



北京航空航天大学
COLLEGE OF SOFTWARE 软件学院
BEIHANG UNIVERSITY

人工智能

第9讲：机器学习-无监督学习

K-均值聚类与主成分分析

张晶

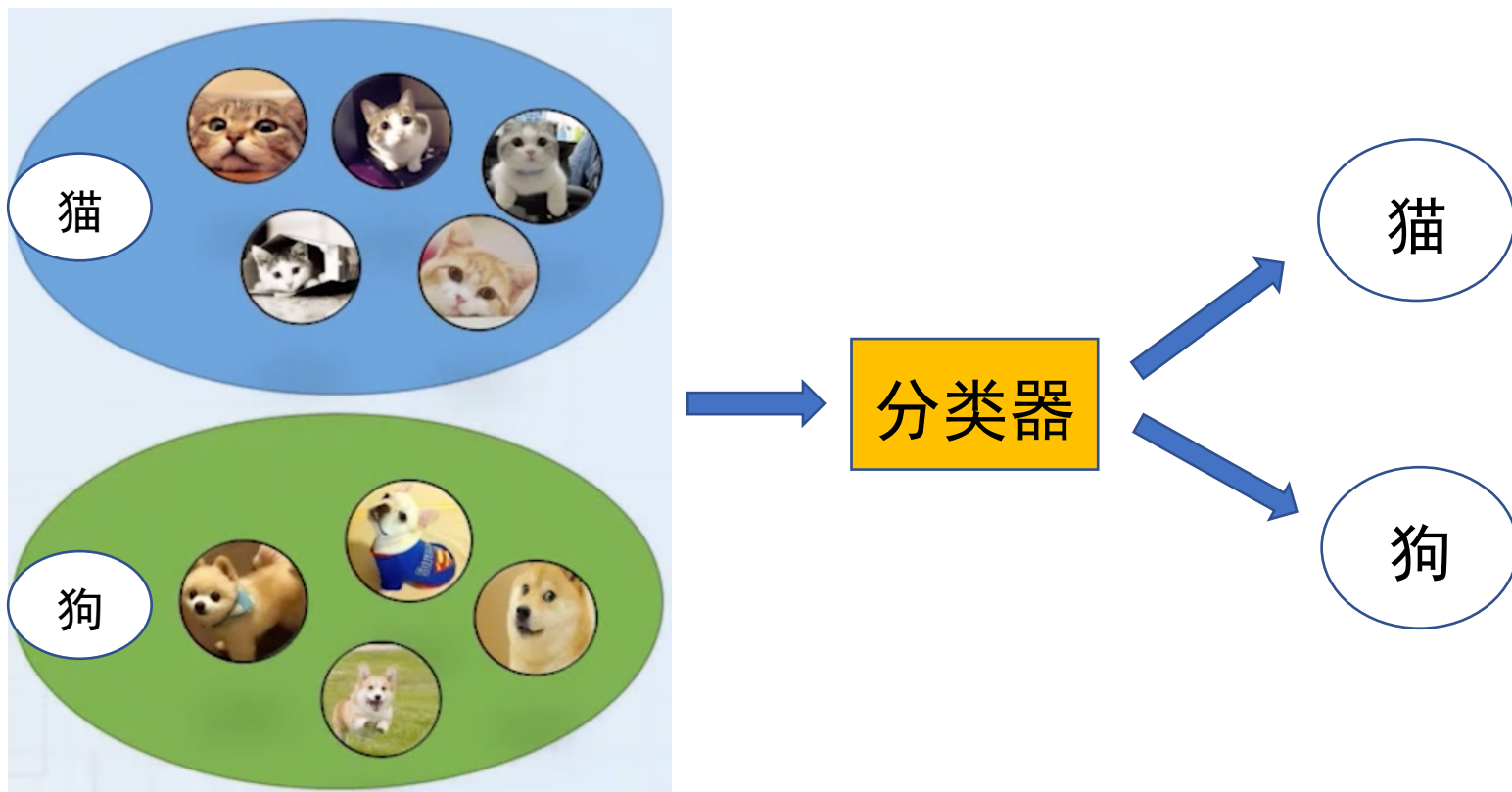
2025年春季

- 参考资料：吴飞，《人工智能导论：模型与算法》，高等教育出版社
- 在线课程：<https://www.icourse163.org/course/ZJU-1003377027?from=searchPage>
- 本部分参考：李宏毅，《机器学习》课程，台湾大学



机器学习的分类（按数据标注情况分类）

- 有监督学习（supervised learning）

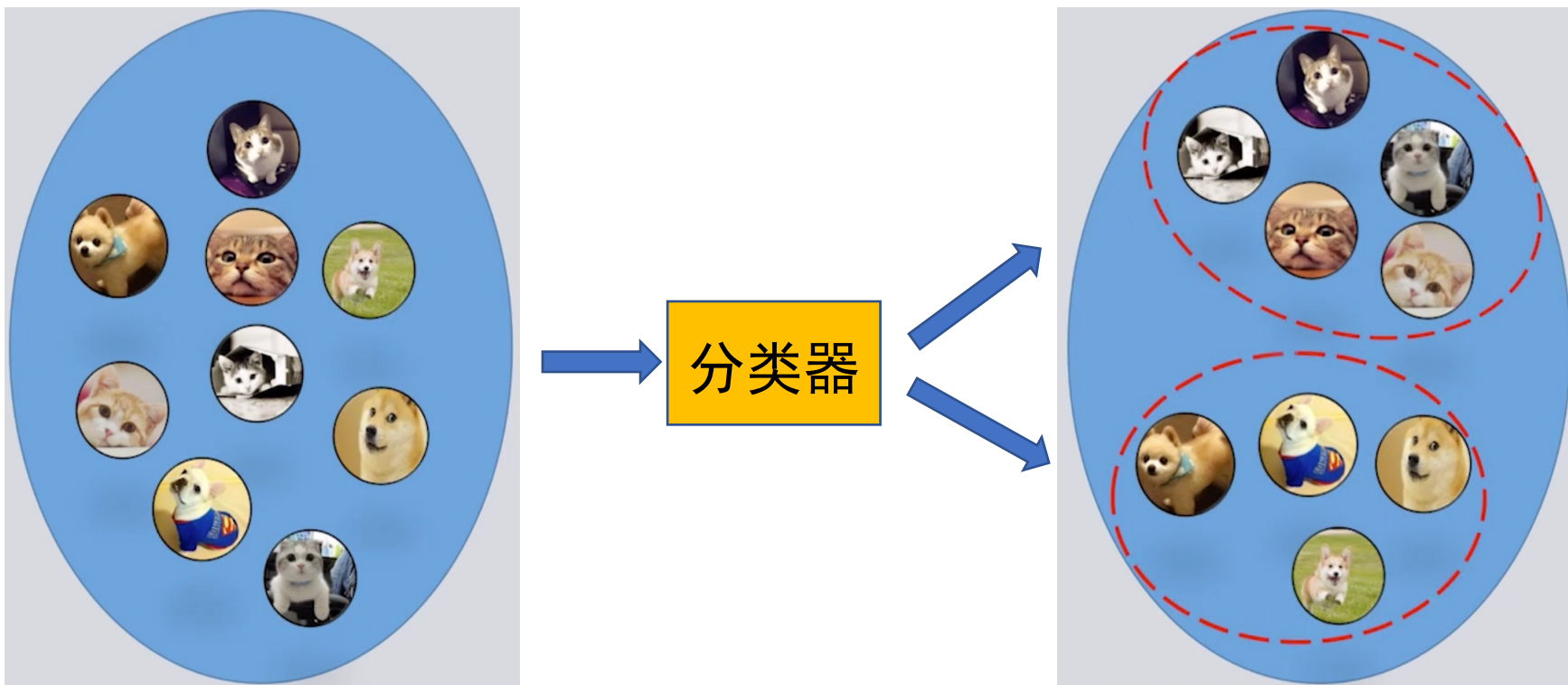




北京航空航天大学
COLLEGE OF SOFTWARE
BEIHANG UNIVERSITY 软件学院

机器学习的分类（按数据标注情况分类）

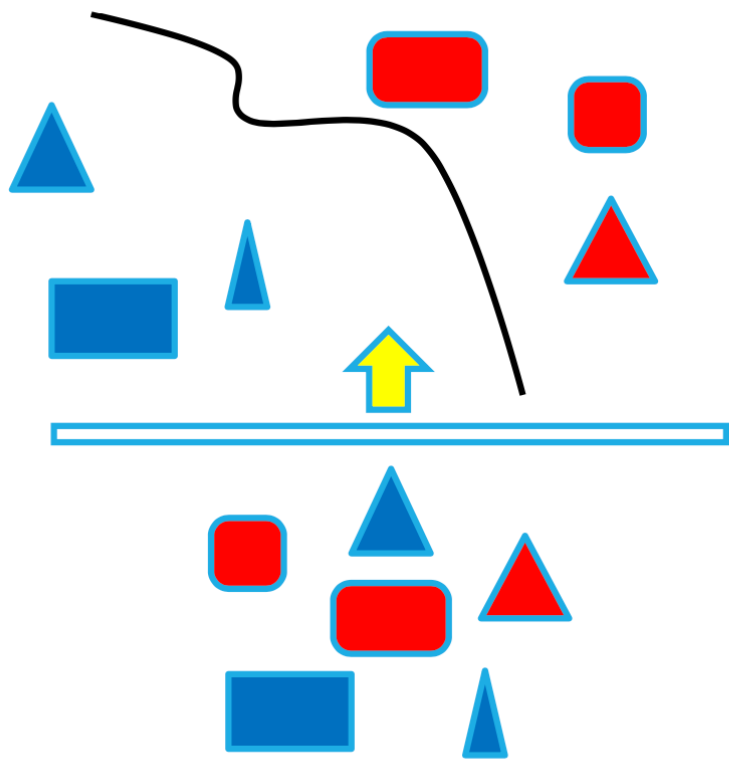
- 无监督学习（unsupervised learning）





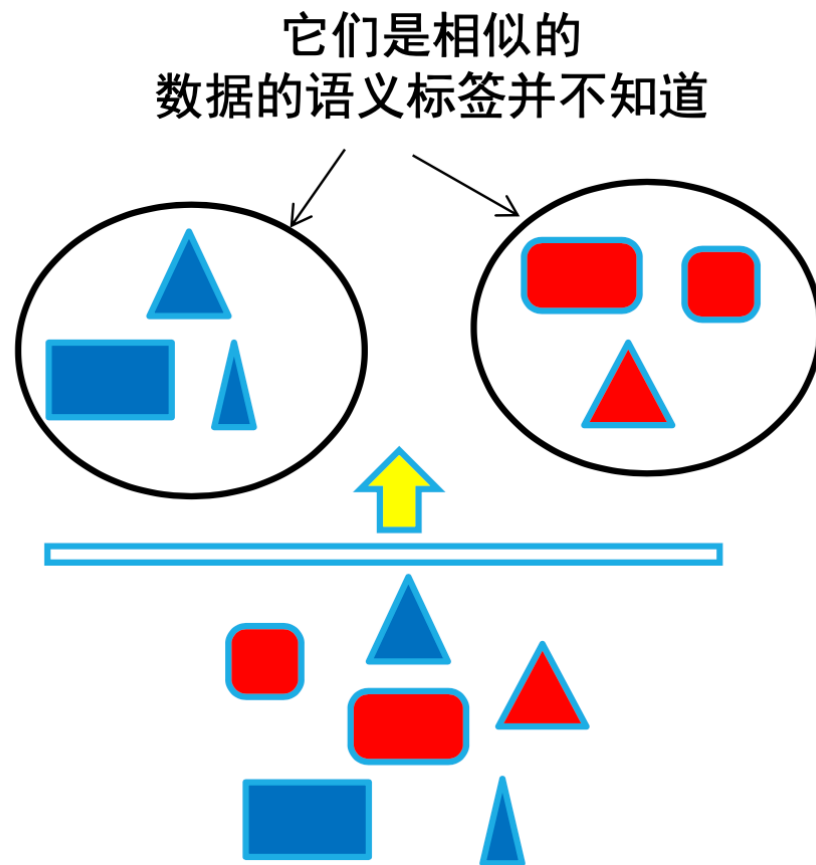
北京航空航天大学
COLLEGE OF SOFTWARE
BEIHANG UNIVERSITY 软件学院

监督学习 v. s. 无监督学习



红色：汽车 蓝色：飞机

左：监督学习



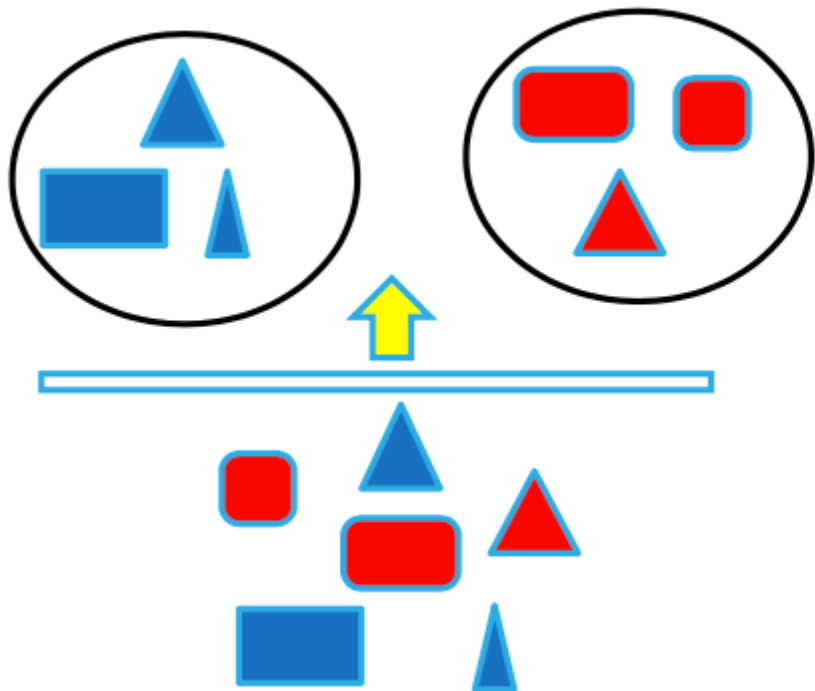
右：无监督学习



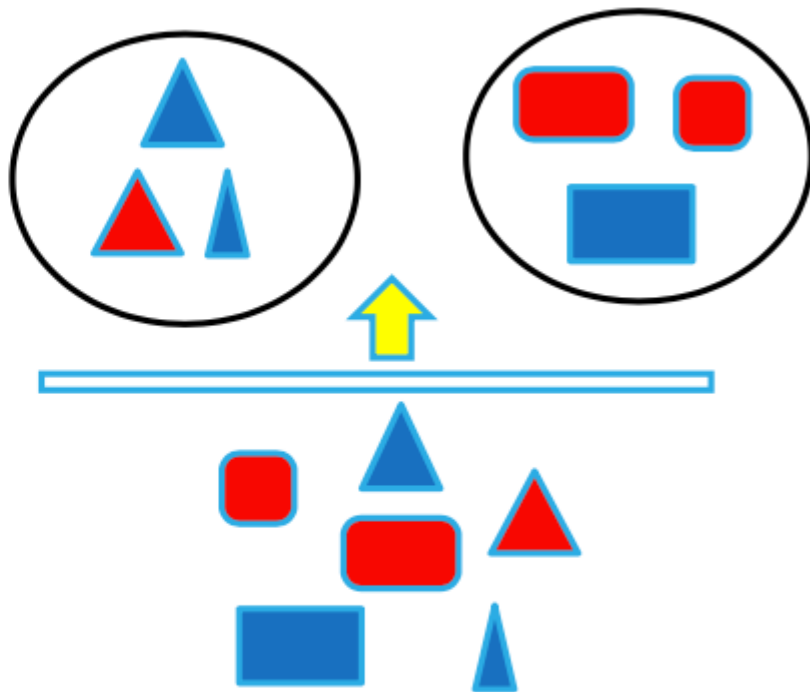
北京航空航天大学
COLLEGE OF SOFTWARE
BEIHANG UNIVERSITY 软件学院

无监督学习：数据特征和相似度函数都很重要

相似度函数：颜色相似



相似度函数：形状相似



无监督学习



北京航空航天大学
COLLEGE OF SOFTWARE
BEIHANG UNIVERSITY 软件学院

提纲

一、K均值聚类

二、主成分分析



K均值聚类 (K-means 聚类)

- 物以类聚，人以群分(《战国策·齐策三》)
- 输入： n 个数据（无任何标注信息）
- 输出： k 个聚类结果
- 目的： 将 n 个数据聚类到 k 个集合（也称为类簇）



北京航空航天大学
COLLEGE OF SOFTWARE
BEIHANG UNIVERSITY 软件学院

K均值聚类 (K-means 聚类)



请思考，想要建模聚类问题，需要考虑哪些基本要素？

作答



K均值聚类算法描述

- 若干定义:

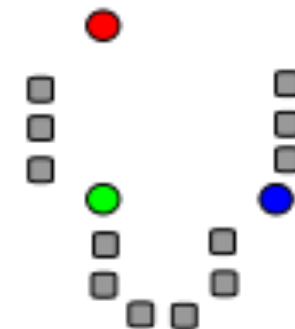
- n 个 m -维数据 $\{x_1, x_2, \dots, x_n\}$, $x_i \in R^m (1 \leq i \leq n)$
- 两个 m 维数据之间的欧氏距离为

$$d(x_i, x_j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \dots + (x_{im} - x_{jm})^2}$$

- $d(x_i, x_j)$ 值越小, 表示 x_i 和 x_j 越相似; 反之越不相似
- 聚类集合数目 k
- 问题: 如何将 n 个数据依据其相似度大小将它们分别聚类到 k 个集合, 使得每个数据仅属于一个聚类集合。



K均值聚类算法：初始化



■ 第一步：初始化聚类质心

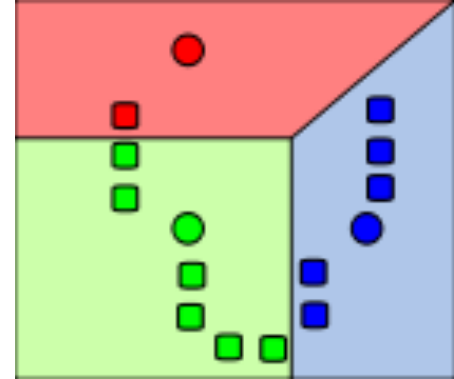
- 初始化 k 个聚类质心

$$C = \{c_1, c_2, \dots, c_k\}, c_j \in R^m (1 \leq j \leq k)$$

- 聚类质心可以从数据 $\{x_1, x_2, \dots, x_n\}$, $x_i \in R^m (1 \leq i \leq n)$ 中采样得到。
- 每个聚类质心 c_j 所在集合记为 G_j 。



K均值聚类算法：对数据进行聚类



■ 第二步：将每个待聚类数据放入唯一一个聚类集合中

- 计算待聚类数据 x_i 和质心 c_j 之间的欧氏距离

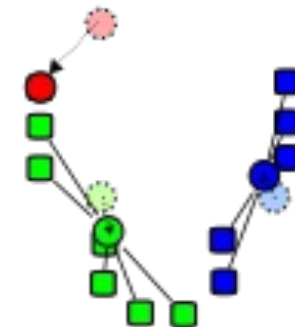
$$d(x_i, c_j) \quad (1 \leq i \leq n, 1 \leq j \leq k)$$

- 将每个 x_i 放入与之距离最近聚类质心所在聚类集合中，即

$$\underset{c_j \in C}{\operatorname{argmin}} d(x_i, c_j)$$



K均值聚类算法：更新聚类质心



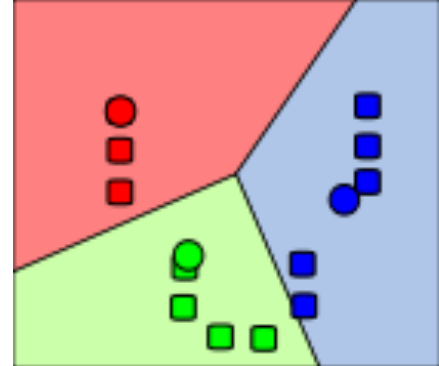
■ 第三步：根据聚类结果、更新聚类质心

- 根据每个聚类集合中所包含的数据，更新该聚类集合质心值，即：

$$c_j = \frac{1}{|G_j|} \sum_{x_i \in G_j} x_i$$



K均值聚类算法：继续迭代



■ 第四步：算法循环迭代，直到满足条件

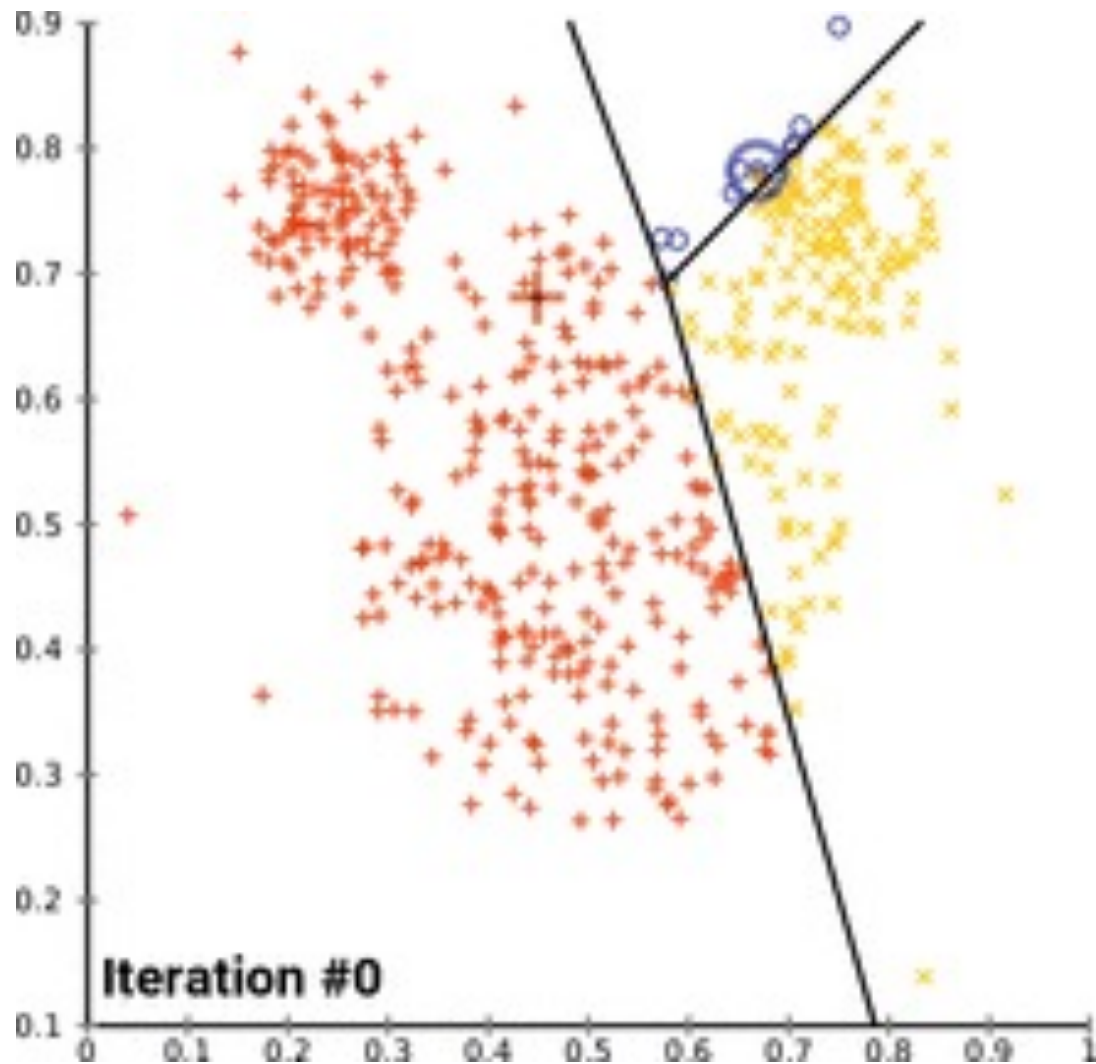
- 在新聚类质心基础上，根据欧氏距离大小，将每个待聚类数据放入唯一一个聚类集合中
- 根据新的聚类结果、更新聚类质心

聚类迭代满足如下任意一个条件，则聚类停止：

- 已经达到了迭代次数上限
- 前后两次迭代中，聚类质心基本保持不变



K均值聚类算：收敛过程





K均值聚类算法的另一个视角：最小化每个类簇的方差

■ 方差：用来计算变量（观察值）与样本平均值之间的差异

$$\arg \min_G \sum_{i=1}^k \sum_{x \in G_i} \|x - G_i\|^2 = \arg \min_G \sum_{i=1}^k |G_i| \text{Var } G_i$$

第*i*个类簇的方差： $\text{var}(G_i) = \frac{1}{|G_i|} \sum_{x \in G_i} \|x - G_i\|^2$

- 欧氏距离与方差量纲相同
- 最小化每个类簇方差将使得最终聚类结果中每个聚类集合中所包含数据呈现出来差异性最小。



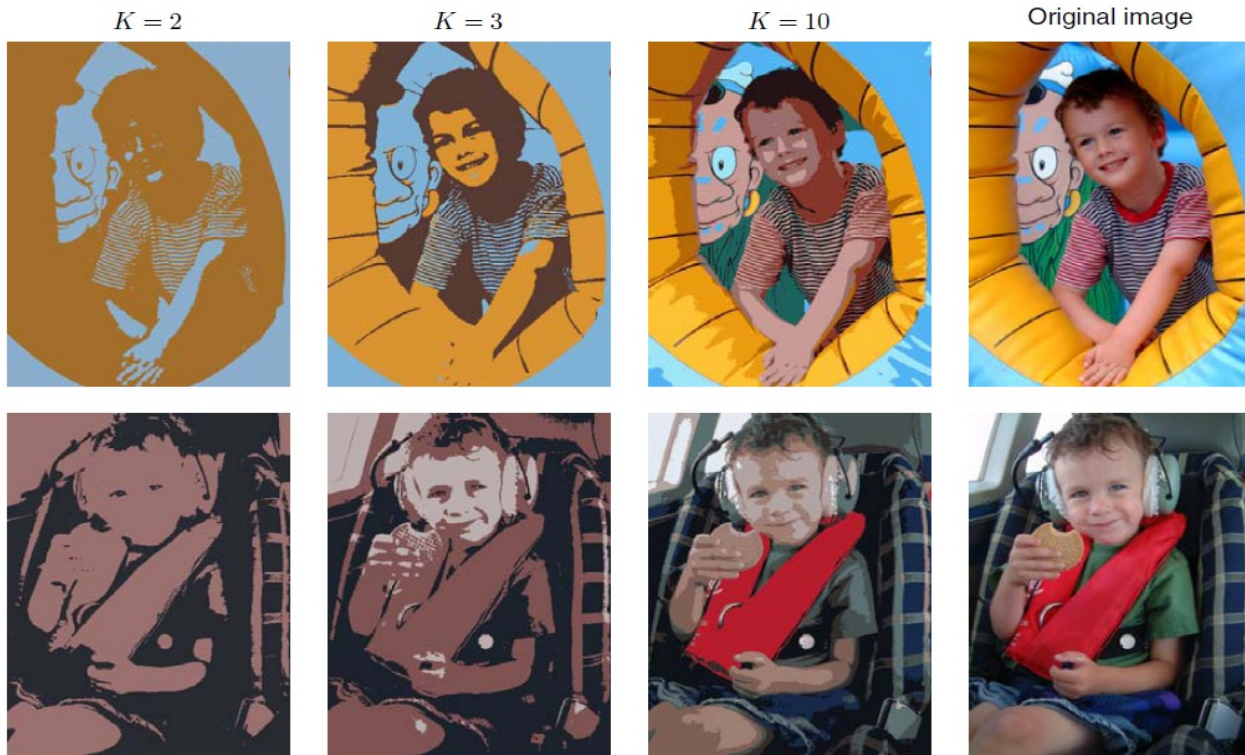
K均值聚类算法的不足

- 需要事先确定**聚类数目**，很多时候我们并不知道数据应被聚类的数目
- 需要**初始化聚类质心**，初始化聚类中心对聚类结果有较大的影响
- 算法是**迭代**执行，时间开销非常大
- **欧氏距离**假设数据每个维度之间的重要性是一样的

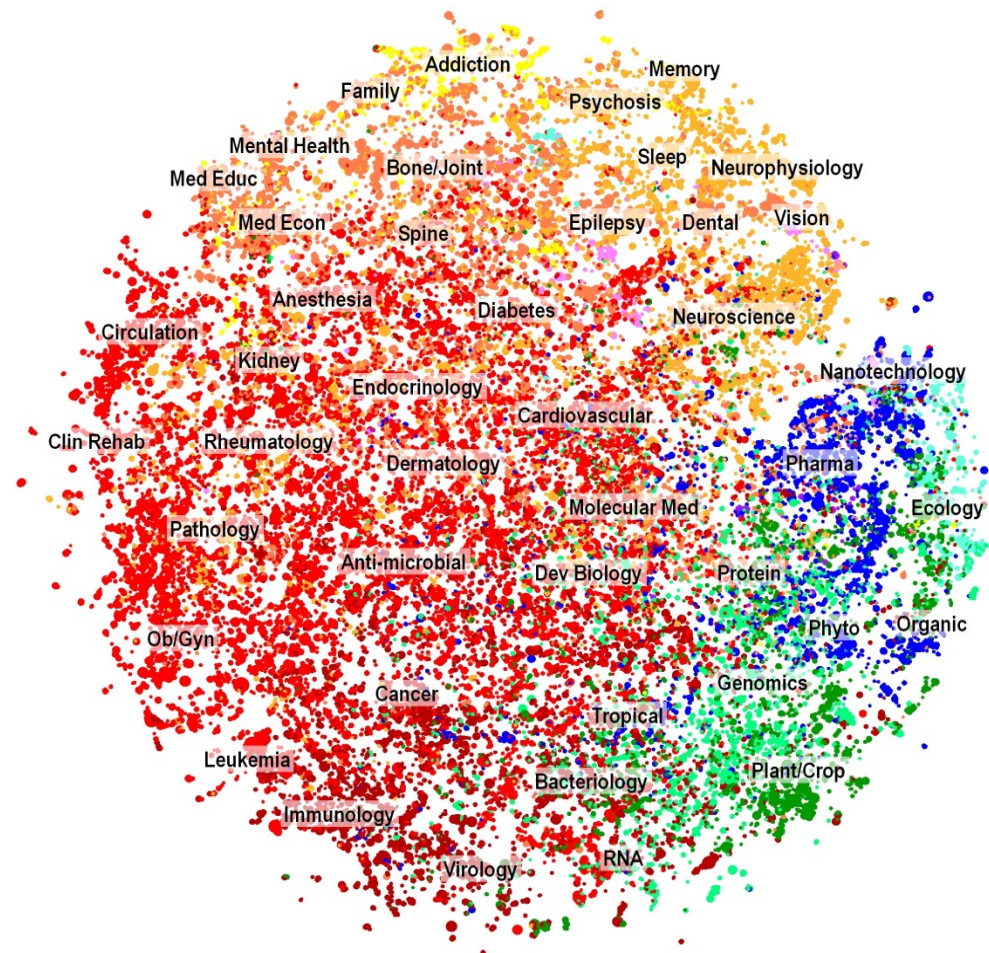


北京航空航天大学
COLLEGE OF SOFTWARE BEIHANG UNIVERSITY
软件学院

K均值聚类算法的应用



图像分割



文本分类：将200多万篇论文聚类到29,000个类别，包括化学、工程、生物、传染疾病、生物信息、脑科学、社会科学、计算机科学等及给出了每个类别中的代表单词



北京航空航天大学
COLLEGE OF SOFTWARE
BEIHANG UNIVERSITY 软件学院

提纲

一、K均值聚类

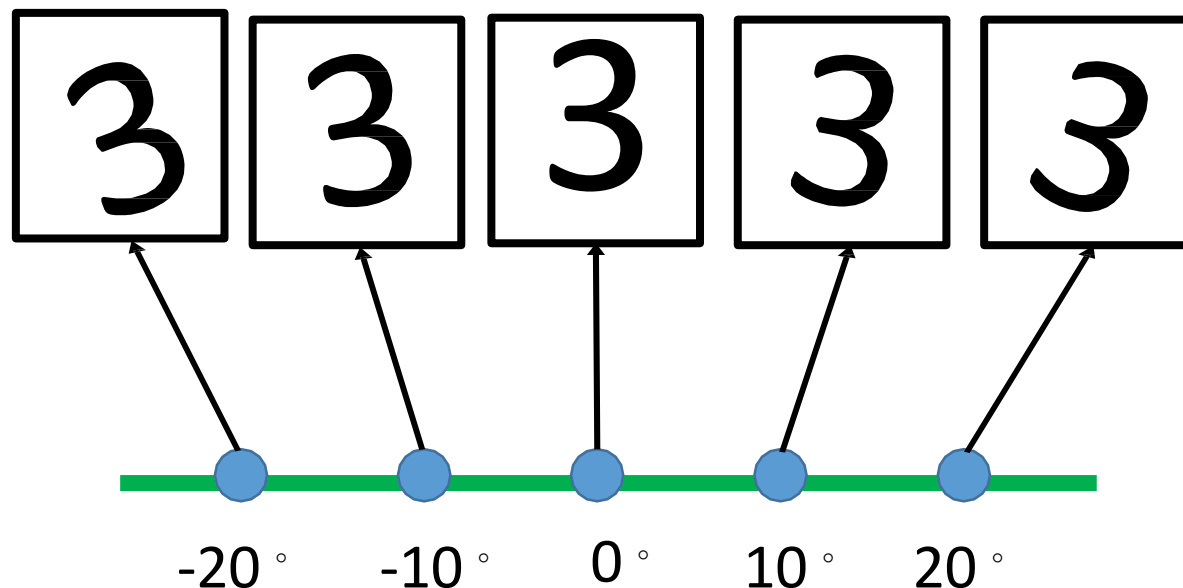
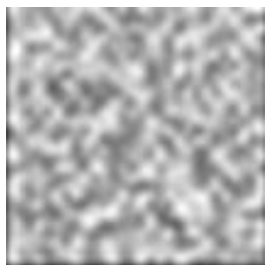
二、主成分分析



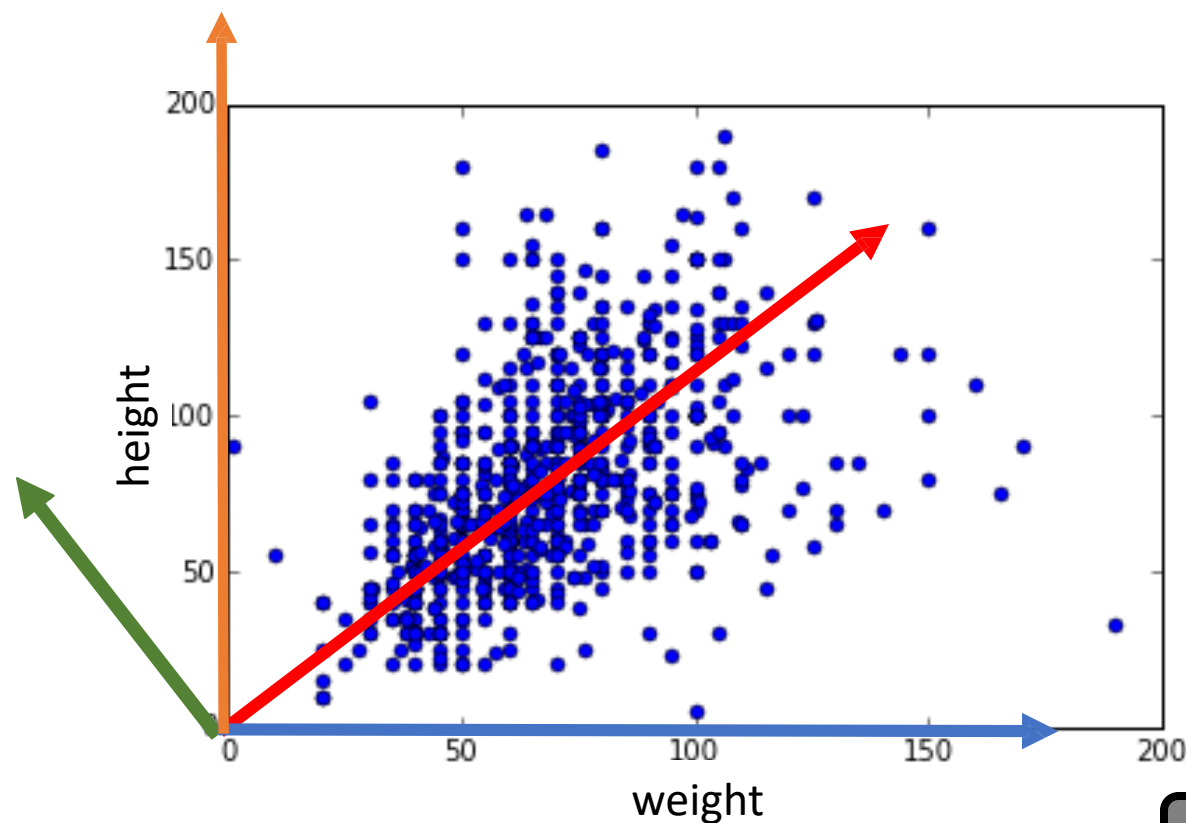
北京航空航天大学
COLLEGE OF SOFTWARE
BEIHANG UNIVERSITY 软件学院

主成分分析: Principle Component Analysis (PCA)

- 主成分分析是一种特征降维方法。人类在认知过程中会主动“化繁为简”
- 奥卡姆剃刀定律 (Occam's Razor): “如无必要, 勿增实体”, 即“简单有效原理”



以下数据为小学生身高体重数据，请分析如果要对该数据降维，往以下4个方向中哪个方向投影更合适？为什么？

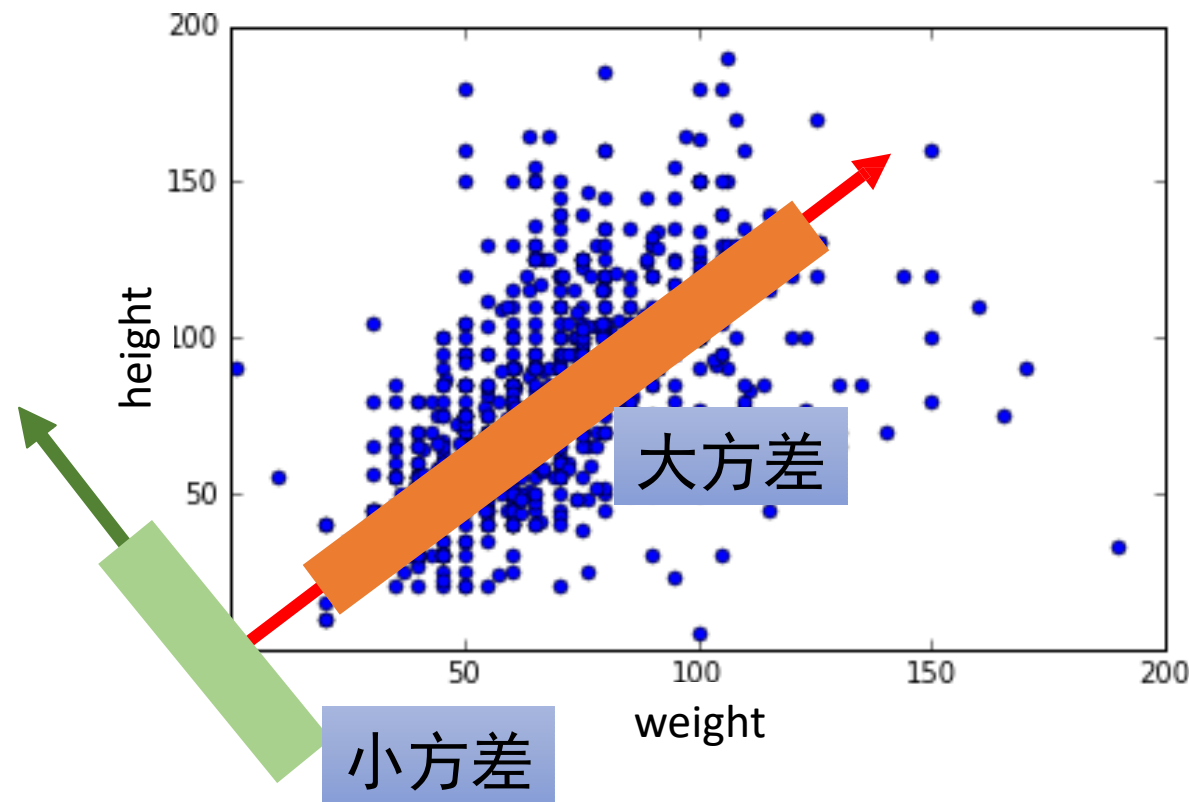


作答



主成分分析: 算法动机

- 在数理统计中，**方差**被经常用来度量数据和其数学期望（即均值）之间偏离程度，这个偏离程度反映了**数据分布结构**。
- 在许多实际问题中，研究数据和其均值之间的偏离程度有着很重要的意义。
- 在降维之中，需要尽可能将数据向**方差最大方向进行投影**，使得数据所蕴含信息没有丢失，彰显个性。
- 如右图所示，向红色方向比绿色方向投影结果在降维这个意义上好。





主成分分析: 算法动机

- 主成分分析思想是将 n 维特征数据映射到 l 维空间（ $n \gg l$ ），去除原始数据之间的冗余性（通过去除相关性手段达到这一目的）。
- 将原始数据向这些数据方差最大的方向进行投影。一旦发现了方差最大的投影方向，则继续寻找保持方差第二的方向且进行投影。
- 将每个数据从 n 维高维空间映射到 l 维低维空间，每个数据所得到最好的 k 维特征就是使得每一维上样本方差都尽可能大。



主成分分析: 若干概念-方差与协方差

数据样本的方差 variance

- 假设有 n 个数据, 记为 $X = \{x_i\} (i = 1, \dots, n)$
 - 方差等于各个数据与样本均值之差的平方和之平均数
 - 方差描述了样本数据的波动程度

$$Var(X) = \frac{1}{n} \sum_{i=1}^n (x_i - E(X))^2$$

- 其中 $E(X)$ 是样本均值, $E(X) = \frac{1}{n} \sum_{i=1}^n x_i$



主成分分析: 若干概念-方差与协方差

数据样本的协方差 covariance

- 假设有 n 个两维变量数据, 记为 $(X, Y) = \{(\mathbf{x}_i, \mathbf{y}_i)\}$ ($i = 1, \dots, n$)
 - 衡量两个变量之间的相关度

$$\text{cov}(X, Y) = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - E(X))(\mathbf{y}_i - E(Y))$$

- 其中 $E(X)$ 和 $E(Y)$ 分别是 X 和 Y 的样本均值, 分别定义如下

$$E(X) = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i, \quad E(Y) = \frac{1}{n} \sum_{i=1}^n \mathbf{y}_i$$

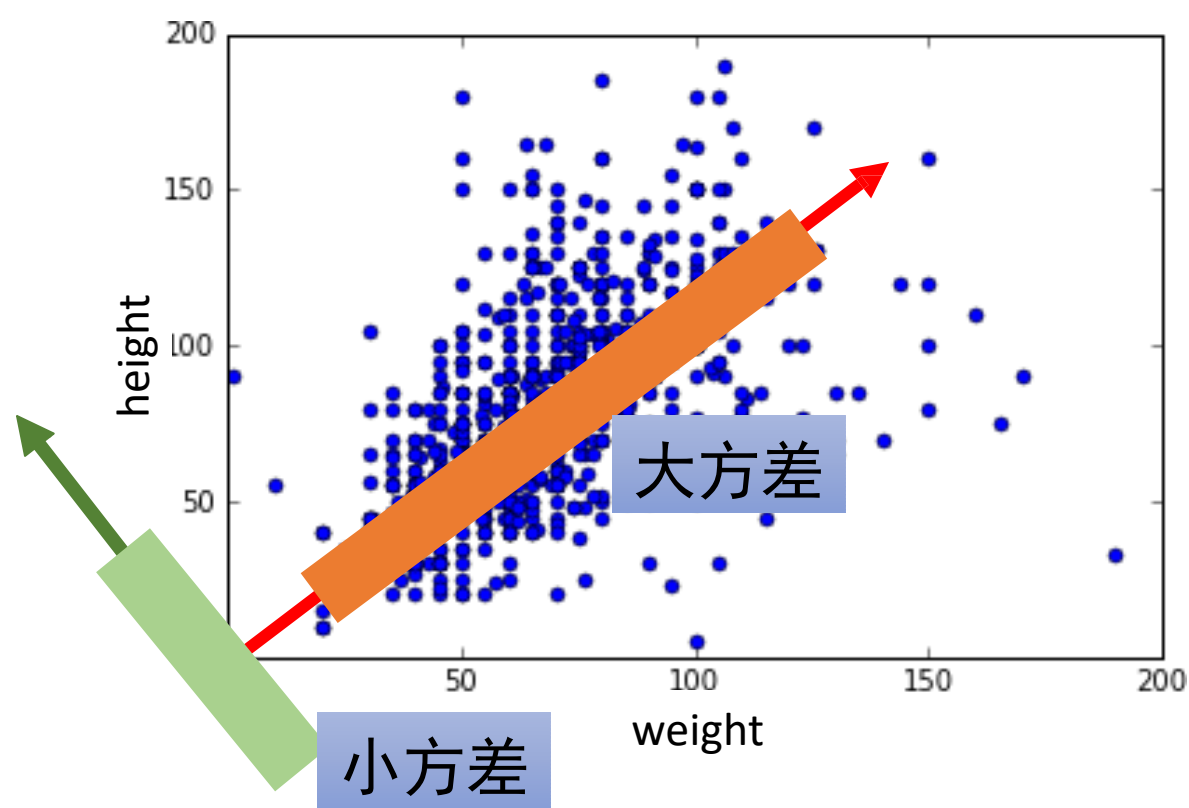
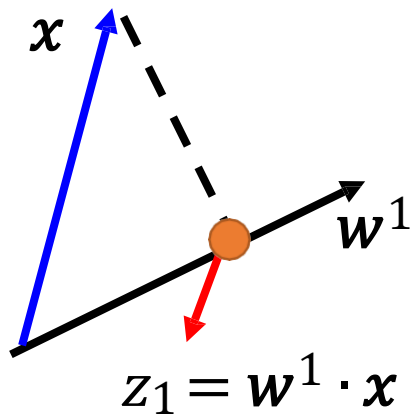


主成分分析：目标函数

$$z = Wx$$

降维到1维：

$$z_1 = (w^1)^T \cdot x$$



- 将所有数据 x 映射到 w^1 , 得到一组新的特征 z_1
- 希望映射后的数据 z_1 的方差尽可能大

$$\text{Var}(z_1) = \frac{1}{N} \sum_{z_1} (z_1 - \bar{z}_1)^2 \quad \|w^1\|_2 = 1$$

请思考, $\|\mathbf{w}^1\|_2 = 1$ 的必要性?

作答



主成分分析：目标函数

$$z = Wx$$

降维到多维：

$$z_1 = (w^1)^T \cdot x$$

$$z_2 = (w^2)^T \cdot x$$

$$W = \begin{bmatrix} (w^1)^T \\ (w^2)^T \\ \vdots \end{bmatrix}$$

正交矩阵

将所有数据 x 映射到 w^1 , 得到一组新的特征 z_1

希望映射后的数据 z_1 的方差尽可能大

$$Var(z_1) = \frac{1}{N} \sum_{z_1} (z_1 - \bar{z}_1)^2 \quad \|w^1\|_2 = 1$$

希望映射后的数据 z_2 的方差尽可能大

$$Var(z_2) = \frac{1}{N} \sum_{z_2} (z_2 - \bar{z}_2)^2 \quad \|w^2\|_2 = 1$$
$$(w^1)^T w^2 = 0$$



主成分分析：推导（略）

- 根据 $z_1 = (\mathbf{w}^1)^T \mathbf{x}$

$$\text{以及 } \bar{z}_1 = \frac{1}{N} \sum z_1 = \frac{1}{N} \sum (\mathbf{w}^1)^T \mathbf{x} = (\mathbf{w}^1)^T \frac{1}{N} \sum \mathbf{x} = (\mathbf{w}^1)^T \bar{\mathbf{x}}$$

- 得到 $Var(z_1) = \frac{1}{N} \sum (z_1 - \bar{z}_1)^2 = \frac{1}{N} \sum ((\mathbf{w}^1)^T \mathbf{x} - (\mathbf{w}^1)^T \bar{\mathbf{x}})^2$
 $= \frac{1}{N} \sum ((\mathbf{w}^1)^T (\mathbf{x} - \bar{\mathbf{x}}))^2 = (\mathbf{w}^1)^T \frac{1}{N} \sum (\mathbf{x} - \bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}})^T \mathbf{w}^1$
 $= (\mathbf{w}^1)^T Cov(\mathbf{x}) \mathbf{w}^1 = (\mathbf{w}^1)^T S \mathbf{w}^1$
- 因此，主成分分析的目标函数为

$$\begin{aligned} & \max (\mathbf{w}^1)^T S \mathbf{w}^1 \\ & s. t. \|\mathbf{w}^1\|_2 = (\mathbf{w}^1)^T \mathbf{w}^1 = 1 \end{aligned}$$



主成分分析：推导（略）

$$\max(\mathbf{w}^1)^T S \mathbf{w}^1 \quad s. t. \|\mathbf{w}^1\|_2 = (\mathbf{w}^1)^T \mathbf{w}^1 = 1$$

- 根据拉格朗日乘子法[Bishop, Appendix E]

$$g(\mathbf{w}^1) = (\mathbf{w}^1)^T S \mathbf{w}^1 - \alpha((\mathbf{w}^1)^T \mathbf{w}^1 - 1)$$

$$\partial g(\mathbf{w}^1) / \partial w_1^1 = 0$$

$$\partial g(\mathbf{w}^1) / \partial w_2^1 = 0$$

\vdots

$$S \mathbf{w}^1 - \alpha \mathbf{w}^1 = 0$$

$$S \mathbf{w}^1 = \alpha \mathbf{w}^1 \quad \mathbf{w}^1 : \text{特征向量}$$

$$(\mathbf{w}^1)^T S \mathbf{w}^1 = \alpha (\mathbf{w}^1)^T \mathbf{w}^1$$

$$= \alpha \quad \text{选择最大的特征值}$$

\mathbf{w}^1 是协方差矩阵 S 的一个特征向量

对应于最大特征值 λ_1



主成分分析：推导（略）

$$\begin{aligned} & \max(\mathbf{w}^2)^T S \mathbf{w}^2 \quad s. t. (\mathbf{w}^2)^T \mathbf{w}^2 = 1 \quad (\mathbf{w}^2)^T \mathbf{w}^1 = 0 \\ g(\mathbf{w}^2) &= (\mathbf{w}^2)^T S \mathbf{w}^2 - \alpha((\mathbf{w}^2)^T \mathbf{w}^2 - 1) - \beta((\mathbf{w}^2)^T \mathbf{w}^1 - 0) \end{aligned}$$

$$\left. \begin{aligned} \partial g(\mathbf{w}^2) / \partial w_1^2 &= 0 \\ \partial g(\mathbf{w}^2) / \partial w_2^2 &= 0 \\ &\vdots \end{aligned} \right\} \begin{aligned} S \mathbf{w}^2 - \alpha \mathbf{w}^2 - \beta \mathbf{w}^1 &= 0 \\ \underline{0} - \alpha \underline{0} - \beta \underline{1} &= 0 \\ &= ((\mathbf{w}^1)^T S \mathbf{w}^2)^T = (\mathbf{w}^2)^T S^T \mathbf{w}^1 \\ &= (\mathbf{w}^2)^T S \mathbf{w}^1 = \lambda_1 (\mathbf{w}^2)^T \mathbf{w}^1 = 0 \end{aligned}$$

$S \mathbf{w}^1 = \lambda_1 \mathbf{w}^1$

$$\beta = 0: \quad S \mathbf{w}^2 - \alpha \mathbf{w}^2 = 0 \quad S \mathbf{w}^2 = \alpha \mathbf{w}^2$$

\mathbf{w}^2 是协方差矩阵 S 的一个特征向量
对应于第二大特征值 λ_2



主成分分析: 算法描述

- 假设有 n 个 d 维样本数据所构成的集合 $X^{org} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ ，其中 $\mathbf{x}_i (1 \leq i \leq n) \in R^d$

- 主成分分析的主要步骤:

- 数据预处理: $X = \frac{X^{org} - \mu}{\sigma}$

- 计算协方差矩阵: $\Sigma = \frac{1}{n-1} X^T X$

- 求得协方差矩阵 Σ 的特征向量和特征根

- 取前 l 个最大特征根所对应的特征向量组成映射矩阵 W

- 给定一个样本 \mathbf{x}_i ，可将 \mathbf{x}_i 从 d 维空间如下映射到 l 维空间: $(\mathbf{x}_i)_{1 \times d} (W)_{d \times l}$

- 将所有降维后数据用 Y 表示，有 $Y = XW$

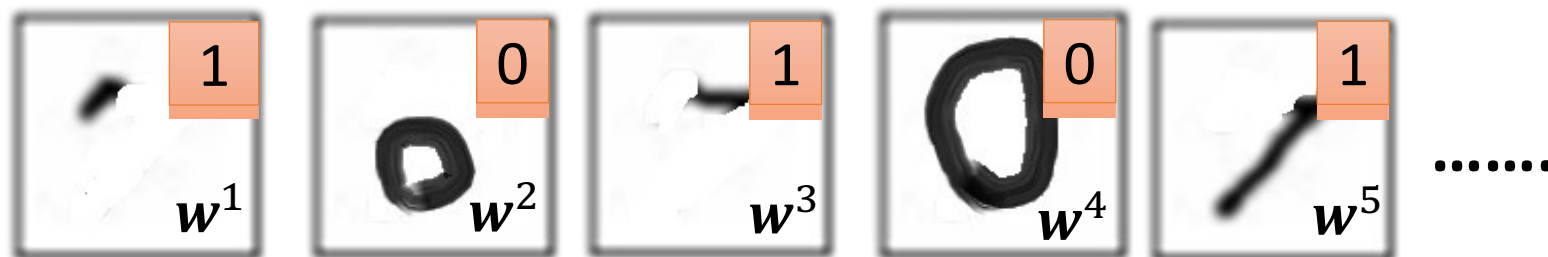
降维 原始映射
结果 数据矩阵

$$\begin{array}{l} Y \in R^{n \times l} \\ X \in R^{n \times d} \\ W \in R^{d \times l} \end{array} \quad \begin{array}{c} \text{蓝色} \\ Y \end{array} = \begin{array}{c} \text{黄色} \\ X \end{array} \times \begin{array}{c} \text{绿色} \\ W \end{array}$$

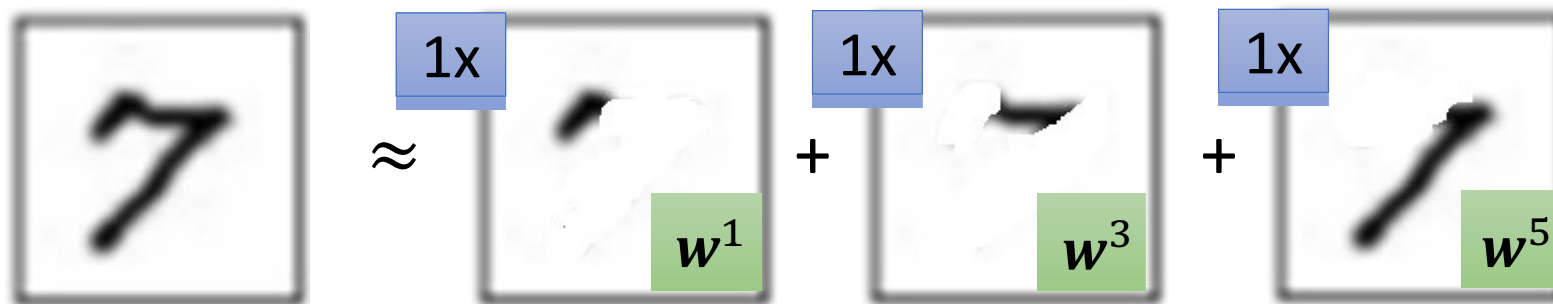


主成分分析: 另一个角度

主成分:



$\begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ \vdots \end{bmatrix}$



$$x \approx y_1 w^1 + y_2 w^2 + \dots + y_K w^K + \bar{x}$$

手写数字
图像的像
素值表达

主成分

$\begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_K \end{bmatrix}$

手写数字
图像新特
征表达



主成分分析: 另一个角度

$$\underset{\uparrow}{x - \bar{x}} \approx y_1 \mathbf{w}^1 + y_2 \mathbf{w}^2 + \cdots + y_K \mathbf{w}^K = \underset{\uparrow}{\hat{x}}$$

重构误差: $\|(x - \bar{x}) - \hat{x}\|_2$

优化问题: 通过最小化 L 得到 \mathbf{w} ,

$$L = \min_{\{\mathbf{w}^1, \dots, \mathbf{w}^K\}} \sum \left\| (x - \bar{x}) - \left(\sum_{k=1}^K y_k \mathbf{w}^k \right) \right\|_2$$



北京航空航天大学
COLLEGE OF SOFTWARE
BEIHANG UNIVERSITY 软件学院

提纲

一、K均值聚类

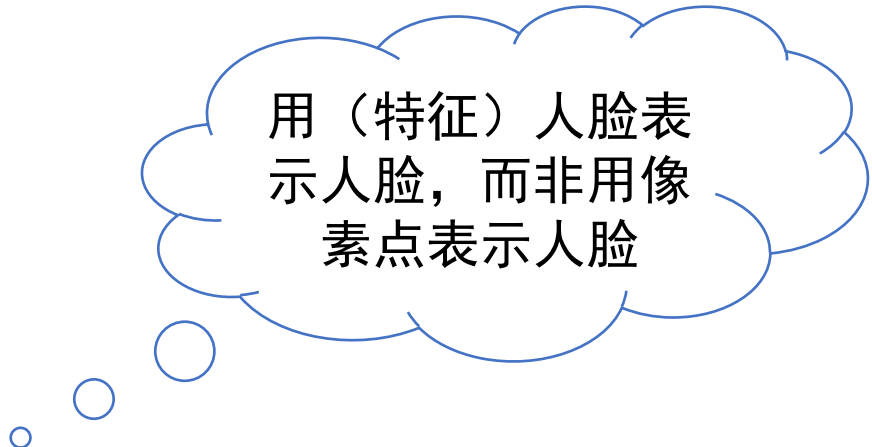
二、主成分分析

应用：特征人脸方法



特征人脸方法: 动机

- 特征人脸方法是一种应用主成分分析来实现人脸图像降维的方法
- 其本质是用一种称为“特征人脸 (eigenface)”的特征向量按照线性组合形式来表达每一张原始人脸图像, 进而实现人脸识别。
- 由此可见, 这一方法的关键之处在于如何得到特征人脸。



用 (特征) 人脸表示人脸, 而非用像素点表示人脸



北京航空航天大学
COLLEGE OF SOFTWARE
BEIHANG UNIVERSITY 软件学院

特征人脸方法: 算法描述



$$\begin{bmatrix} 45 & \dots & 68 \\ \dots & \dots & \dots \\ 36 & \dots & 86 \end{bmatrix}_{32 \times 32 = 1024}$$



$$\begin{bmatrix} 45 \\ \dots \\ 68 \\ \dots \\ \dots \\ 36 \\ \dots \\ 86 \end{bmatrix}_{1024 \times 1}$$

- 将每幅人脸图像转换成列向量
- 如将一幅 32×32 的人脸图像转成 1024×1 的列向量



特征人脸: 算法描述

$$Y \in R^{n \times l} \quad X \in R^{n \times d} \quad W \in R^{d \times l}$$

- 输入: n 个1024维人脸样本数据所构成的矩阵 X , 降维后的维数 l
- 输出: 映射矩阵 $W = \{w_1, w_2, \dots, w_l\}$ (其中每个 $w_j (1 \leq j \leq l)$ 是一个特征人脸)
- 算法步骤:

1: 对于每个人脸样本数据 x_i 进行中心化处理: $x_i = x_i - \mu, \mu = \frac{1}{n} \sum_{j=1}^n x_j$

2: 计算原始人脸样本数据的协方差矩阵: $\Sigma = \frac{1}{n-1} X^T X$

3: 对协方差矩阵 Σ 进行特征值分解, 对所得特征根从小到大排序 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d$

4: 取前 l 个最大特征根所对应特征向量 w_1, w_2, \dots, w_l 组成映射矩阵 W

5: 将每个人脸图像 x_i 按照如下方法降维: $(x_i)_{1 \times d} (W)_{d \times l} = 1 \times l$



特征人脸: 算法描述

- 每个人脸特征向量 w_i 与原始人脸数据 x_i 的维数是一样的，均为1024。
- 可将每个特征向量还原为 32×32 的人脸图像，称之为特征人脸，因此可得到 l 个特征人脸。

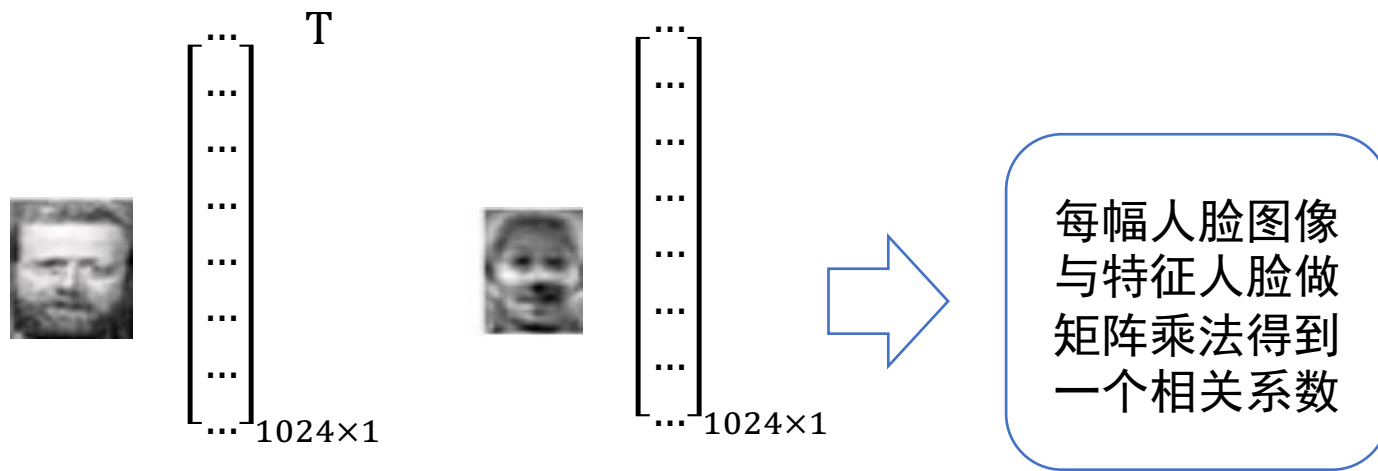


400个人脸（左）和与之对应的36个特征人脸



基于特征人脸的降维

- 将每幅人脸分别与每个特征人脸做矩阵乘法，得到一个相关系数
- 每幅人脸得到 l 个相关系数 \Rightarrow 每幅人脸从1024维约减到 l 维





基于特征人脸的降维

- 由于每幅人脸是所有特征人脸的线性组合，因此就实现人脸从“像素点表达”到“特征人脸表达”的转变。每幅人脸从1024维约减到 l 维。

$$x_i = \alpha_{i1} \times \text{face}_1 + \alpha_{i2} \times \text{face}_2 + \dots + \alpha_{il} \times \text{face}_l \quad \Rightarrow \quad (\alpha_{i1}, \dots, \alpha_{il})$$

x_i

使用 l 个特征人脸的线性组合来表达原始人脸数据 x_i

x_i 的像素点
空间表达
32×32

x_i 的人脸子
空间的 l 个系
数表达

在后续人脸识别分类中，就使用这 l 个系数来表示原始人脸图像。即计算两张人脸是否相似，不是去计算两个32×32矩阵是否相似，而是计算两个人脸所对应的 l 个系数是否相似



北京航空航天大学
COLLEGE OF SOFTWARE
BEIHANG UNIVERSITY 软件学院

人脸表达后的分析与处理

