# 第9章 降维

1. $k$近邻学习
2. 主成分分析

机器学习-第10章（10.1和10.3）

统计学习方法-第3、16章全部

SUN YAT-SEN UNIVERSITY 中山大學

# 第9章 降维

1. $k$近邻学习

2. 主成分分析

机器学习-第10章（10.1和10.3）

统计学习方法-第3、16章全部

# $k$近邻学习

$k$近邻(k-Nearest Neighbor，简称kNN)学习是一种常用的监督学习方法，

- 工作机制：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$k$个训练样本，然后基于这$k$个"邻居"的信息来进行预测

# $k$近邻学习

$k$近邻(k-Nearest Neighbor，简称kNN)学习是一种常用的监督学习方法，

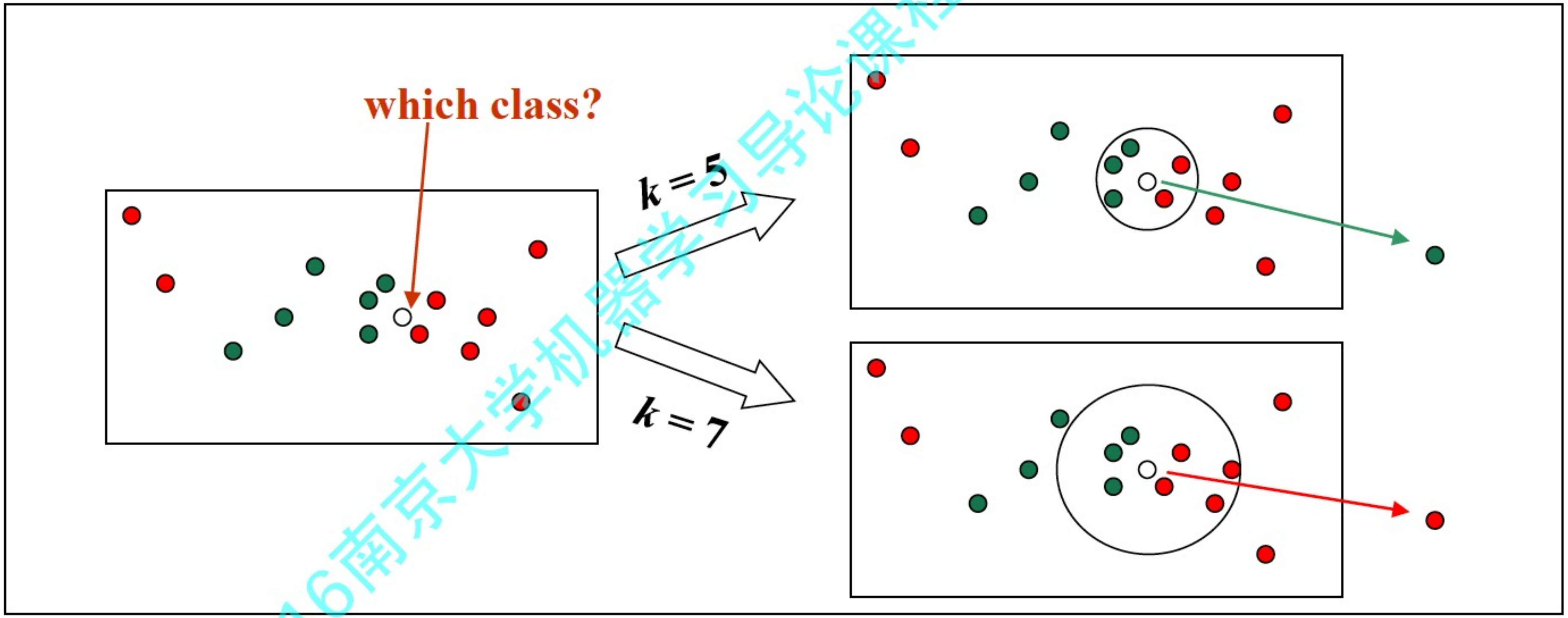
- 工作机制：给定测试样本，基于某种距离度量找出训练集中与其最靠近的$k$个训练样本，然后基于这$k$个"邻居"的信息来进行预测
- 在分类任务中：使用“投票法”，即选择这$k$个样本中出现最多的类别标记作为预测结果
- 在回归任务中：使用“平均法”，即将这$k$个样本的实值输出标记的平均值作为预测结果；还可基于距离远近进行加权平均或加权投票，距离越近的样本权重越大.

# $k$近邻学习

$k$ 近邻 ($k$-Nearest Neighbor, $k$NN)

懒惰学习 (lazy learning) 的代表

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



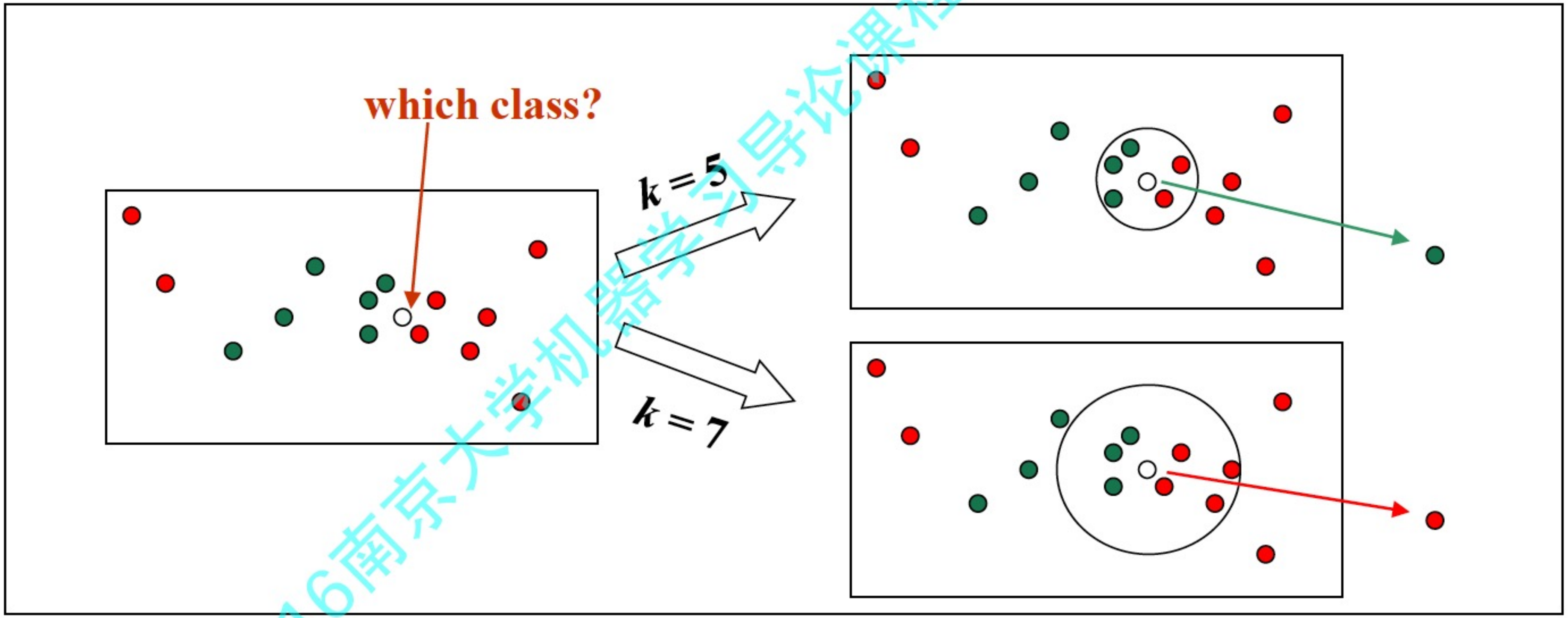
# $k$近邻学习

$k$ 近邻 ($k$-Nearest Neighbor, $k$NN)

懒惰学习 (lazy learning) 的代表

基本思路：
近朱者赤，近墨者黑

(投票法；平均法)




关键： $k$ 值选取；距离计算

# $k$近邻学习（用于分类问题时）

$k$近邻（$k$-Nearest Neighbor, kNN）学习是一种**监督学习方法**

数据：$D=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_m,y_m)\}$，$\boldsymbol{x}_i\in\mathbb{R}^d, y_i\in\mathcal{Y}$

模型：用训练集、$k$值、距离度量及分类规则对特征空间的划分

策略：不具有显示的学习过程

# $k$近邻学习（用于分类问题时）

$k$近邻（$k$-Nearest Neighbor, kNN）学习是一种**监督学习方法**

数据：$D=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_m,y_m)\}$，$\boldsymbol{x}_i\in\mathbb{R}^d, y_i\in\mathcal{Y}$

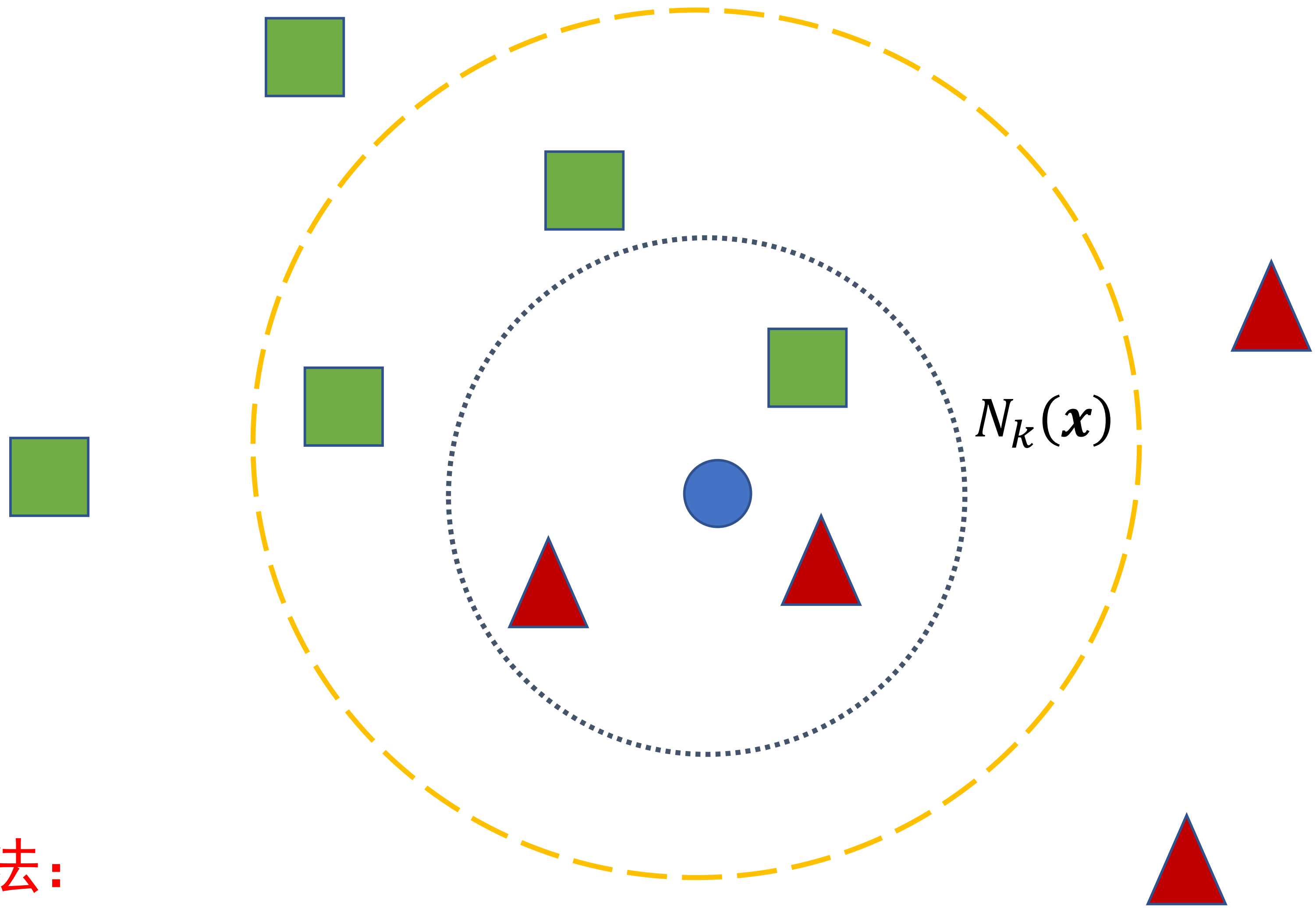
模型：用训练集、$k$值、距离度量及分类规则对特征空间的划分

策略：不具有显示的学习过程

算法：

- 根据给定的距离度量，在训练集$D$中找出与$\boldsymbol{x}$最邻近的$k$个点，涵盖这$k$个点的邻域记作$N_k(\boldsymbol{x})$
- 在$N_k(\boldsymbol{x})$中根据分类决策规则（如多数表决）决定$\boldsymbol{x}$的类别$y$

# kNN基本想法




算法：

- 根据给定的距离度量，在训练集 $D$ 中找出与 $\boldsymbol{x}$ 最邻近的 $k$ 个点，涵盖这$k$个点的邻域记作$N_k(\boldsymbol{x})$
- 在$N_k(\boldsymbol{x})$中根据分类决策规则（如多数表决）决定$\boldsymbol{x}$的类别$y$

# $k$近邻学习（用于分类问题时）

$k$近邻（$k$-Nearest Neighbor, kNN）学习是一种**监督学习方法**

数据：$D=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_m,y_m)\}$，$\boldsymbol{x}_i\in\mathbb{R}^d, y_i\in\mathcal{Y}$

模型：用训练集、$k$值、距离度量及分类规则对特征空间的划分

策略：不具有显示的学习过程

算法：

- 根据给定的距离度量，在训练集$D$中找出与$\boldsymbol{x}$最邻近的$k$个点，涵盖这$k$个点的邻域记作$N_k(\boldsymbol{x})$
- 在$N_k(\boldsymbol{x})$中根据分类决策规则（如多数表决）决定$\boldsymbol{x}$的类别$y$

核心思想：
如果一个样本在特征空间中的$k$个最相邻的样本大多都属于某一个类别，则该样本也属于这个类别，并具有这个类别样本的性质

# k近邻

## 算法要素

- ➢K值：$k$是一个重要参数，当$k$取不同值时，分类结果会有显著不同.
- ➢距离度量方式：不同的距离计算方式，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.
- ➢分类决策规则

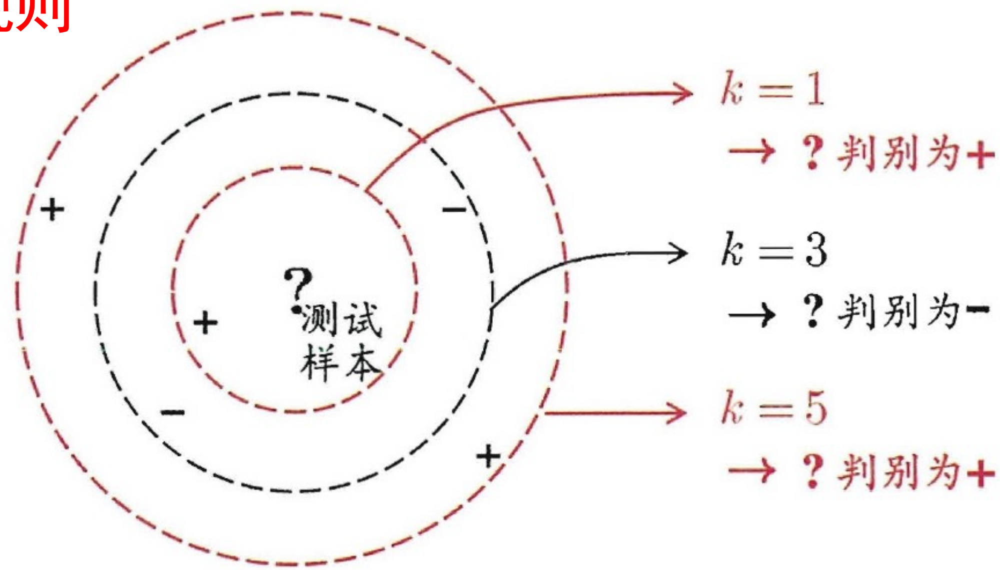
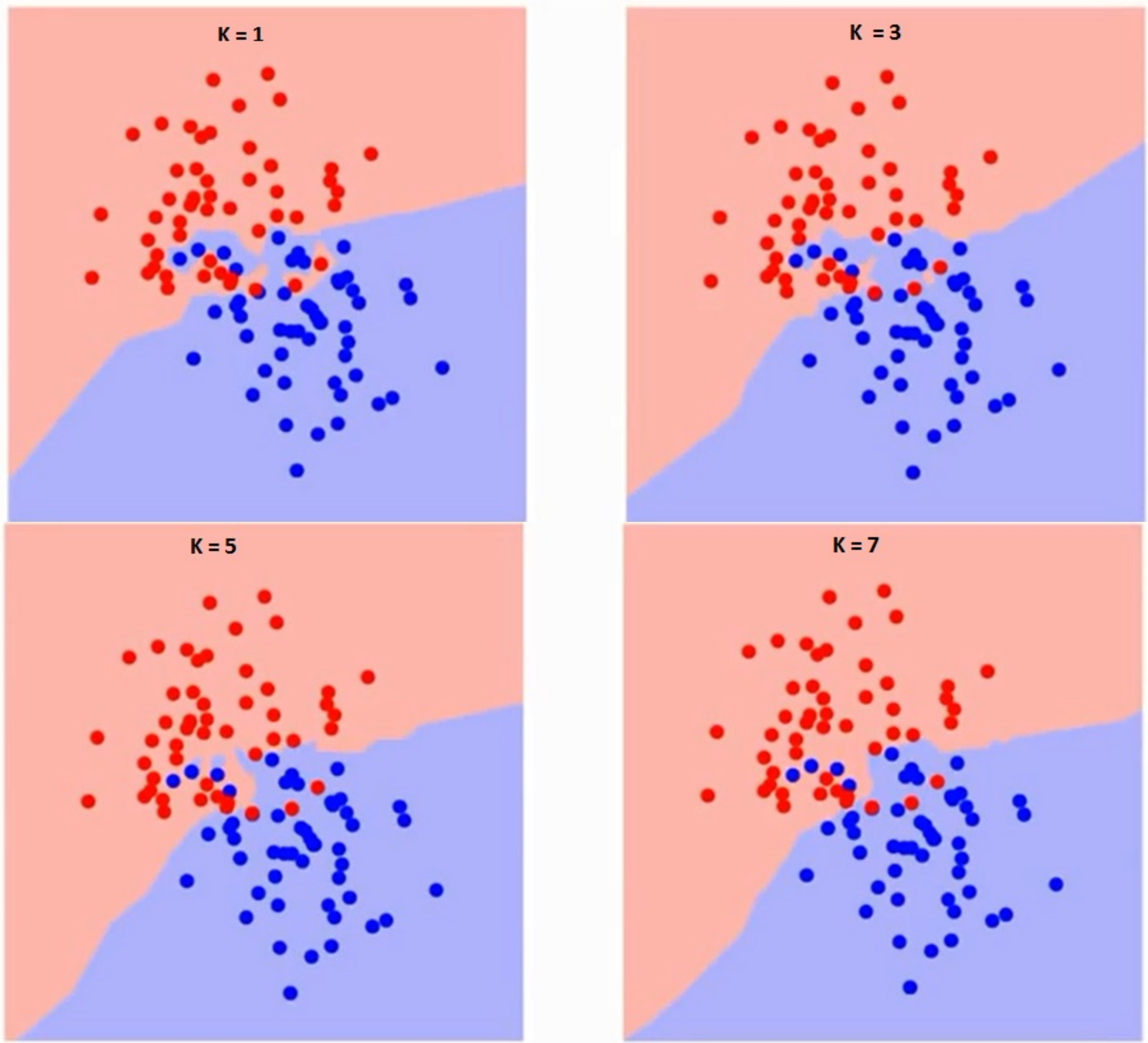



**图 10.1** $k$ 近邻分类器示意图. 虚线显示出等距线; 测试样本在 $k=1$ 或 $k=5$ 时被判别为正例, $k=3$ 时被判别为反例.




**K值选取**：

根据样本分布

较小$k$值：训练误差减小，容易发生过拟合

较大$k$值：减少泛化误差，但训练误差增加

https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/?#?&utm_source=coding-window-blog&source=coding-window-blog

# kNN算法特点

(1) 不需要提前训练

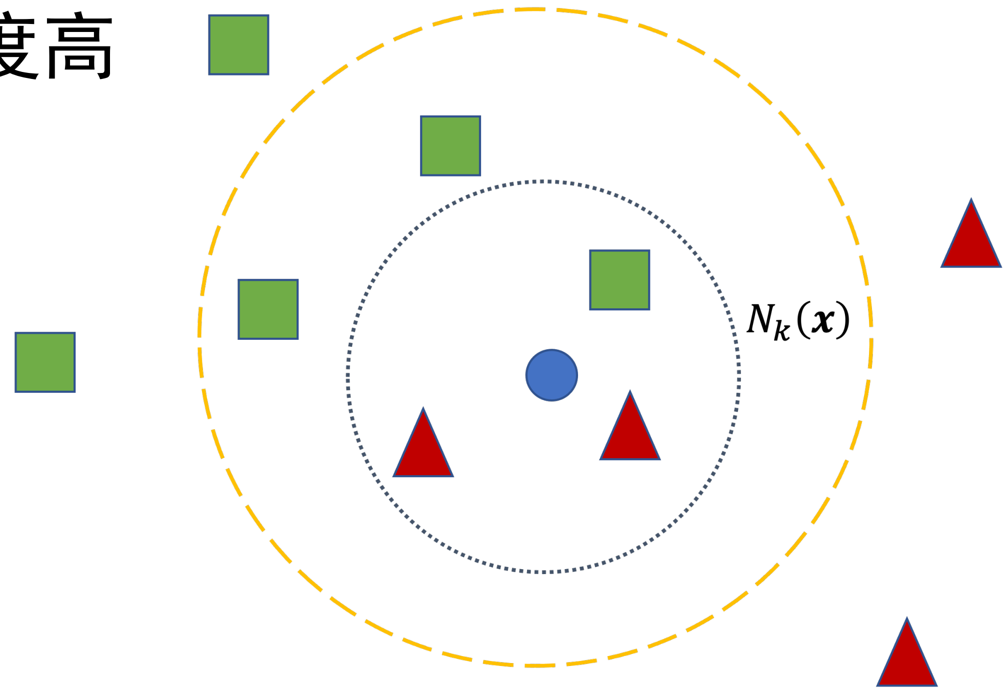
(2) 基于样本之间的距离和决策机制

(3) 简单易实现

(4) 精度高，对异常值不太敏感

(5) 计算复杂度高，空间复杂度高

(6) 理想$k$值难决定



# 第9章 降维

1. $k$近邻学习

2. 主成分分析

机器学习-第10章（10.1和10.3）

统计学习方法-第3、16章全部

# 绪论——无监督学习

## 无监督学习：从无标注数据中学习分析模型的机器学习问题

无标注数据是“自然”得到的数据，分析模型表示数据的类别、转换等

无标注数据

训练集（training set）$D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_m\}$

特征（属性）

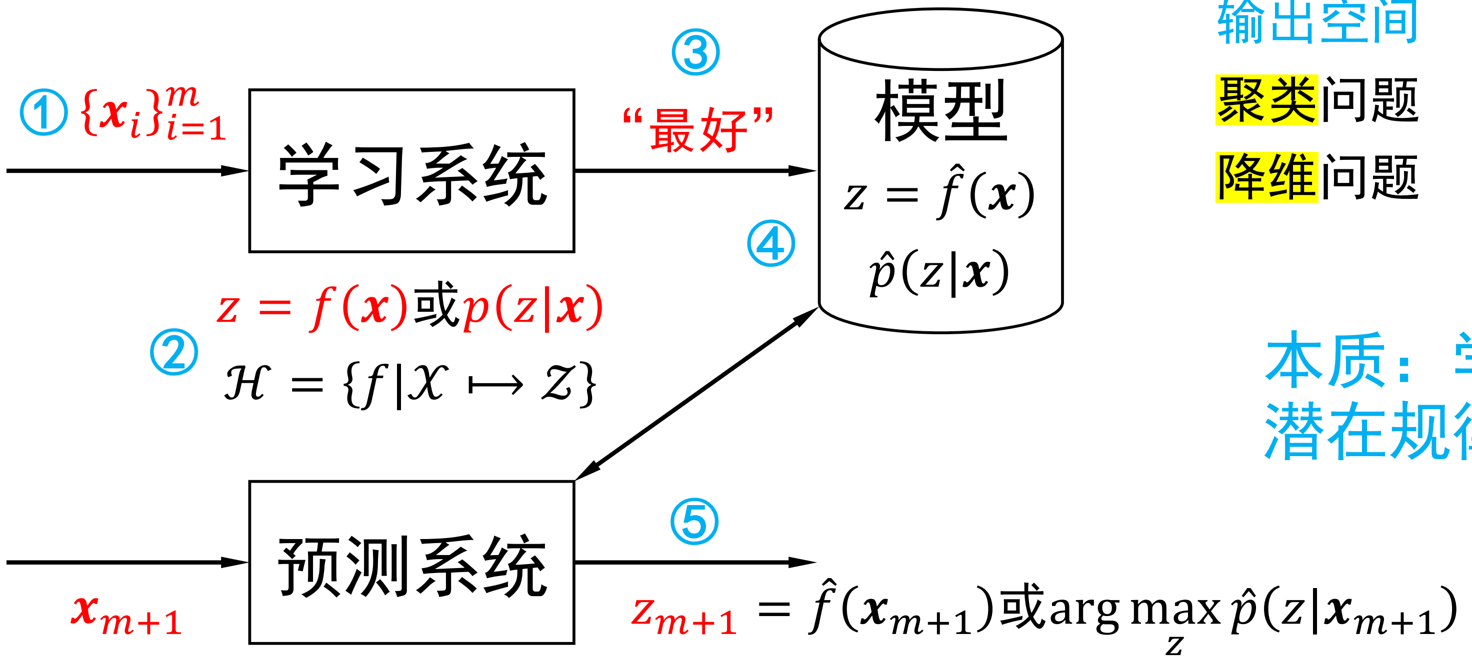
示/实例（instance），特征向量（ feature vector ）$\boldsymbol{x}_i = (x_{i1}; x_{i2}; \cdots; x_{in}) \in \mathcal{X} \subset \mathbb{R}^n$

输出$z_i$表示为对应输入分析所得的类别、转换等$z_i \in \mathcal{Z} \subset \mathbb{R}^?$

输入空间（特征空间）

输出空间

模型实际上都是定义在特征空间上的




聚类问题

降维问题

本质：学习数据中的潜在规律或结构

①数据、②模型、③策略、④算法、⑤应用

# 维数灾难 (curse of dimensionality)

- kNN精度依据一个重要的假设：任意测试样本 $x$ 附近的任意小的$\delta$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：

# 维数灾难 (curse of dimensionality)

- kNN精度依据一个重要的假设：任意测试样本$x$附近的任意小的$\delta$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：
  - 当 $\delta = 0.001$ ，仅考虑单个属性，则需1000个样本点平均分布在归一化后的属性取值范围内

# 维数灾难 (curse of dimensionality)

- kNN精度依据一个重要的假设：任意测试样本$x$附近的任意小的$\delta$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：
  - 当$\delta$=0.001，仅考虑单个属性，则需1000个样本点平均分布在归一化后的属性取值范围内
  - 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。

# 维数灾难 (curse of dimensionality)

- kNN精度依据一个重要的假设：任意测试样本$\boldsymbol{x}$ 附近的任意小的$\delta$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：
  - 当$\delta$ =0.001，仅考虑单个属性，则需1000个样本点平均分布在归一化后的属性取值范围内
  - 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。
  - 高维空间会给距离计算带来很大的麻烦，例如当维数很高时甚至连计算内积都不再容易。

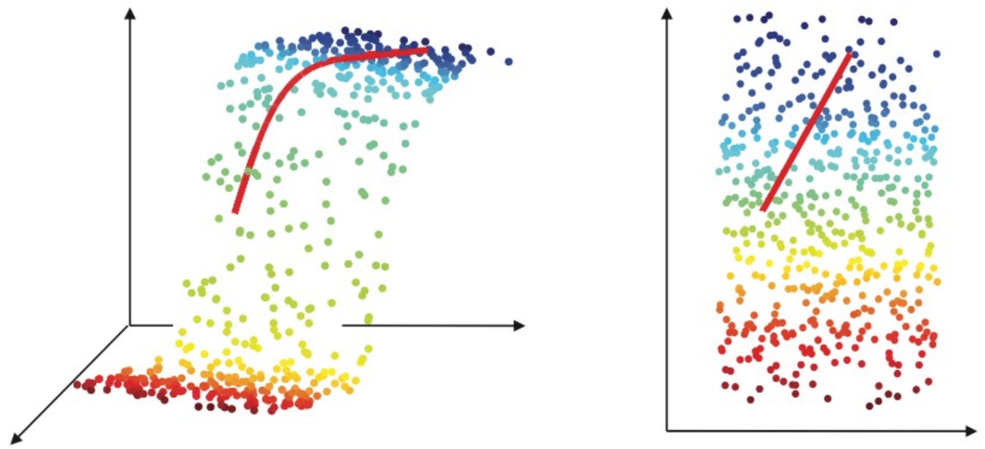
# 维数灾难 (curse of dimensionality)

- kNN精度依据一个重要的假设：任意测试样本$\boldsymbol{x}$ 附近的任意小的 $\delta$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：
  - ➢ 当 $\delta$=0.001，仅考虑单个属性，则需1000个样本点平均分布在归一化后的属性取值范围内
  - ➢ 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。
  - ➢ 高维空间会给距离计算带来很大的麻烦，例如当维数很高时甚至连计算内积都不再容易。
  - ➢ 在高维情形下出现的数据样本稀疏、距离计算困难等问题，是所有机器学习方法共同面临的严重障碍，被称为"维数灾难"。

# 降维

降维（dimension reduction）是缓解维数灾难的一个重要途径

- 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



(a) 三维空间中观察到的样本点　　(b) 二维空间中的曲面

图 10.2　低维嵌入示意图

数据样本虽然是高维的，但与学习任务密切相关的也许仅是某个低维分布，即高维空间中的一个低维"嵌入"（embedding），因而可以对数据进行有效的降维

# 举例

考虑以下三维空间中的六个点

| | | | | | |
|---|---|---|---|---|---|
| 1 | 2 | 4 | 3 | 5 | 6 |
| 2 | 4 | 8 | 6 | 10 | 12 |
| 3 | 6 | 12 | 9 | 15 | 18 |

如果存储每个数字需要一个字节，我们需要18=3*6个字节

# 举例

这6个点在空间上有一定的关联性：方向相同

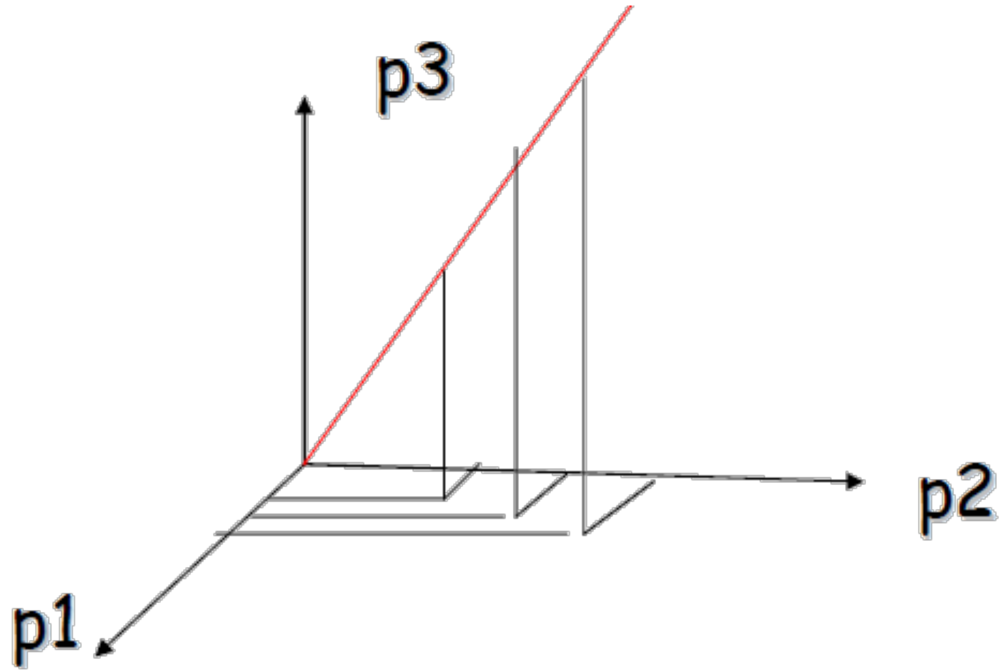
$$\begin{bmatrix}1\\2\\3\end{bmatrix} = 1 \times \begin{bmatrix}1\\2\\3\end{bmatrix} \qquad \begin{bmatrix}4\\8\\12\end{bmatrix} = 4 \times \begin{bmatrix}1\\2\\3\end{bmatrix} \qquad \begin{bmatrix}5\\10\\15\end{bmatrix} = 5 \times \begin{bmatrix}1\\2\\3\end{bmatrix}$$

$$\begin{bmatrix}2\\4\\6\end{bmatrix} = 2 \times \begin{bmatrix}1\\2\\3\end{bmatrix} \qquad \begin{bmatrix}3\\6\\9\end{bmatrix} = 3 \times \begin{bmatrix}1\\2\\3\end{bmatrix} \qquad \begin{bmatrix}6\\12\\18\end{bmatrix} = 6 \times \begin{bmatrix}1\\2\\3\end{bmatrix}$$

存储这6个点，我们实际只需要9=3+6个字节

# 举例

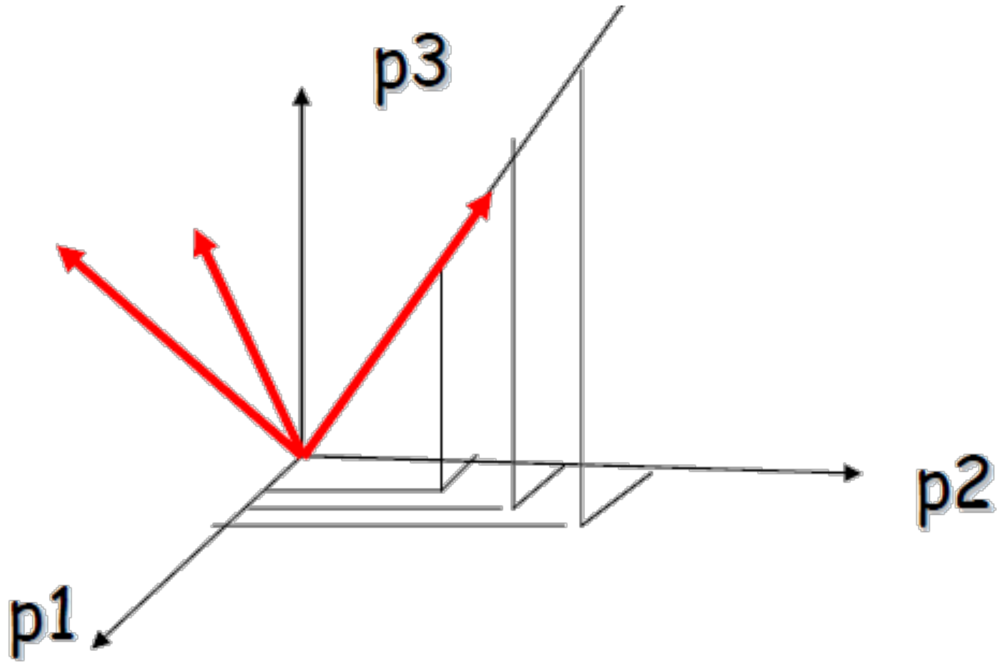
在这个例子中，6个点正好在同一条直线上



# 举例

我们可以在一个新坐标系中表示这6个点

新坐标系的第一个坐标方向是这6个点的方向




在新坐标系中，每个点只有一个非零坐标

我们只需要存储这个坐标方向和每个点的非零坐标

# 预备知识：基变换与降维

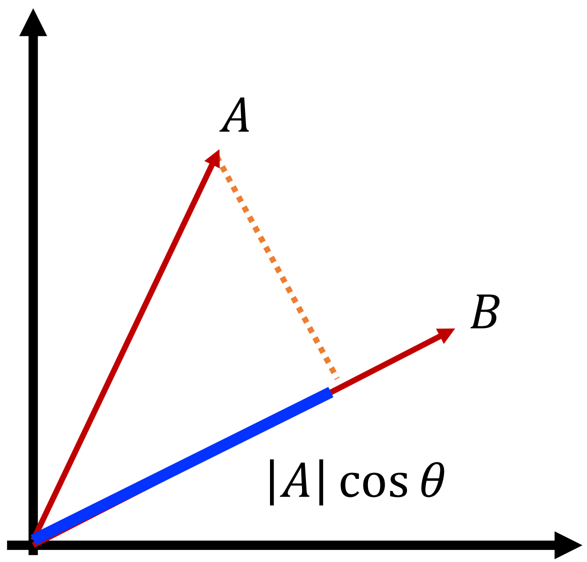
# 内积

- 内积
  - 将两个向量映射为一个实数

$$A = (a_1; a_2; \ldots; a_m)$$
$$B = (b_1; b_2; \ldots; b_m)$$

$$A \cdot B = a_1 b_1 + a_2 b_2 + \cdots + a_m b_m = A^{\mathrm{T}} B$$

  - 几何解释
    - ✓ 向量为n维空间中的一条从原点发射的有向线段
    - ✓ 假设为二维



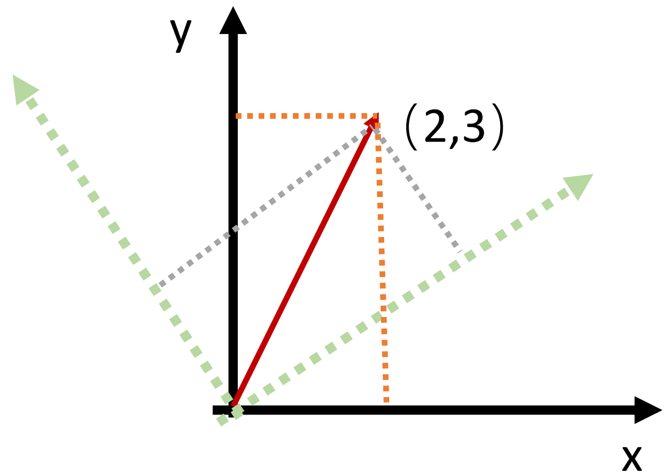

$$A \cdot B = |A||B| \cos\theta$$

如果向量B的模为1，则内积为A向B所在的直线投影的矢量长度

# 基

- 基：
  - 一个二维向量可以对应二维笛卡尔直角坐标系中从原点出发的一个有向线段
  - 向量的表示$(2,3)^{\mathrm{T}}$
  - 线性组合表示

    $$x(1,0)^{\mathrm{T}}+y(0,1)^{\mathrm{T}}$$
  - (1,0)和(0,1) 为二维空间中的一组基

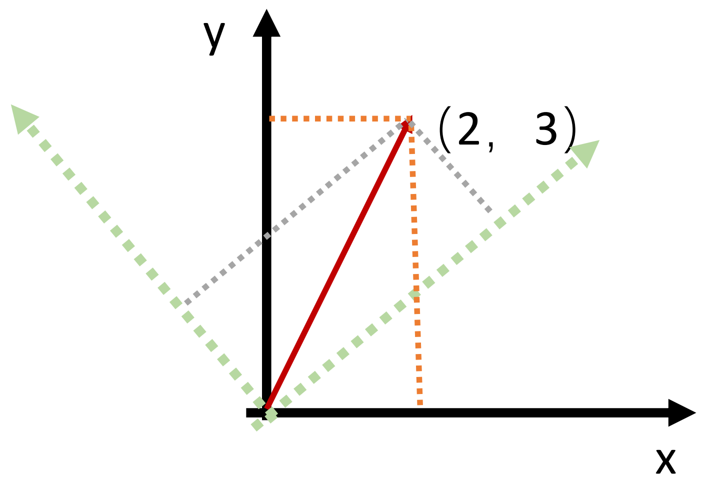



要准确描述一个向量，首先要确定一组基，然后对基所在的直线投影

一般情况下以(1,0)和(0,1)为基，一般基的模为1

# 基变换与坐标变换

- 新基 $(1,1)$和$(-1,1)$
  - $\left(\frac{1}{\sqrt{2}},\frac{1}{\sqrt{2}}\right)$ 和 $\left(-\frac{1}{\sqrt{2}},\frac{1}{\sqrt{2}}\right)$



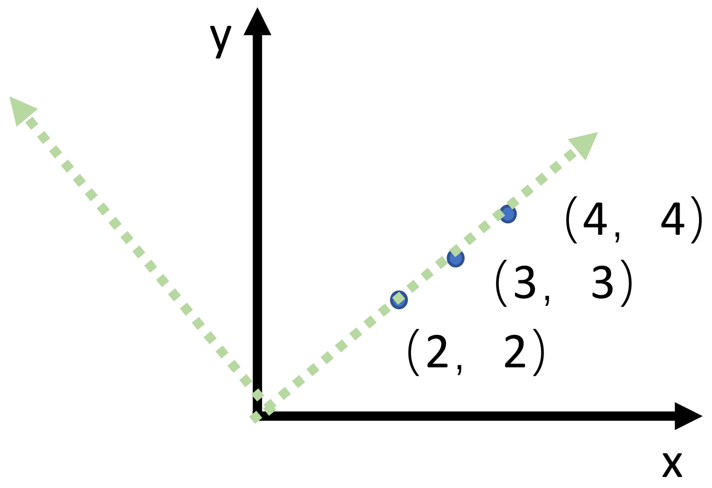

- 怎么求新基下的坐标?
  - $\overrightarrow{(2,3)}\cdot\overrightarrow{\left(\frac{1}{\sqrt{2}},\frac{1}{\sqrt{2}}\right)}=\frac{5}{\sqrt{2}}$
  - $\overrightarrow{(2,3)}\cdot\overrightarrow{\left(-\frac{1}{\sqrt{2}},\frac{1}{\sqrt{2}}\right)}=\frac{1}{\sqrt{2}}$

$$\begin{pmatrix}\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\end{pmatrix}\begin{pmatrix}2\\3\end{pmatrix}=\begin{pmatrix}\frac{5}{\sqrt{2}}\\ \frac{1}{\sqrt{2}}\end{pmatrix}$$

过度矩阵

# 由坐标变换到降维

- 新基 $(1,1)$和$(-1,1)$
  - ➢ $\left(\frac{1}{\sqrt{2}},\frac{1}{\sqrt{2}}\right)$ 和 $\left(-\frac{1}{\sqrt{2}},\frac{1}{\sqrt{2}}\right)$




$$\begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}\begin{pmatrix} 2 \\ 2 \end{pmatrix} = \boxed{\phantom{00}}$$

$$\begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}\begin{pmatrix} 3 \\ 3 \end{pmatrix} = \boxed{\phantom{00}}$$

$$\begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}\begin{pmatrix} 4 \\ 4 \end{pmatrix} = \boxed{\phantom{00}}$$

$$\begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}\begin{pmatrix} 2 & 3 & 4 \\ 2 & 3 & 4 \end{pmatrix} =$$

$$\boxed{\begin{pmatrix} \frac{4}{\sqrt{2}} & \frac{6}{\sqrt{2}} & \frac{8}{\sqrt{2}} \\ 0 & 0 & 0 \end{pmatrix}}$$

降维

# 数据预处理

- 训练数据集

$$\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2) \ldots, (\mathbf{x}_m, y_m)\}$$

- 均一化操作
  - 特征 $j$ 均值

$$u_j = \frac{1}{m}\sum_{i=1}^{m} x_i^j$$

  - 对于每一个样本数据$(\mathbf{x}_i, y_i)$, 用 $x_i^j - u_j$来代替$x_i^j$

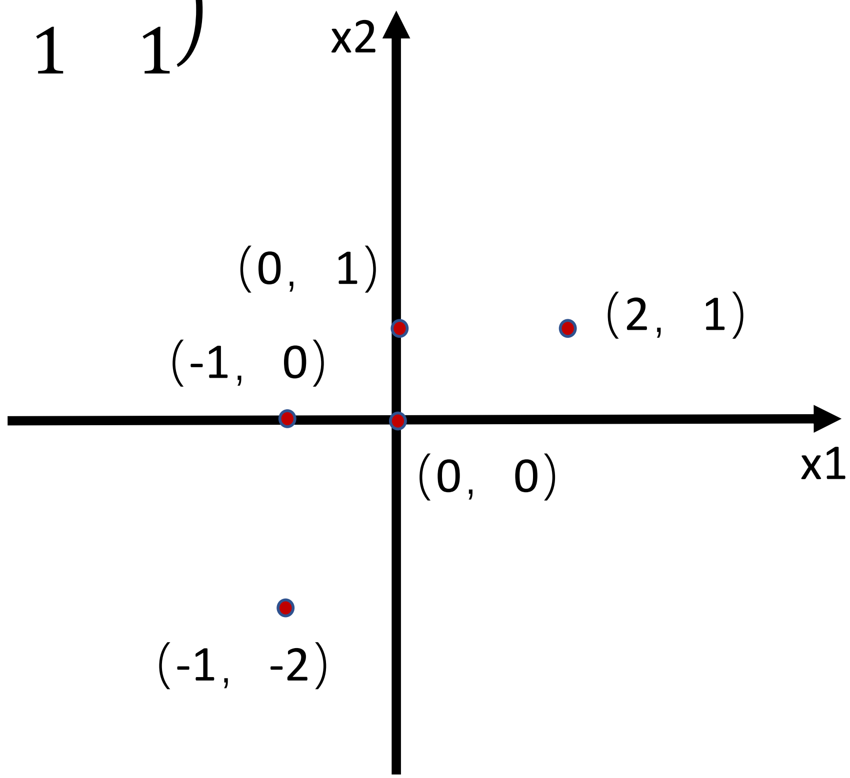
# 例子

5个点 $x_1, x_2$ $\begin{pmatrix} 1 & 1 & 2 & 4 & 2 \\ 1 & 3 & 3 & 4 & 4 \end{pmatrix}$

均值2,3

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

用 $x_i^j - u_j$来代替$x_i^j$



# 主成分问题建模与求解

# 主成分分析

- 主成分分析

由线性相关变量表示的观测数据

⬇ 利用正交变换转换

少数几个由线性无关变量表示的数据

⬇

线性无关的变量称为主成分

# 主成分分析

## ◻ 主成分分析

由线性相关变量表示的观测数据

⇩ 利用正交变换转换

少数几个由线性无关变量表示的数据

⇩

线性无关的变量称为主成分

主成分的个数通常小于原始变量的个数，所以主成分分析属于降维方法

# 主成分分析

- 主成分分析

由线性相关变量表示的观测数据

利用正交变换转换

少数几个由线性无关变量表示的数据

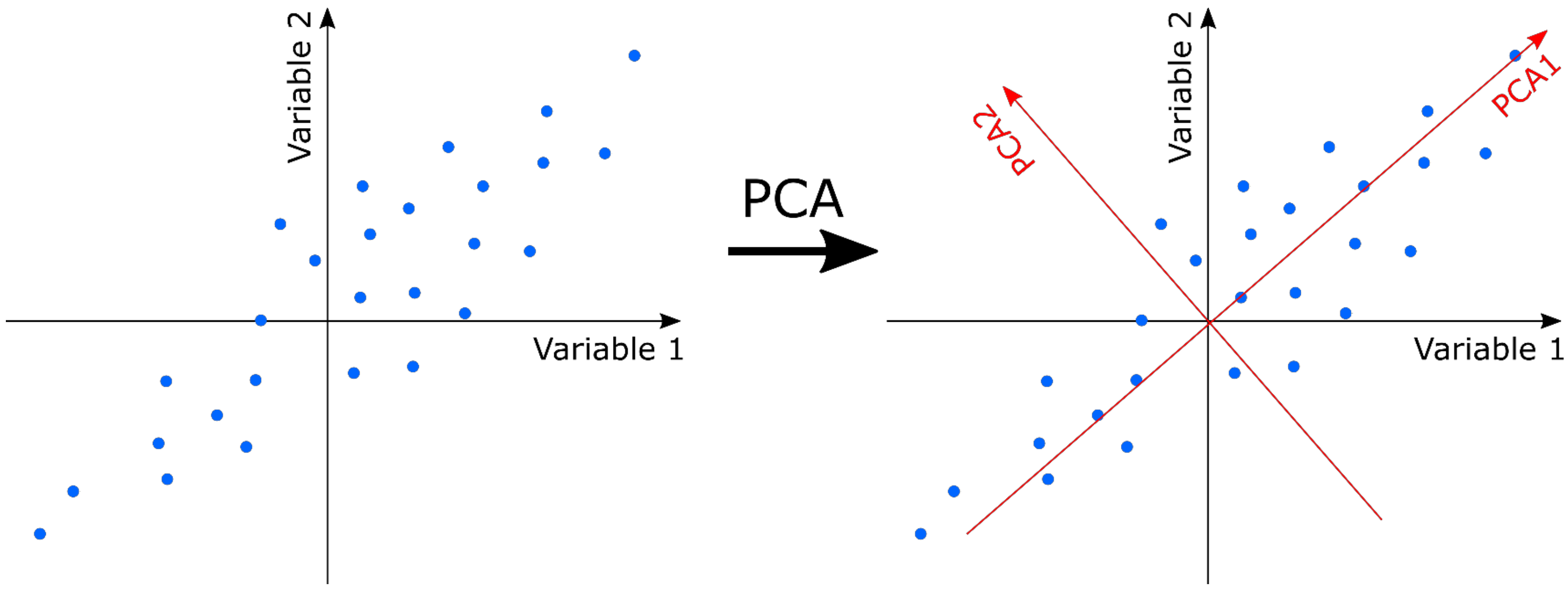
线性无关的变量称为主成分

主成分的个数通常小于原始变量的个数，所以主成分分析属于降维方法

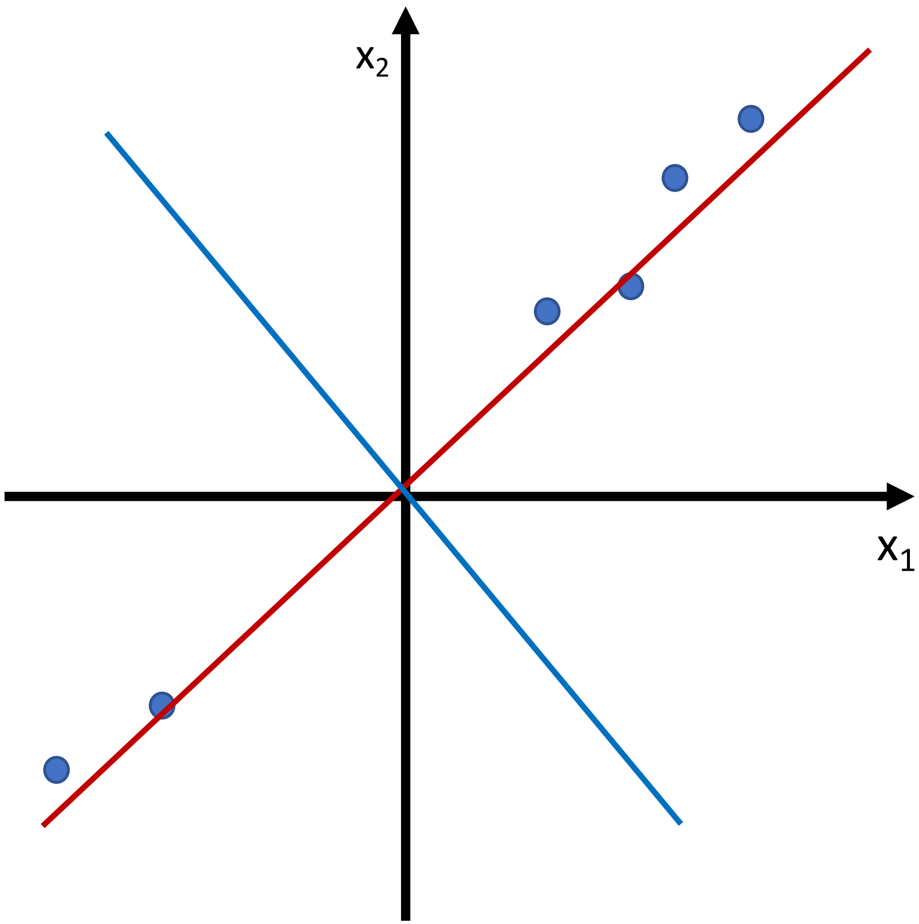
主成分分析主要用于近似地表示原始数据，发现数据的基本结构，也可以把数据由少数主成分表示，对数据降维

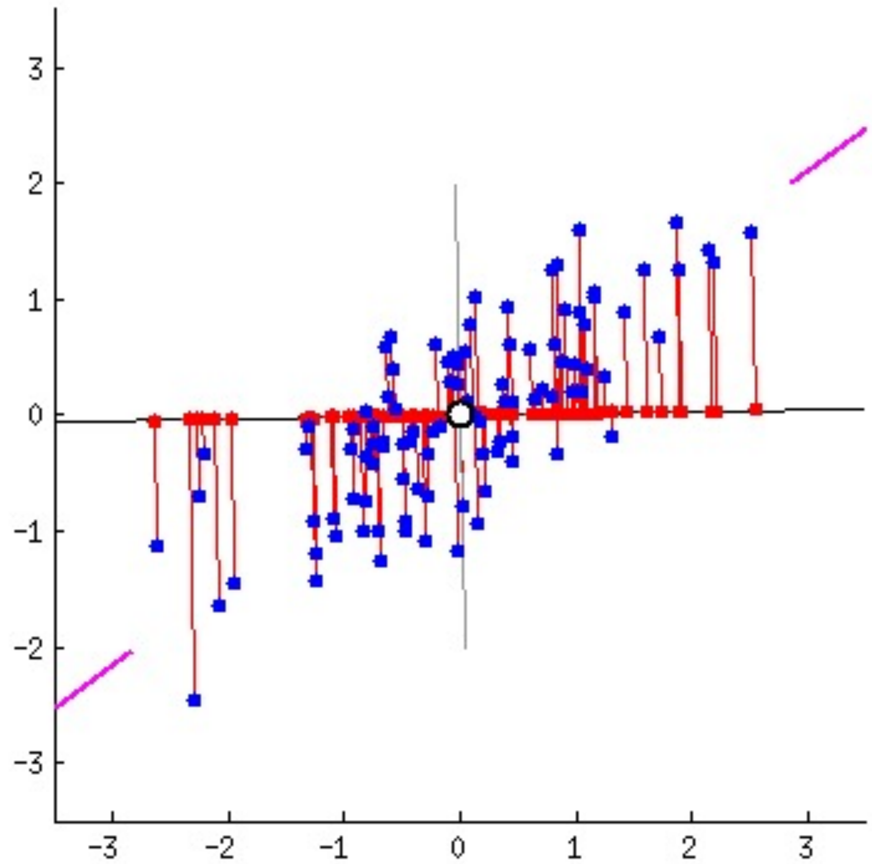
# 举例

给定一组样本点，如何用类似的方法进行降维？



# 主成分分析



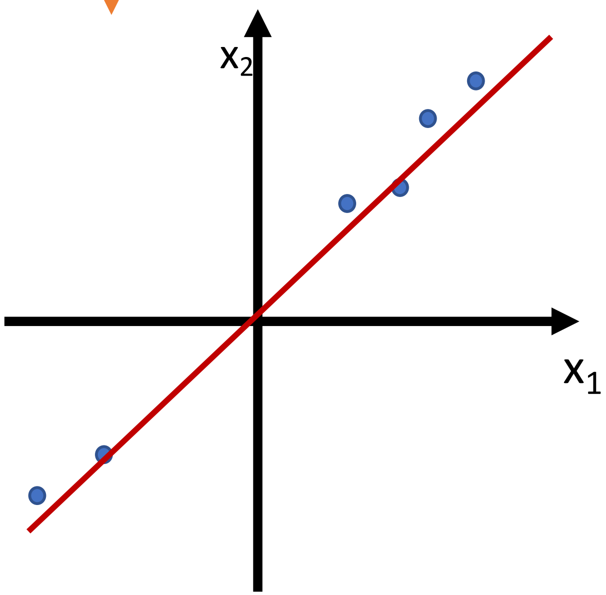



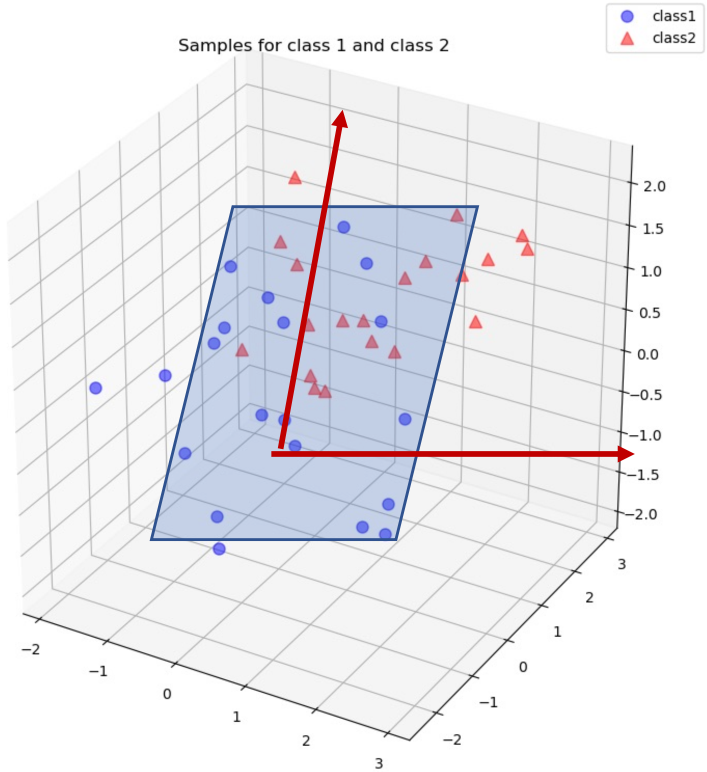

图片摘自互联网

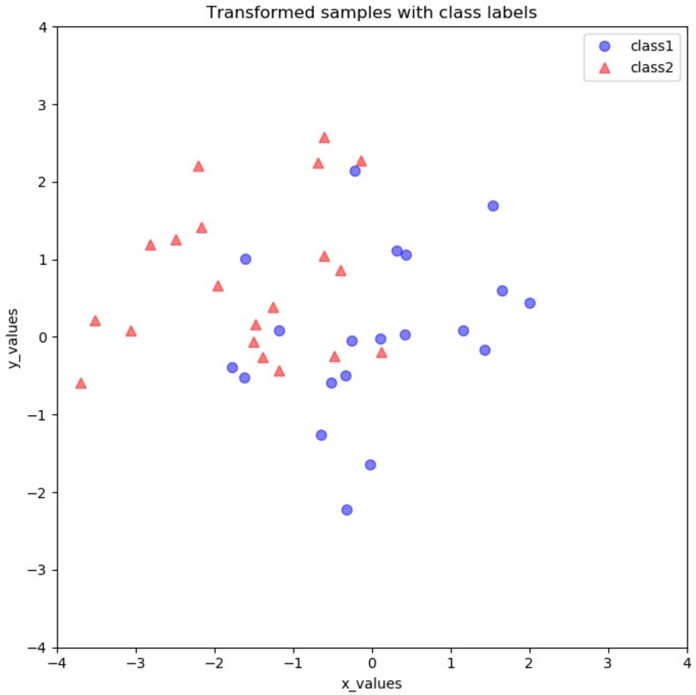
# 主成分分析

从2维到1维: 找到一个向量 $w$ 来最小化 Projection Error

$d$维到$k$维: 找到一个超平面（$k$维）来最小化 Projection Error



# 主成分分析

## ◻ 主成分分析（Principal Component Analysis, PCA）

主成分分析是最常用的一种降维方法.

对于正交属性空间中的样本点，如何用一个超平面（直线的高维推广）对所有样本进行恰当的表达？

# 主成分分析

## ■ 主成分分析（Principal Component Analysis, PCA）

主成分分析是最常用的一种降维方法.

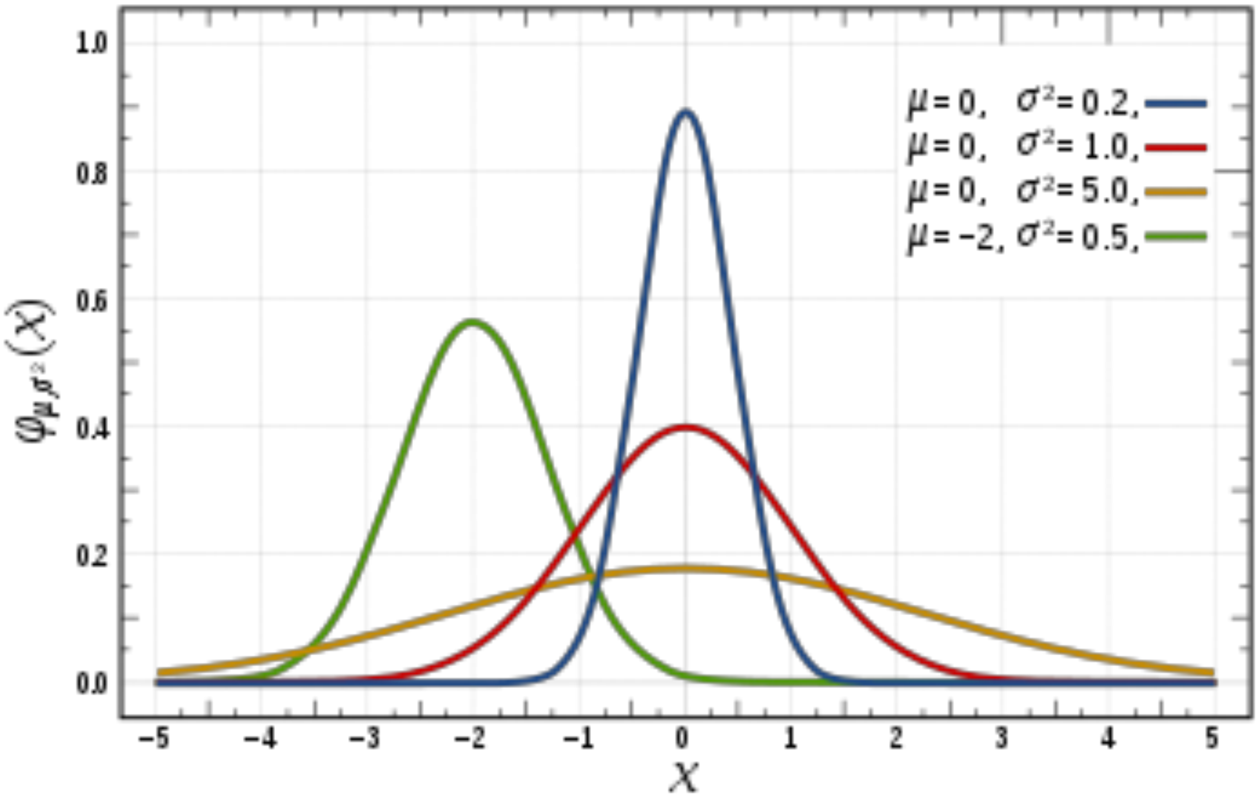
对于正交属性空间中的样本点，如何用一个超平面（直线的高维推广）对所有样本进行恰当的表达？

➢ 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

# 方差

- 方差
  - 一个随机变量的**方差**描述的是它的**离散程度**，也就是该变量离其期望值的距离

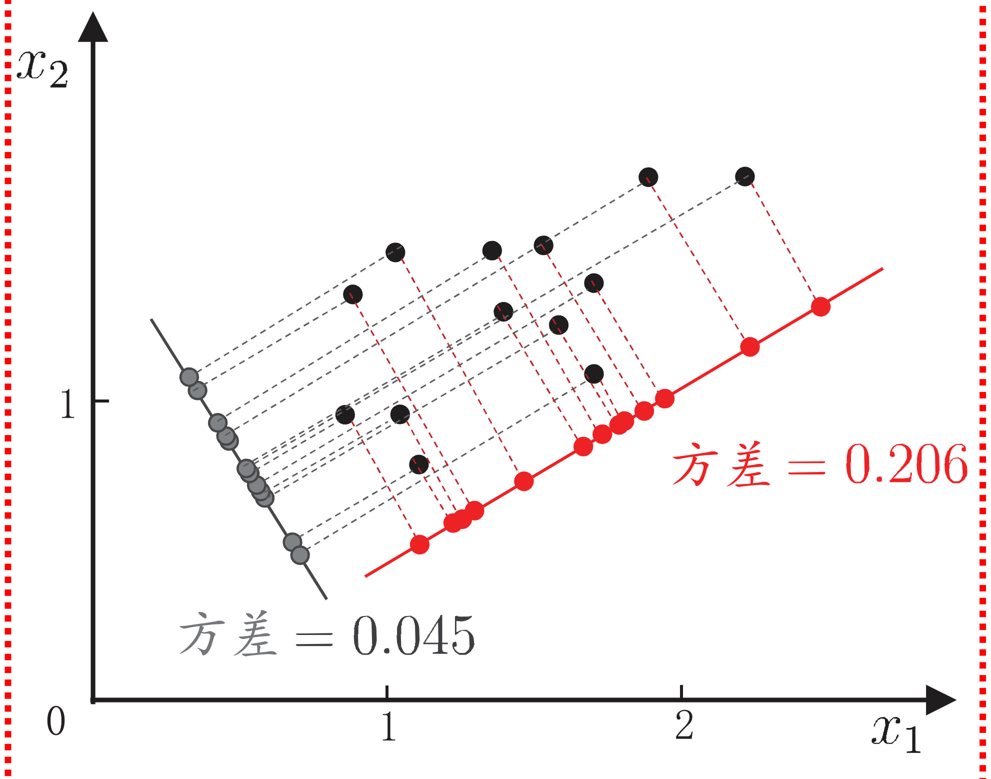



$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

## 主成分分析

超平面，大概应具有这样的性质：

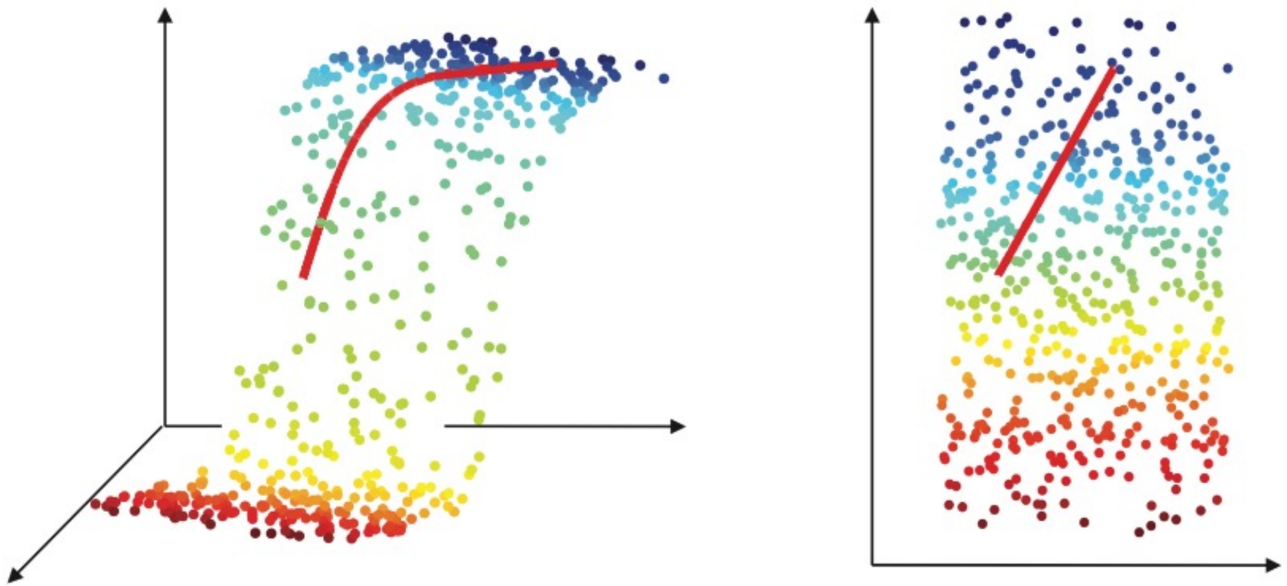
- 最近重构性：样本点到这个超平面的距离都足够近；
- 最大可分性：样本点在这个超平面上的投影尽可能分开



# 主成分分析

## 主成分分析（Principal Component Analysis, PCA）

数据集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（对给定数据进行规范化，使得数据每一变量的平均值为0，方差为1 ）得到的数据分布在原点附近



(a) 三维空间中观察到的样本点　　(b) 二维空间中的曲面

图 10.2　低维嵌入示意图

# 主成分分析

## 主成分分析（Principal Component Analysis, PCA）

数据集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（对给定数据进行规范化，使得数据每一变量的平均值为0，方差为1 ）得到的数据分布在原点附近

1. 假定数据样本进行了中心化

$$\sum_{i=1}^{m} \boldsymbol{x}_i = \mathbf{0}$$

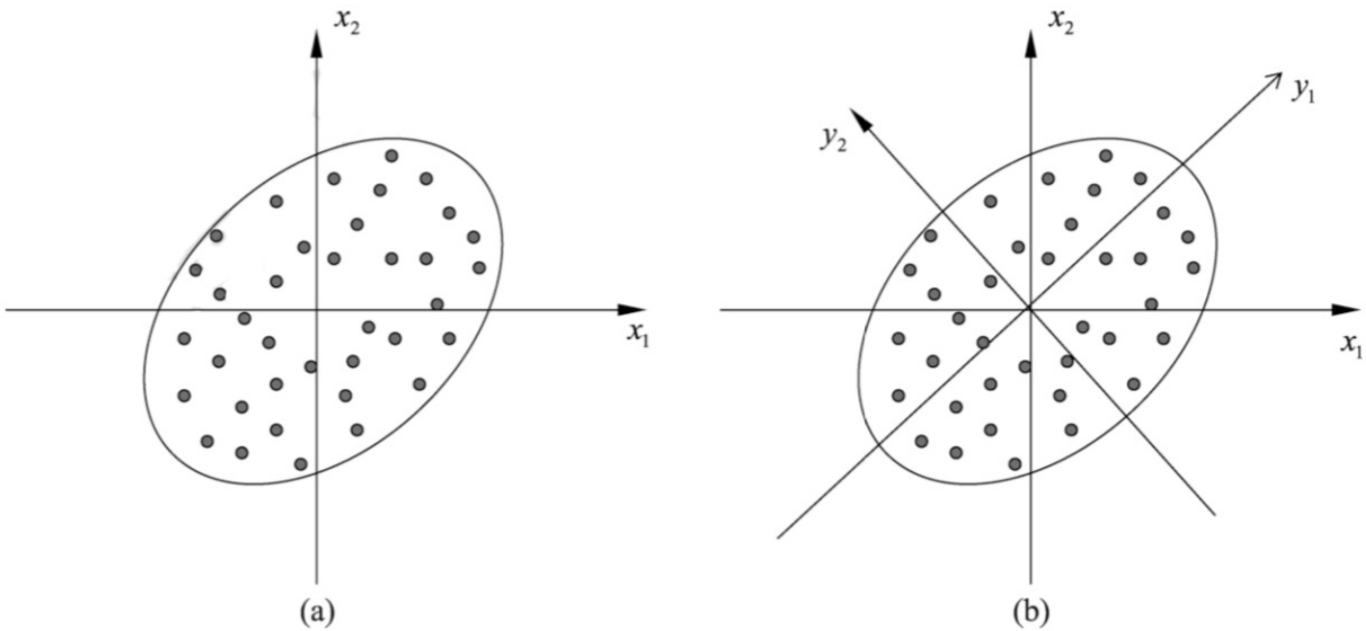
# 主成分分析

## □ 主成分分析

对原坐标系中的数据进行主成分分析等价于进行坐标系旋转变换，将数据投影到新坐标系的坐标轴上；

数据由线性相关的两个变量$x_1$和$x_2$表示，

- 主成分分析对数据进行正交变换，
- 对原坐标系进行旋转变换，
- 并将数据在坐标系表示

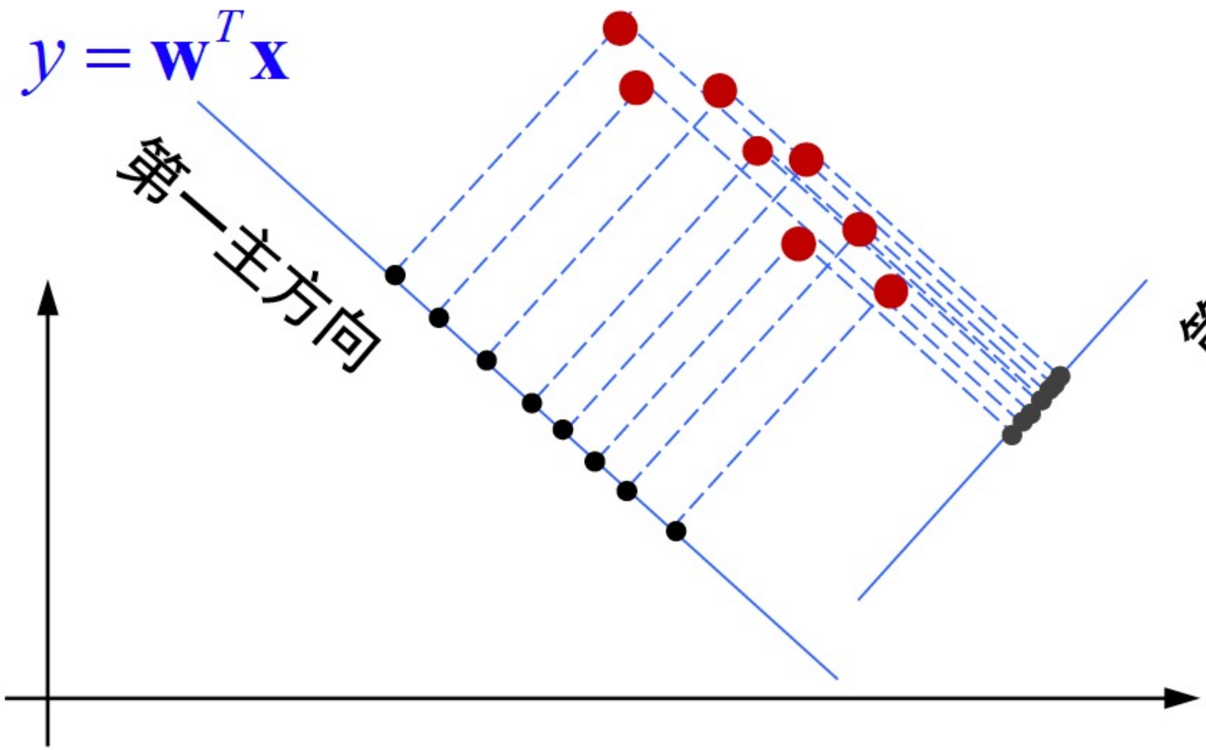


# 主成分分析

## ◘ 主成分分析

对原坐标系中的数据进行主成分分析等价于进行坐标系旋转变换，将数据投影到新坐标系的坐标轴上；

新坐标系的第一坐标轴($y_1$)、第二坐标轴($y_2$)等分别表示第一主成分、第二主成分等



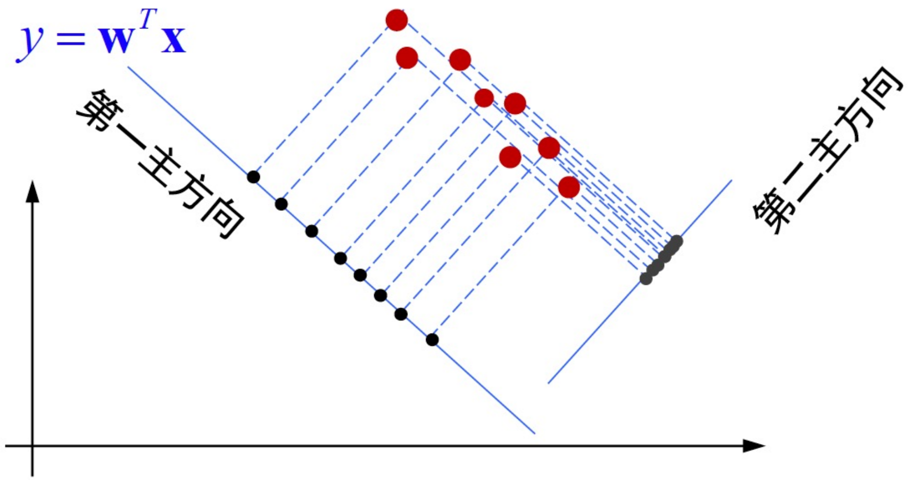
# 主成分分析

## 主成分分析

对原坐标系中的数据进行主成分分析等价于进行坐标系旋转变换，将数据投影到新坐标系的坐标轴上；

新坐标系的第一坐标轴、第二坐标轴等分别表示第一主成分、第二主成分等

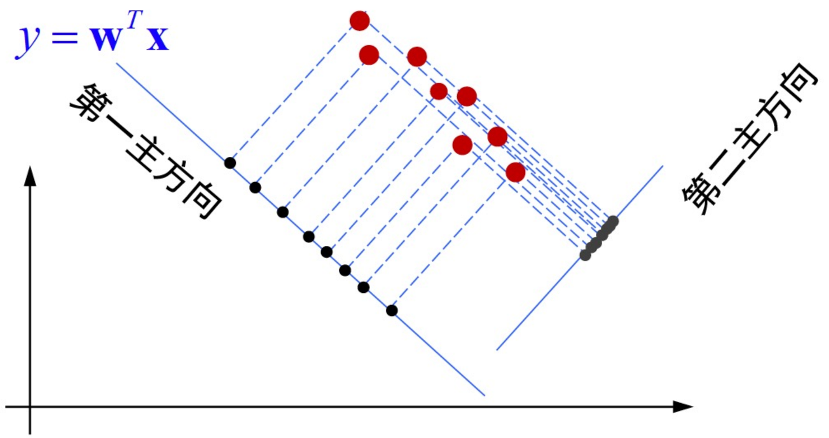
2. 假定投影变换后得到的新坐标系为$\{\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \cdots, \boldsymbol{\omega}_d\}$，$\boldsymbol{\omega}_i$是标准正交基向量，$\|\boldsymbol{\omega}_i\| = 1$，$\boldsymbol{\omega}_i^{\mathrm{T}}\boldsymbol{\omega}_j = 0\ (i \neq j)$



# 主成分分析

## □ 主成分分析

- 假定投影变换后得到的新坐标系为$\{\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \cdots, \boldsymbol{\omega}_d\}$，$\boldsymbol{\omega}_i$是标准正交基向量，$\|\boldsymbol{\omega}_i\| = 1$，$\boldsymbol{\omega}_i^{\mathrm{T}}\boldsymbol{\omega}_j = 0 \ (i \neq j)$
- 若丢弃新坐标系中的部分坐标，即将维度降低到 $d' < d$，则样本点在低维坐标系中的投影是




$$\boldsymbol{z}_i = (z_{i1}; z_{i2}; \cdots; z_{id'})$$

$z_{ij} = \boldsymbol{\omega}_j^{\mathrm{T}}\boldsymbol{x}_i$是$\boldsymbol{x}_i$在低维坐标下第$j$维的坐标

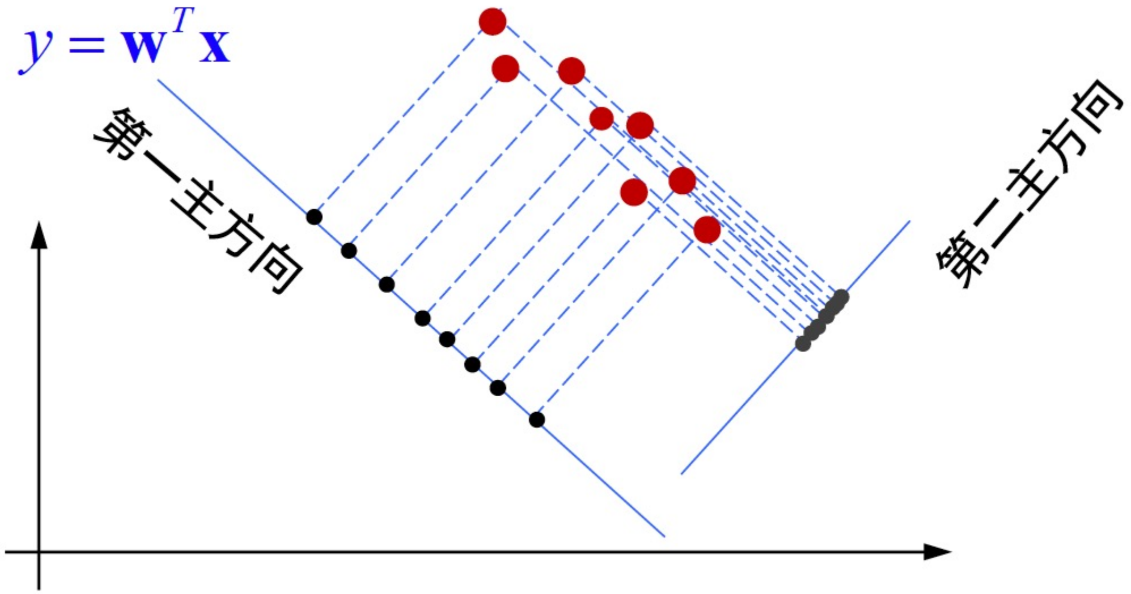
若基于$\boldsymbol{z}_i$来重构$\boldsymbol{x}_j$，则会得到$\hat{\boldsymbol{x}}_i = \sum_{j=1}^{d'} z_{ij}\boldsymbol{\omega}_j$

# 主成分分析

## 主成分分析

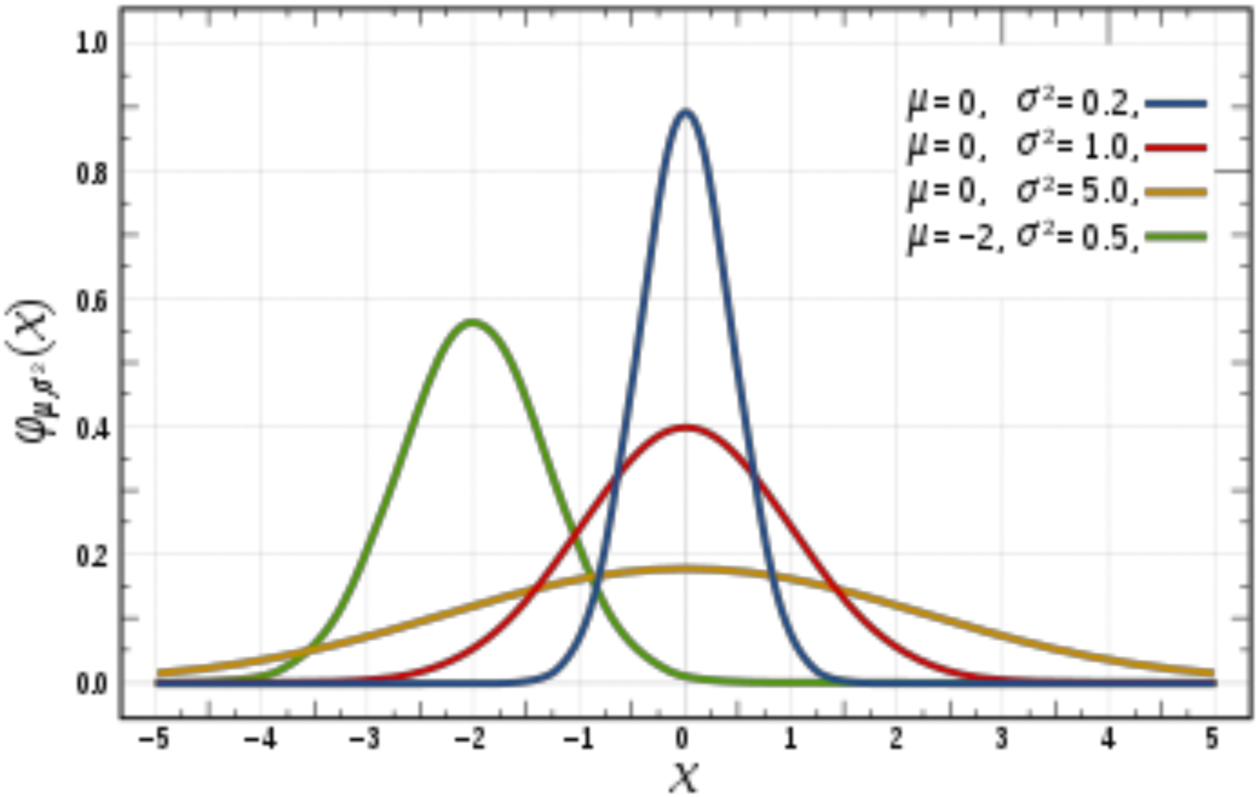
数据在每一轴上的坐标值的平方表示相应变量的方差，
这个坐标系是在所有可能的新的坐标系中，
坐标轴上的方差的和最大的

- 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



# 方差

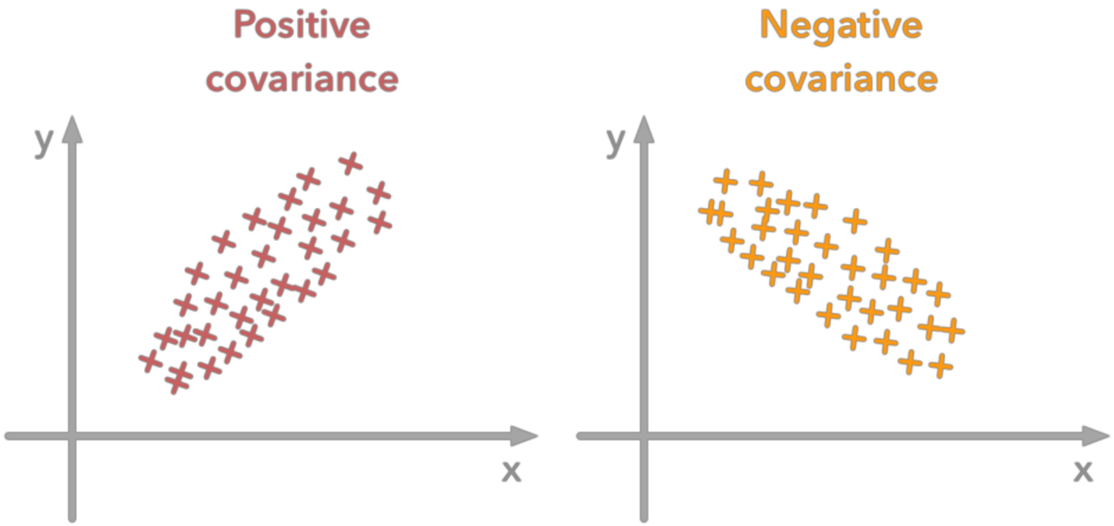
- 方差
  - 一个随机变量的**方差**描述的是它的离散程度，也就是该变量离其期望值的距离




$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

# 协方差

- 协方差
  - **协方差**描述的是不同随机变量之间的关系

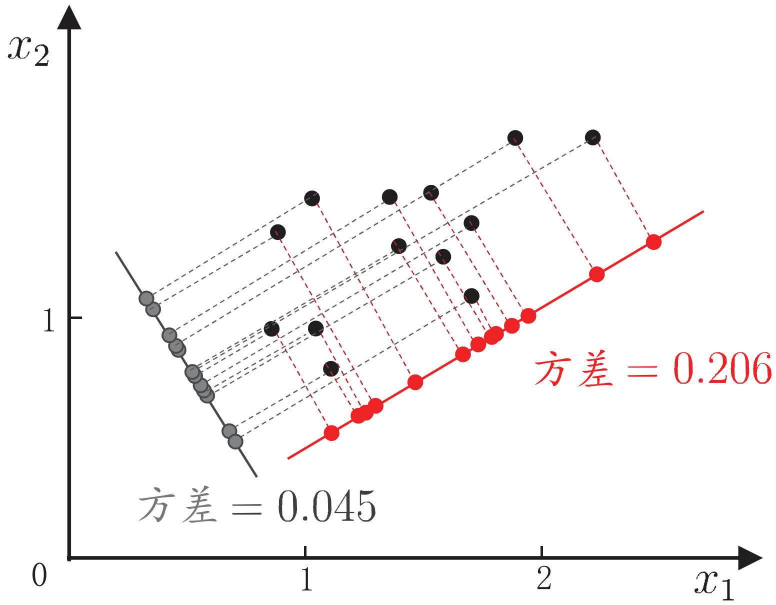



- 协方差矩阵

$$COV = \begin{bmatrix} COV(X,X) & COV(X,Y) & COV(X,Z) \\ COV(Y,X) & COV(Y,Y) & COV(Y,Z) \\ COV(Z,X) & COV(Z,Y) & COV(Z,Z) \end{bmatrix}$$

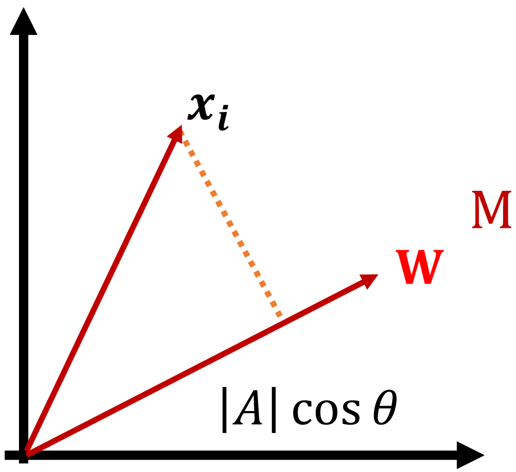
# 最大可分性

- 样本点在这个超平面上的投影能尽可能分开
- 方差：
  - ➢ 每个元素与均值的差的平方和




$$var = \frac{1}{m}\sum_{i=1}^{m}|\boldsymbol{x_i}|^2$$

$$\boldsymbol{x_i} \cdot \mathrm{W} = \boldsymbol{x_i^T}\mathbf{W}$$




$A \cdot B = |A||B|\cos\theta$

$$\mathrm{Max}_{\mathrm{w}}\ var = \frac{1}{m}\sum_{i=1}^{m}\left|\boldsymbol{x_i^T}\mathbf{W}\right|^2$$

$$= \frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{x_i^T}\mathbf{W}\right)^T\boldsymbol{x_i^T}\mathbf{W}$$

$$= \frac{1}{m}\sum_{i=1}^{m}\mathbf{W}^{\boldsymbol{T}}\boldsymbol{x_i}\boldsymbol{x_i^T}\mathbf{W}$$

$$= \frac{1}{m}\mathbf{W}^{\boldsymbol{T}}\sum_{i=1}^{m}\boldsymbol{x_i}\boldsymbol{x_i^T}\mathbf{W}$$

# PCA的求解

$$var = \frac{1}{m}\mathbf{W}^T \sum_{i=1}^{m} \boldsymbol{x}_i \boldsymbol{x}_i^T \mathbf{W} = \mathbf{W}^T (\mathbf{X}\mathbf{X}^T)\mathbf{W}$$

$$\mathrm{Max}_{\mathrm{w}} \ \ \mathrm{tr}(\mathrm{W}^T\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathrm{W}) \quad \mathrm{s.t.}\ \mathrm{W}^T\mathrm{W} = \mathbf{I}$$

$$\Rightarrow \quad \max_{\mathbf{W}} \ \ \mathrm{tr}(\mathbf{W}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{W}) \qquad \mathrm{s.t.} \ \ \mathbf{W}^{\mathrm{T}}\mathbf{W} = \mathbf{I}.$$

- 对优化式使用拉格朗日乘子法可得

$$\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{W} = \lambda\mathbf{W}$$

只需对协方差矩阵$\mathbf{X}\mathbf{X}^{\mathrm{T}}$进行特征值分解，

并将求得的特征值排序：$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ ，

再取前 $d'$ 个特征值对应的特征向量构成 $\mathbf{W} = (\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \cdots, \boldsymbol{\omega}_{d'})$ ，

这就是主成分分析的解

# 主成分分析

**输入:** 样本集 $D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$;
低维空间维数 $d'$.

**过程:**

1: 对所有样本进行中心化: $\boldsymbol{x}_i \leftarrow \boldsymbol{x}_i - \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{x}_i$;
2: 计算样本的协方差矩阵 $\mathbf{X}\mathbf{X}^{\mathrm{T}}$;
3: 对协方差矩阵 $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ 做特征值分解;
4: 取最大的 $d'$ 个特征值所对应的特征向量 $\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_{d'}$.

**输出:** 投影矩阵 $\mathbf{W} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_{d'})$.

图 10.5　PCA 算法

降维后低维空间的维数$d'$通常是由用户事先指定，

还可从重构的角度设置一个重构阈值，例如$t = 95\%$，然后选取使下式成立的最小$d'$值

$$\frac{\sum_{i=1}^{d'} \lambda_i}{\sum_{i=1}^{d} \lambda_i} \geq t$$

# 主成分分析

## 特性

$$\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{W} = \lambda\mathbf{W}$$

- 主成分分析仅需保留$\mathbf{W}$与样本的均值向量，即可通过简单的向量减法和矩阵-向量乘法将新样本投影至低维空间中

# 主成分分析

## □ 特性

$$\mathbf{XX}^{\mathrm{T}}\mathbf{W} = \lambda\mathbf{W}$$

- 主成分分析仅需保留**W**与样本的均值向量，即可通过简单的向量减法和矩阵-向量乘法将新样本投影至低维空间中

- 降维虽然会导致信息的损失，

但一方面，舍弃这些信息后能使得样本的采样密度增大，
另一方面，当数据受到噪声影响时，最小的特征值所对应的特征向量往往与噪声有关（异常点、离散点），舍弃可以起到去噪效果

谢谢！