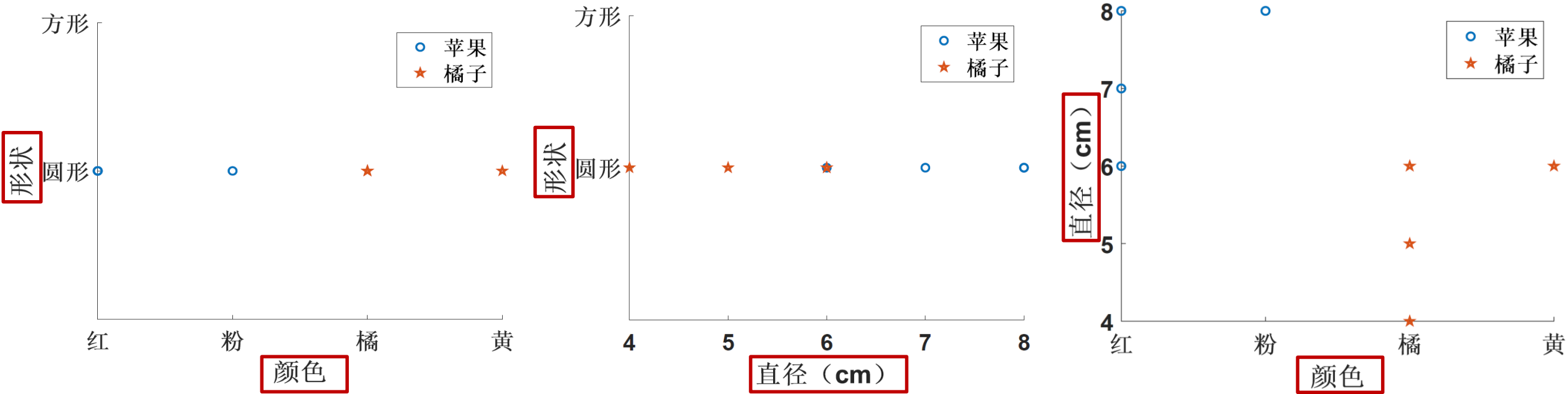
例如：样本 $X = (x_1, x_2, x_3, \dots, x_8)$ ，样本向量属性维数为3（**颜色**，**直径**，**形状**），原始分布如图所示。



04 主成分分析

35

降维后分布如图所示



04 主成分分析

36

统计分析中，数据的变量之间可能存在相关性，以致增加了分析的难度。于是，考虑由少数不相关的变量来代替相关的变量，用来表示数据，并且要求能够保留数据中的大部分信息。

对于正交属性空间中的样本点，如何用一个超平面对所有样本进行恰当的表达？

若存在这样的超平面，那么它大概应具有这样的性质：

- **最近重构性**：样本点到这个超平面的距离都足够近；
- **最大可分性**：样本点在这个超平面上的投影能尽可能分开。

04 主成分分析

37

基本思想

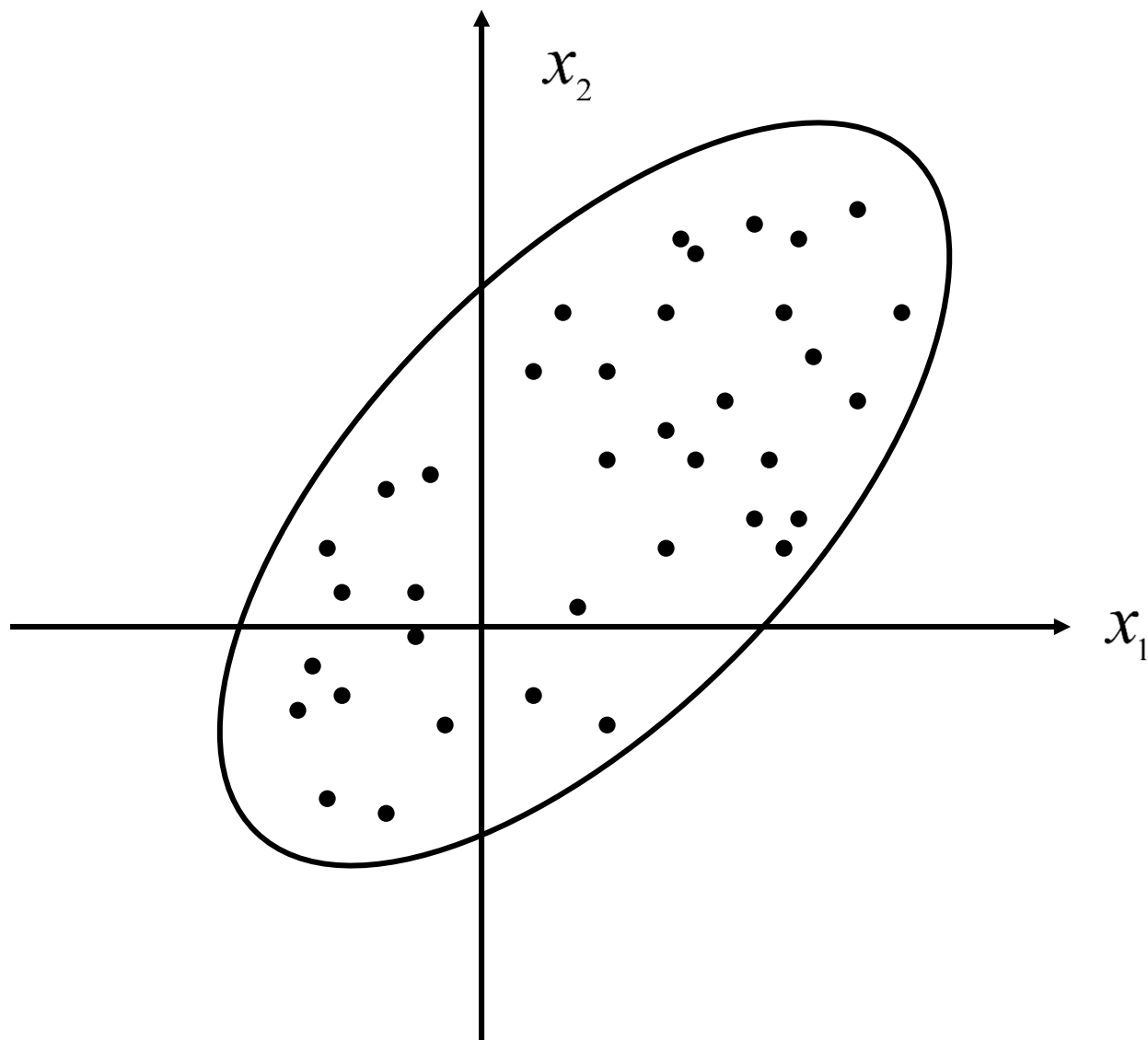
在主成分分析中，首先对给定数据进行规范化，是的数据每一变量的平均值为0，方差为1。之后对数据进行正交变换，原来由线性相关变量表示的数据，通过正交变换变成由若干个线性无关的新变量表示的数据。新变量是可能的正交变换中变量的方差的和（信息保存）最大的，方差表示在新变量上信息的大小。将新变量依次称为第一主成分、第二主成分等。

04 主成分分析

38

直观解释

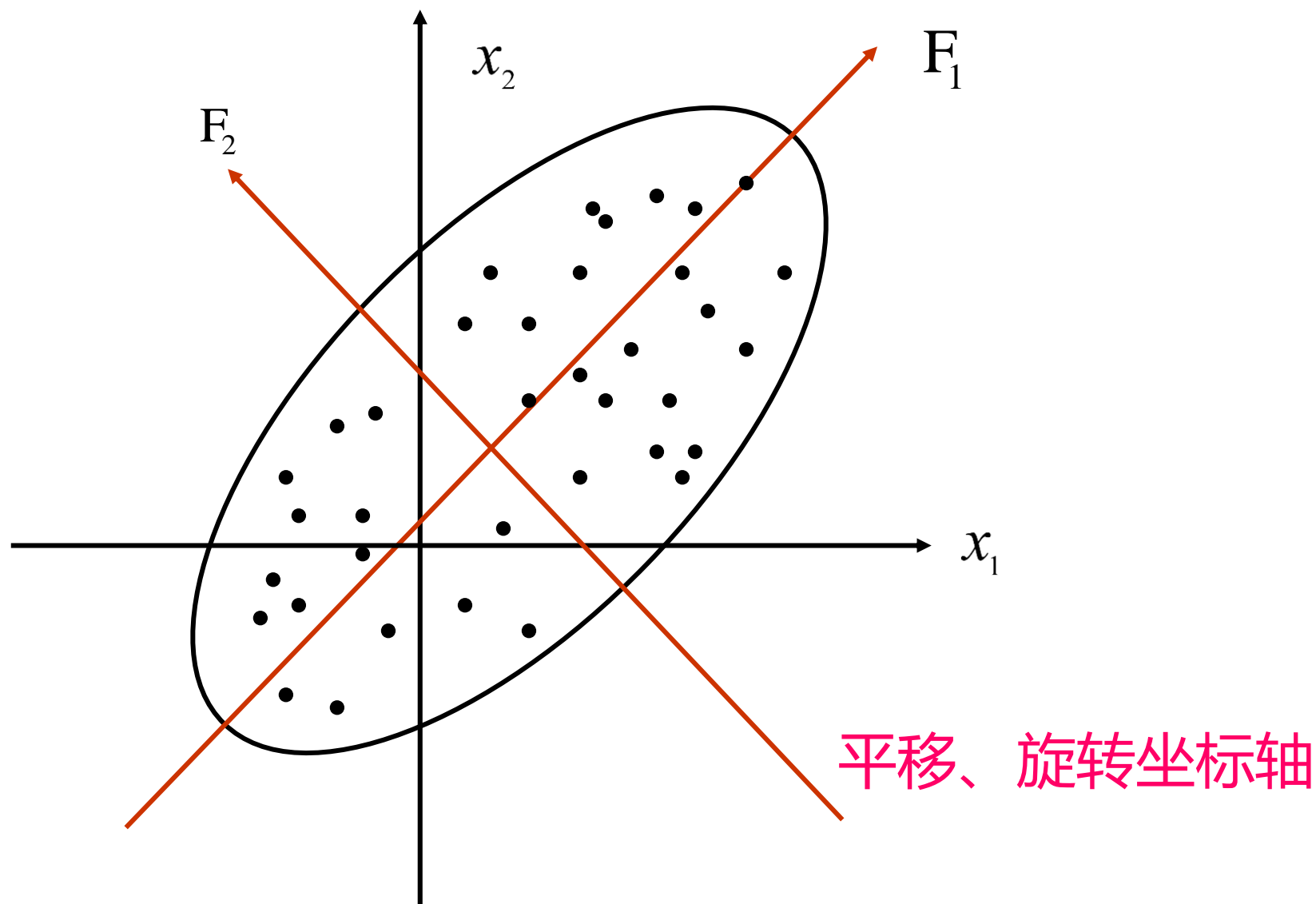
数据结合中的样本由实数空间（正交坐标系）中的点表示，空间的一个坐标轴表示一个变量，规范化处理后得到的数据分布在原点附近。



04 主成分分析

39

主成分分析的几何解释



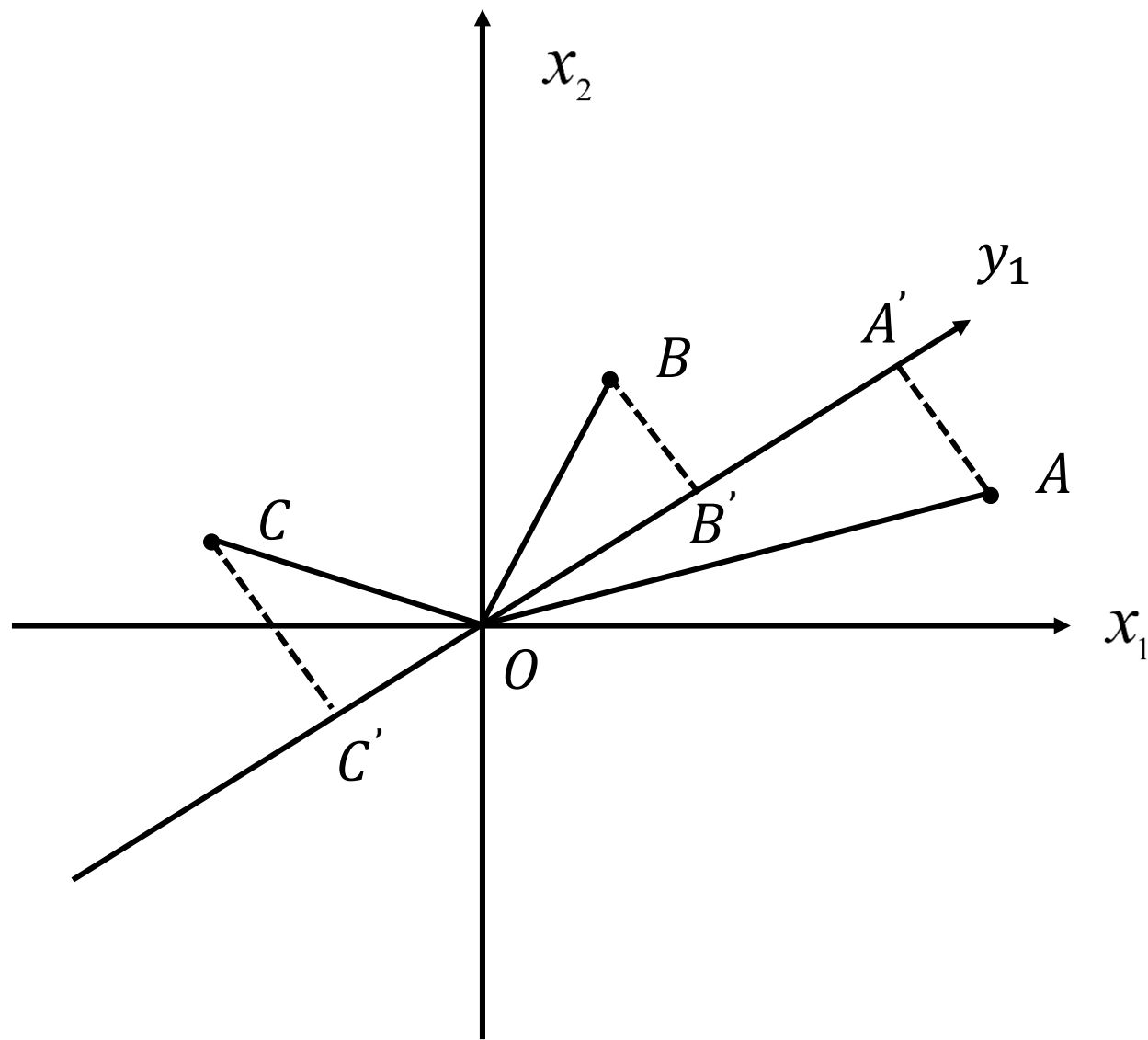
04 主成分分析

40

假设有两个变量 x_1 和 x_2 ，三个样本点 A, B, C ，样本分布在由 x_1 和 x_2 轴组成的坐标系中。

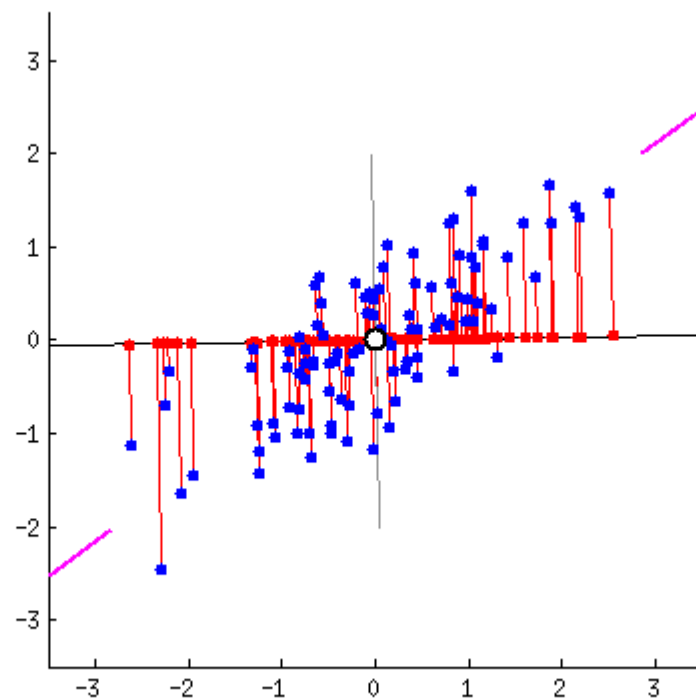
样本点 A, B, C 在 y_1 轴上投影，得到 y_1 轴的坐标 A', B', C' 。

主成分分析在旋转变换中选取样本点的距离平方和最小的轴，作为第一主成分。第二主成分等的选取，在保证与已选坐标轴正交的条件下，类似地进行。



04 主成分分析

41



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。

04 主成分分析

42

1. 定义和导出

$\mathbf{x} = (x_1, x_2, \dots, x_m)^T$ 是 m 维随机变量，其均值向量是 $\boldsymbol{\mu}$

$$\boldsymbol{\mu} = E(\mathbf{x}) = (\mu_1, \mu_2, \dots, \mu_m)^T$$

协方差矩阵是 Σ

$$\Sigma = cov(\mathbf{x}, \mathbf{x}) = E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T]$$

考虑由 m 维随机变量 \mathbf{x} 到 m 维随机变量 $\mathbf{y} = (y_1, y_2, \dots, y_m)^T$ 的线性变换

$$y_i = \alpha_i^T \mathbf{x} = \alpha_{1i}x_1 + \alpha_{2i}x_2 + \dots + \alpha_{mi}x_m$$

其中 $\alpha_i^T = (\alpha_{1i}, \alpha_{2i}, \dots, \alpha_{mi})$, $i = 1, 2, \dots, m$

04 主成分分析

43

由随机变量的性质可知，

$$E(y_i) = \alpha_i^T, \quad i = 1, 2, \dots, m$$

$$\text{var}(y_i) = \alpha_i^T \Sigma \alpha_i, \quad i = 1, 2, \dots, m$$

$$\text{cov}(y_i, y_j) = \alpha_i^T \Sigma \alpha_j, \quad i = 1, 2, \dots, m, \quad j = 1, 2, \dots, m$$

在数据总体（population）上进行的主成分分析称为**总体主成分分析**，在有限样本上进行的主成分分析称为**样本主成分分析**，前者是后者的基础。

04 主成分分析

44

总体主成分

给定一个的 $y_i = \alpha_i^T \mathbf{x} = \alpha_{1i}x_1 + \alpha_{2i}x_2 + \dots + \alpha_{mi}x_m$ 的线性变换，如果他们满足下列条件：

- (1) 系数向量 α_i^T 是单位向量，即 $\alpha_i^T \alpha_i = 1, i = 1, 2, \dots, m$ ；
- (2) 变量 y_i 与 y_j 互不相关，即 $cov(y_i, y_j) = 0 (i \neq j)$ ；
- (3) 变量 y_1 是 \mathbf{x} 的所有线性变换中方差最大的； y_2 是与 y_1 不相关的 \mathbf{x} 的所有线性变换中方差最大的；一般地， y_i 是与 $y_1, y_2, \dots, y_{i-1} (i = 1, 2, \dots, m)$ 都不相关的 \mathbf{x} 的所有线性变换中方差最大的；这时分别称 y_1, y_2, \dots, y_m 为 \mathbf{x} 的第一主成分、第二主成分、...、第 m 主成分。

04 主成分分析

45

定义中的条件（1）表明线性变换是正交变换， $\alpha_1, \alpha_2, \dots, \alpha_m$ 是其一组标准正交基，

$$\alpha_i^T \alpha_j = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

条件（2）（3）给出了一个求主成分的方法：第一步，在 \mathbf{x} 的所有线性变换中，

$$\alpha_1^T \mathbf{x} = \sum_{i=1}^m \alpha_{i1} x_i$$

在 $\alpha_1^T \alpha_1 = 1$ 条件下，求方差最大的，得到 \mathbf{x} 的第一主成分。

04 主成分分析

46

第二步，在与 $\alpha_1^T \mathbf{x}$ 不相关的 \mathbf{x} 的所有线性变换中，

$$\alpha_2^T \mathbf{x} = \sum_{i=1}^m \alpha_{i2} x_i$$

在 $\alpha_2^T \alpha_2 = 1$ 条件下，求方差最大的，得到 \mathbf{x} 的第二主成分。

第 k 步，在与 $\alpha_1^T \mathbf{x}$ ， $\alpha_2^T \mathbf{x}$ ， \dots ， $\alpha_{k-1}^T \mathbf{x}$ 不相关的 \mathbf{x} 的所有线性变换中，

$$\alpha_k^T \mathbf{x} = \sum_{i=1}^m \alpha_{ik} x_i$$

在 $\alpha_2^T \alpha_2 = 1$ 条件下，求方差最大的，得到 \mathbf{x} 的第 k 主成分。如此继续下去，知道得到 \mathbf{x} 的第 m 主成分。

04 主成分分析

47

2. 主要性质

【定理】设 \mathbf{x} 是 m 维随机变量， Σ 是 \mathbf{x} 的协方差矩阵， Σ 的特征值分别是 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m \geq 0$ ，特征值对应的单位特征向量分别是 $\alpha_1, \alpha_2, \dots, \alpha_m$ ，则 \mathbf{x} 的第 k 主成分是

$$y_k = \alpha_k^T \mathbf{x} = \alpha_{1k}x_1 + \alpha_{2k}x_2 + \dots + \alpha_{mk}x_m$$

\mathbf{x} 的第 k 主成分的方差是

$$\text{var}(y_k) = \alpha_k^T \Sigma \alpha_k = \lambda_k, \quad k = 1, 2, \dots, m$$

即协方差矩阵 Σ 的第 k 个特征值。

04 主成分分析

48

证明：采用拉格朗日乘子法求出主成分

首先求 \mathbf{x} 的第一主成分 $y_1 = \alpha_1^T \mathbf{x}$ ，即求系数向量 α_1 。由总体主成分定义知，第一主成分 α_1 是在 $\alpha_1^T \alpha_1 = 1$ 条件下， \mathbf{x} 的所有线性变换中使方差

$$\text{var}(\alpha_1^T \mathbf{x}) = \alpha_1^T \Sigma \alpha_1$$

达到最大。

求第一主成分就是求解约束最优化问题：

$$\max_{\alpha_1} \alpha_1^T \Sigma \alpha_1$$

$$\text{s.t. } \alpha_1^T \alpha_1 = 1$$

04 主成分分析

49

定义拉格朗日函数

$$\alpha_1^T \Sigma \alpha_1 - \lambda(\alpha_1^T \alpha_1 - 1)$$

其中 λ 是拉格朗日乘子。将拉格朗日函数对 α_1 求导，并令其为0，得

$$\Sigma \alpha_1 - \lambda \alpha_1 = 0$$

因此， λ 是 Σ 的特征值， α_1 是对应的单位特征向量。于是，目标函数

$$\alpha_1^T \Sigma \alpha_1 = \alpha_1^T \lambda \alpha_1 = \lambda \alpha_1^T \alpha_1 = \lambda$$

假设 α_1 是 Σ 的最大特征值 λ_1 对应的单位特征向量，显然 α_1 与 λ_1 是最优化问题的解。所以， $\alpha_1^T x$ 构成第一主成分，其方差等于协方差矩阵的最大特征值。

$$\text{var}(\alpha_1^T \mathbf{x}) = \alpha_1^T \Sigma \alpha_1 = \lambda_1$$

04 主成分分析

50

接着求 x 的第二主成分 $y_2 = \alpha_2^T x$ 。第二主成分的 α_2 是在 $\alpha_2^T \alpha_2 = 1$ ，且 $\alpha_2^T x$ 与 $\alpha_1^T x$ 不相关的条件下， x 的所有线性变换中使方差

$$\text{var}(\alpha_2^T x) = \alpha_2^T \Sigma \alpha_2$$

达到最大的。

求第二主成分需要求解约束最优化问题

$$\max_{\alpha_2} \alpha_2^T \Sigma \alpha_2$$

$$\text{s.t. } \alpha_1^T \Sigma \alpha_2 = 0, \quad \alpha_2^T \Sigma \alpha_1 = 0$$

$$\alpha_2^T \alpha_2 = 1$$

04 主成分分析

51

注意到

$$\alpha_1^T \Sigma \alpha_2 = \alpha_2^T \Sigma \alpha_1 = \alpha_2^T \lambda_1 \alpha_1 = \lambda_1 \alpha_2^T \alpha_1 = \lambda_1 \alpha_1^T \alpha_2$$

以及

$$\alpha_1^T \alpha_2 = 0, \quad \alpha_2^T \alpha_1 = 0$$

定义拉格朗日函数

$$\alpha_2^T \Sigma \alpha_2 - \lambda (\alpha_2^T \alpha_2 - 1) - \phi \alpha_2^T \alpha_1$$

其中 λ , ϕ 是拉格朗日乘子。对 α_2 求导, 并令其为0, 得

$$2\Sigma \alpha_2 - 2\lambda \alpha_2 - \phi \alpha_1 = 0$$

将方程左乘以 α_1^T 有

$$2\alpha_1^T \Sigma \alpha_2 - 2\lambda \alpha_1^T \alpha_2 - \phi \alpha_1^T \alpha_1 = 0$$

04 主成分分析

52

此式前两项为0，且 $\alpha_1^T \alpha_1 = 1$ ，导出 $\phi = 0$

$$2\alpha_1^T \Sigma \alpha_2 - 2\lambda \alpha_1^T \alpha_2 - \phi \alpha_1^T \alpha_1 = 0$$

因此

$$\Sigma \alpha_2 - \lambda \alpha_2 = 0$$

由此， λ 是 Σ 的特征值， α_2 是对应的单位特征向量。于是，目标函数

$$\alpha_2^T \Sigma \alpha_2 = \alpha_2^T \lambda \alpha_2 = \lambda \alpha_2^T \alpha_2 = \lambda$$

假设 α_2 是 Σ 的第二大特征值 λ_2 对应的单位特征向量，显然 α_2 与 λ_2 是以上问题的最优化问题的解。于是 $\alpha_2^T \mathbf{x}$ 构成第二主成分，其方差等于协方差矩阵的第二大特征值，

$$\text{var}(\alpha_2^T \mathbf{x}) = \alpha_2^T \Sigma \alpha_2 = \lambda_2$$

04 主成分分析

53

依此类推， \mathbf{x} 的第 k 主成分是 $\alpha_k^T \mathbf{x}$ ，并且 $\text{var}(\alpha_k^T \mathbf{x}) = \lambda_k$ ，这里 λ_k 是 Σ 的第 k 个特征值并且 α_k 是对应的单位特征向量。

按照上述方法求得第一、第二、直到第 m 主成分，其系数向量 $\alpha_1, \alpha_2, \dots, \alpha_m$ 分别是 Σ 的第一个、第二个、直到第 m 个单位特征向量， $\lambda_1, \lambda_2, \dots, \lambda_m$ 分别是对应的特征值。并且，第 k 主成分的方差等于 Σ 的第 k 个特征值，

$$\text{var}(\alpha_k^T \mathbf{x}) = \alpha_k^T \Sigma \alpha_k = \lambda_k, \quad k = 1, 2, \dots, m$$

04 主成分分析

54

【推论】 m 维随机变量 $\mathbf{y} = (y_1, y_2, \dots, y_m)^T$ 的分量依次是 \mathbf{x} 的第一主成分到第 m 主成分的充要条件是：

(1) $\mathbf{y} = A^T \mathbf{x}$, A 为正交矩阵

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mm} \end{bmatrix}$$

(2) \mathbf{y} 的协方差矩阵为对角矩阵

$$\text{cov}(\mathbf{y}) = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_m)$$

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m \geq 0$$

其中 λ_k 是 Σ 的第 k 个特征值， α_k 是对应的单位特征向量， $k = 1, 2, \dots, m$ 。

04 主成分分析

55

总体主成分的性质

(1) 总体主成分 \mathbf{y} 的协方差矩阵为对角矩阵

$$\text{cov}(\mathbf{y}) = \Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_m)$$

(2) 总体主成分 \mathbf{y} 的协方差之和等于随机变量 \mathbf{x} 的方差之和，即

$$\sum_{i=1}^m \lambda_i = \sum_{i=1}^m \sigma_{ii}$$

其中 σ_{ii} 是随机变量 x_i 的方差，即协方差矩阵 Σ 的对角元素。事实上，利用矩阵的迹 (trace) 的性质，可知

$$\sum_{i=1}^m \text{var}(x_i) = \text{tr}(\Sigma^T) = \text{tr}(A\Lambda A^T) = \text{tr}(A^T \Lambda A) = \text{tr}(\Lambda) = \sum_{i=1}^m \lambda_i = \sum_{i=1}^m \text{var}(y_i)$$

04 主成分分析

56

总体主成分的性质

(3) 第 k 个主成分 y_k 与变量 x_i 的相关系数 $\rho(y_k, x_i)$ 称为因子负荷量 (factor loading) , 它表示第 k 个主成分 y_k 与变量 x_i 的相关关系。计算公式

$$\rho(y_k, x_i) = \frac{\sqrt{\lambda_k} \alpha_{ik}}{\sqrt{\sigma_{ii}}} , \quad k = 1, 2, \dots, m$$

因为

$$\rho(y_k, x_i) = \frac{\text{cov}(y_k, x_i)}{\sqrt{\text{var}(y_k) \text{var}(x_i)}} = \frac{\text{cov}(\alpha_k^T \mathbf{x}, e_i^T \mathbf{x})}{\sqrt{\lambda_k} \sqrt{\sigma_{ii}}}$$

其中 e_i 为基本单位向量, 其第 i 个分量为1, 其余为0。再由协方差的性质

$$\text{cov}(\alpha_k^T \mathbf{x}, e_i^T \mathbf{x}) = \alpha_k^T \Sigma e_i = e_i^T \Sigma \alpha_k = \lambda_k e_i^T \alpha_k = \lambda_k \alpha_{ik}$$

由此得到 $\rho(y_k, x_i) = \frac{\sqrt{\lambda_k} \alpha_{ik}}{\sqrt{\sigma_{ii}}} , \quad k = 1, 2, \dots, m。$

04 主成分分析

57

总体主成分的性质

(4) 第 k 个主成分 y_k 与 m 个变量的因子负荷量满足

$$\sum_{i=1}^m \sigma_{ii} \rho^2(y_k, x_i) = \lambda_k$$

由式子 $\rho(y_k, x_i) = \frac{\sqrt{\lambda_k} \alpha_{ik}}{\sqrt{\sigma_{ii}}}$, $k = 1, 2, \dots, m$ 有

$$\sum_{i=1}^m \sigma_{ii} \rho^2(y_k, x_i) = \sum_{i=1}^m \lambda_k \alpha_{ik}^2 = \lambda_k \alpha_k^T \alpha_k = \lambda_k$$

04 主成分分析

58

总体主成分的性质

(5) m 个主成分与 i 个变量 x_i 的因子负荷量满足

$$\sum_{k=1}^m \rho^2(y_k, x_i) = 1$$

由于 y_1, y_2, \dots, y_m 互不相关，故

$$\rho^2(x_i, (y_1, y_2, \dots, y_m)) = \sum_{k=1}^m \rho^2(y_k, x_i)$$

又因 x_i 可以表为 y_1, y_2, \dots, y_m 的线性组合，所以 x_i 与 y_1, y_2, \dots, y_m 的相关系数的平方为 1，即

$$\rho^2(x_i, (y_1, y_2, \dots, y_m)) = 1$$

故得 $\sum_{k=1}^m \rho^2(y_k, x_i) = 1$ 。

04 主成分分析

59

主成分的个数

【方差贡献率】 第 k 主成分 y_k 的方差贡献率定义为 y_k 的方差与所有方差之和的比，记作 η_k

$$\eta_k = \frac{\lambda_k}{\sum_{i=1}^m \lambda_i}$$

k 主成分 y_k 的累计方差贡献率定义为 k 个方差之和与所有方差之和的比

$$\sum_{i=1}^k \eta_k = \frac{\sum_{i=1}^k \lambda_i}{\sum_{i=1}^m \lambda_i}$$

04 主成分分析

60

通常取 k 使得累计方差贡献率达到规定的百分比以上，例如70%~80%以上。累计方差贡献率放映了主成分保留信息的比例，但它不能放映对某个原有变量 x_i 保留信息的比例，这时通常利用 k 个主成分 y_1, y_2, \dots, y_k 对原有变量 x_i 的贡献率。

【对 x_i 的贡献率】 k 个主成分 y_1, y_2, \dots, y_k 对原有变量 x_i 的贡献率定义为 x_i 与 (y_1, y_2, \dots, y_k) 的相关系数的平方，记作 v_i

$$v_i = \rho^2(x_i, (y_1, y_2, \dots, y_k))$$

计算公式如下：

$$v_i = \rho^2(x_i, (y_1, y_2, \dots, y_k)) = \sum_{j=1}^k \rho^2(x_i, y_j) = \sum_{j=1}^k \frac{\lambda_j \alpha_{ij}^2}{\sigma_{ii}}$$

04 主成分分析

61

样本主成分

- 总体主成分分析，是定义在样本总体上的。
- 在实际问题中，需要在观测数据上进行主成分分析，这就是样本主成分分析。
- 样本主成分也和总体主成分具有相同的性质。

04 主成分分析

62

样本主成分

给定样本矩阵 X 。样本第一主成分 $y_1 = \alpha_1^T x$ 是在 $\alpha_1^T \alpha_1 = 1$ 条件下，使得 $\alpha_1^T x_j$ ($j = 1, 2, \dots, n$) 的样本方差 $\alpha_1^T S \alpha_1$ 最大的 x 的线性变换；样本第二主成分 $y_2 = \alpha_2^T x$ 是在 $\alpha_2^T \alpha_2 = 1$ 和 $\alpha_2^T x_j$ 与 $\alpha_1^T x_j$ ($j = 1, 2, \dots, n$) 的样本协方差 $\alpha_1^T S \alpha_2 = 0$ 条件下，使得 $\alpha_2^T x_j$ ($j = 1, 2, \dots, n$) 最大的 x 的线性变换；一般地，样本第 i 主成分 $y_i = \alpha_i^T x$ 是在 $\alpha_i^T \alpha_i = 1$ 和 $\alpha_i^T x_j$ 和 $\alpha_k^T x_j$ ($k < i, j = 1, 2, \dots, n$) 的样本协方差 $\alpha_k^T S \alpha_i = 0$ 条件下，使得 $\alpha_i^T x_j$ ($j = 1, 2, \dots, n$) 的样本方差 $\alpha_i^T S \alpha_i$ 最大的 x 的线性变换。

04 主成分分析

63

PCA的算法两种实现方法

(1) 基于特征值分解协方差矩阵实现PCA算法

(2) 基于奇异值分解 (Singular Value Decomposition , SVD) 分解协方差矩阵实现PCA算法 (课后了解)

04 主成分分析

64

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。

04 主成分分析

65

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 $P^{-1} = P^T$

04 主成分分析

66

主成分分析具体步骤

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) 对观测数据进行规范化处理，得到规范化数据矩阵，仍以 X 表示。

变换过程如下：

$$x_{ij}^* = \frac{x_{ij} - \bar{x}_i}{\sqrt{s_{ii}}}, \quad i = 1, 2, \dots, m, \quad j = 1, 2, \dots, n$$

其中

$$\bar{x}_i = \frac{1}{n} \sum_{j=1}^n x_{ij}, \quad i = 1, 2, \dots, m$$

$$s_{ii} = \frac{1}{n-1} \sum_{j=1}^n (x_{ij} - \bar{x}_i)^2, \quad i = 1, 2, \dots, m$$

为了方便，以下将规范化变量 x_{ij}^* 仍记作 x_{ij} ，规范化的样本矩阵仍记作 X 。

04 主成分分析

67

主成分分析具体步骤

(2) 依据规范化数据矩阵，计算样本相关矩阵 R

$$R = [r_{ij}]_{m \times m} = \frac{1}{n-1} X X^T$$

其中

$$r_{ij} = \frac{1}{n-1} \sum_{l=1}^n x_{il} x_{lj}, \quad i, j = 1, 2, \dots, m$$

04 主成分分析

68

主成分分析具体步骤

(3) 求样本相关矩阵 R 的 k 个特征值和对应的 k 个单位特征向量

求解 R 的特征方程 $|R - \lambda I| = 0$

得 R 的 m 个特征值

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m$$

求方差贡献率 $\sum_{i=1}^k \eta_i$ 达到预定值的主成分个数 k

求前 k 个特征值对应的单位特征向量 $\alpha_i = (\alpha_{1i}, \alpha_{2i}, \dots, \alpha_{mi})^T$, $i = 1, 2, \dots, k$

04 主成分分析

69

主成分分析具体步骤

(4) 求 k 个样本主成分

以 k 个单位特征向量为系数进行线性变换，求出 k 个样本主成分

$$y_i = \alpha_i^T \mathbf{x}, i = 1, 2, \dots, k$$

(5) 计算 k 个主成分 y_j 与原变量 x_i 的相关系数 $\rho(x_i, y_j)$ ，以及 k 个主成分对原变量 x_i 的贡献率 v_i 。

04 主成分分析

70

主成分分析具体步骤

(6) 计算 n 个样本的 k 个主成分值

将规范化样本数据代入 k 个主成分式 $y_i = \alpha_i^T \mathbf{x}$, $i = 1, 2, \dots, k$

得到 n 个样本的主成分值

第 j 个样本 $\mathbf{x}_j = (x_{1j}, x_{2j}, \dots, x_{mj})^T$ 的第 i 主成分值是

$$y_{ij} = (a_{1i}, a_{2i}, \dots, a_{mi})(x_{1j}, x_{2j}, \dots, x_{mj})^T = \sum_{l=1}^m a_{li}x_{lj}$$
$$i = 1, 2, \dots, m, \quad j = 1, 2, \dots, n$$

04 主成分分析

71

【例子1】

某金融服务公司为了了解贷款客户的信用程度，评价客户的信用等级，采用信用评级常用的 5C (能力，品格，担保，资本，环境) 方法对15名客户进行打分，由此判断客户违约的可能性。

由于各项指标的难易程度不同，因此需要对5项指标进行赋权，以便能够更加合理的对15名客户进行评价。

试对数据进行主成分分析。

04 主成分分析

客户编号	能力	品格	担保	资本	环境
1	66	64	65	65	65
2	65	63	63	65	64
3	57	58	63	59	66
4	67	69	65	68	64
5	61	61	62	62	63
6	64	65	63	63	63
7	64	63	63	63	64
8	63	63	63	63	63
9	65	64	65	66	64
10	67	69	69	68	67
11	62	63	65	64	64
12	68	67	65	67	65
13	65	65	66	65	64
14	62	63	64	62	66
15	64	66	66	65	67

04 主成分分析

数据规范化处理

(1) 均值归一化

计算每一列的平均值

能力	品格	担保	资本	环境
64	64.2	64.46667	64.33333	64.6

计算每一列的标准差

能力	品格	担保	资本	环境
2.77746	2.858571	1.76743	2.43975	1.352247

04 主成分分析

74

数据标准化处理

客户编号	能力	品格	担保	资本	环境
1	0.720082	-0.06997	0.301756	0.273252	0.295804
2	0.360041	-0.41979	-0.82983	0.273252	-0.44371
3	-2.52029	-2.16892	-0.82983	-2.18602	1.035314
4	1.080123	1.679161	0.301756	1.502886	-0.44371
5	-1.08012	-1.11944	-1.39562	-0.95638	-1.18322
6	0	0.27986	-0.82983	-0.5465	-1.18322
7	0	-0.41979	-0.82983	-0.5465	-0.44371
8	-0.36004	-0.41979	-0.82983	-0.5465	-1.18322
9	0.360041	-0.06997	0.301756	0.68313	-0.44371
10	1.080123	1.679161	2.564929	1.502886	1.774824
11	-0.72008	-0.41979	0.301756	-0.13663	-0.44371
12	1.440165	0.97951	0.301756	1.093008	0.295804
13	0.360041	0.27986	0.86755	0.273252	-0.44371
14	-0.72008	-0.41979	-0.26404	-0.95638	1.035314
15	0	0.629685	0.86755	0.273252	1.774824

04 主成分分析

(2) 计算协方差矩阵：计算相关系数矩阵

	能力	品格	担保	资本	环境
能力	1	0.88166	0.552924	0.927601	0.019018
品格	0.88166	1	0.715371	0.911523	0.206959
担保	0.552924	0.715371	1	0.706762	0.621637
资本	0.927601	0.911523	0.706762	1	0.129904
环境	0.019018	0.206959	0.621637	0.129904	1

(3) 计算特征值：对样本相关矩阵进行特征值分解，得到相关矩阵的特征值，并按
照大小排序。

λ_1	能力	3.435
λ_2	品格	1.223
λ_3	担保	0.179
λ_4	资本	0.099
λ_5	环境	0.046

04 主成分分析

【例子2】

- 假设有 n 个学生参加四门课程的考试，将学生们的考试成绩看作随机变量的取值，对考试成绩数据进行标准化处理，得到样本相关矩阵 R

课程	语文	外语	数学	物理
语文	1	0.44	0.29	0.33
外语	0.44	1	0.35	0.32
数学	0.29	0.35	1	0.60
物理	0.33	0.32	0.60	1

- 试对数据进行主成分分析

04 主成分分析

77

- 设变量 x_1, x_2, x_3, x_4 分别表示语文、外语、数学、物理的成绩。对样本相关矩阵进行特征值分解，得到相关矩阵的特征值，并按大小排序，

$$\lambda_1 = 2.17, \lambda_2 = 0.87, \lambda_3 = 0.57, \lambda_4 = 0.39$$

- 这些特征值就是各主成分的方差贡献率。假设要求主成分的累计方差贡献率大于75%，那么只需取前两个主成分即可，即 $k = 2$ ，因为

$$\frac{\lambda_1 + \lambda_2}{\sum_{i=1}^4 \lambda_i} = 0.76$$

04 主成分分析

78

- 求出对应于特征值 λ_1 , λ_2 的单位特征向量

项目	x_1	x_2	x_3	x_4	方差贡献率
y_1	0.460	0.476	0.523	0.537	0.543
y_2	0.574	0.486	-0.476	-0.456	0.218

由 $y_i = \alpha_i^T \mathbf{x}$, $i = 1, 2, \dots, k$ 可得第一主成分 y_1 、第二主成分 y_2

$$y_1 = 0.460x_1 + 0.476x_2 + 0.523x_3 + 0.537x_4$$

$$y_2 = 0.574x_1 + 0.486x_2 - 0.476x_3 - 0.456x_4$$

04 主成分分析

79

- 接下来由特征值和单位特征向量求出第一、第二主成分的因子负荷量，以及第一、第二主成分对变量 x_i 的贡献率

项目	x_1	x_2	x_3	x_4
y_1	0.678	0.701	0.770	0.791
y_2	0.536	0.453	-0.444	-0.425
y_1, y_2 对 x_i 的贡献率	0.747	0.697	0.790	0.806

04 主成分分析

80

- 第一主成分 y_1 对应的因子负荷量 $\rho(y_1, x_i), i = 1,2,3,4$ 均为正数，表明各门课程成绩提高都可使 y_1 提高
- 也就是说，第一主成分 y_1 反映了学生的整体成绩
- 因子负荷量的数值相近，且 $\rho(y_1, x_4)$ 的数值最大，这 表明物理成绩在整体成绩中占最重要位置

项目	x_1	x_2	x_3	x_4
y_1	0.678	0.701	0.770	0.791
y_2	0.536	0.453	-0.444	-0.425
y_1, y_2 对 x_i 的贡献率	0.747	0.697	0.790	0.806

04 主成分分析

81

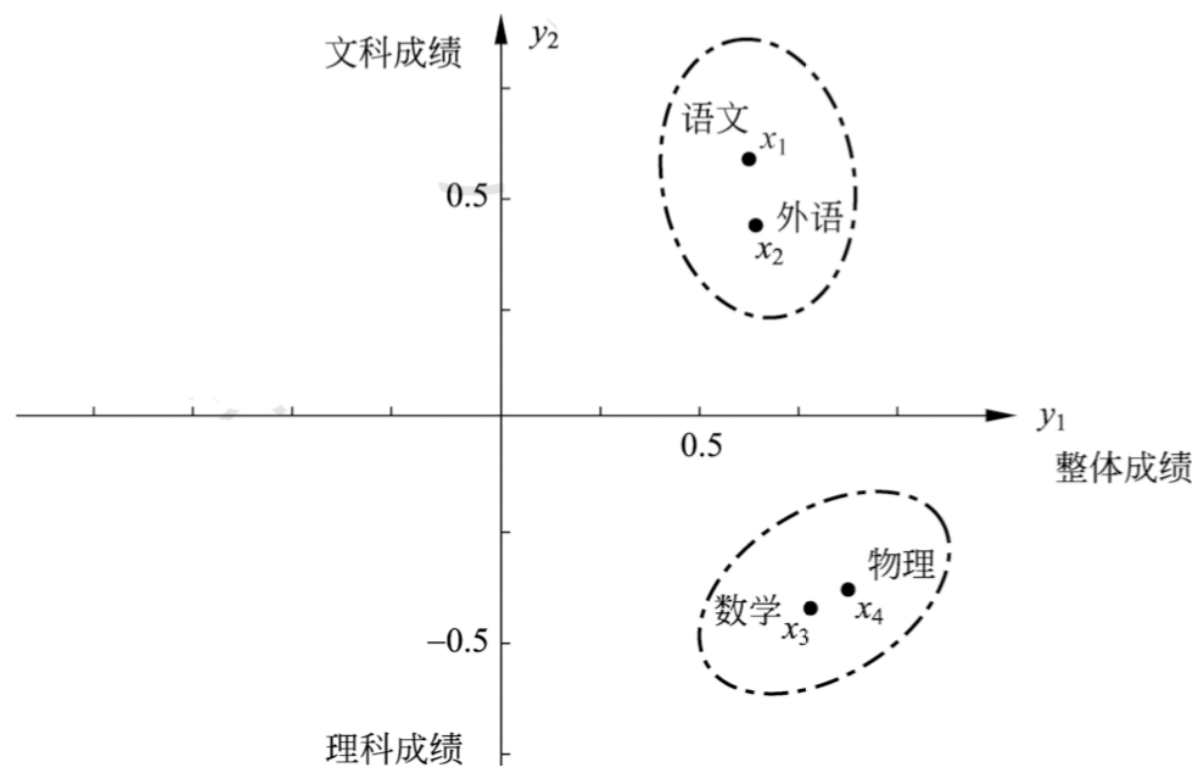
- 第二主成分 y_2 对应的因子负荷量 $\rho(y_2, x_i), i = 1,2,3,4$ 有正有负
- 正的是语文和外语，负的是数学和物理
- 表明文科成绩提高都可使 y_2 提高，理科成绩提高都可使 y_2 降低
- 也就是说，第二主成分 y_2 反映了学生的文科成绩与理科成绩的关系。

项目	x_1	x_2	x_3	x_4
y_1	0.678	0.701	0.770	0.791
y_2	0.536	0.453	-0.444	-0.425
y_1, y_2 对 x_i 的贡献率	0.747	0.697	0.790	0.806

04 主成分分析

82

- 将原变量 x_1, x_2, x_3, x_4 （语文、外语、数学、物理）和主成分 y_1, y_2 （整体成绩、文科对理科成绩）的因子负荷量在平面坐标系中表示。
- 4个原变量聚成了两类：因子负荷量相近的语文、外语为一类，数学、物理为一类，前者反映文科课程成绩，后者反映理科课程成绩。



04 主成分分析

83

【例子3】

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

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：

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

04 主成分分析

84

然后求 Σ 的特征值和特征向量：

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

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： $\Sigma_1 \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \Sigma_2 \begin{pmatrix} -1 \\ 1 \end{pmatrix}$

04 主成分分析

85

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：

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

04 主成分分析

86

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：

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

04 主成分分析

87

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



谢谢！