《人工智能及应用》

第六讲：机器学习-无监督学习

授课人：张 鑫 zhangxin@uestc.edu.cn

专　业：机器人工程

提纲

1、无监督学习的概念

2、K均值聚类

3、主成份分析

1、有监督学习与无监督学习

有监督学习

Supervised learning

训练集

（带标签）

编号	就业 (x_1)	教育水平 (x_2)	居住时间 (x_3)	信用卡? (y)
1	是	研究生	5	Yes
2	是	高中	2	No
3	否	大学	1	No
4	是	高中	10	Yes
…				

学习算法

学习模型

$$y = f(x_1, x_2, x_3, \theta)$$

预测结果　输入变量　模型参数

从带标签的训练数据中，学习从数据到其标签的映射关系

数据的标注作为监督信号

最优化目标为预测结果与标注尽可能一致

1、有监督学习与无监督学习

有监督学习

Supervised learning

训练集

（带标注）

	x_1	x_2	x_3	y
编号	就业	教育水平	居住时间	信用卡?
1	是	研究生	5	Yes
2	是	高中	2	No
3	否	大学	1	No
4	是	高中	10	Yes
…				

归纳

学习算法

学习模型

$$y = f(x_1, x_2, x_3, \theta)$$

应用模型

（演绎）

$$y = f(x_1, x_2, x_3, \theta)$$

推理阶段

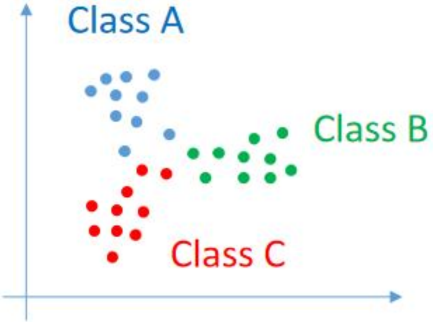
编号	就业	教育水平	居住时间	信用卡?
1	否	研究生	3	?
2	是	大学	2	?
…				

1、有监督学习与无监督学习

有监督学习

Supervised learning

训练数据带标签，在标签的监督下，发掘数据与标签之间的映射关系



老虎

狮子

大象

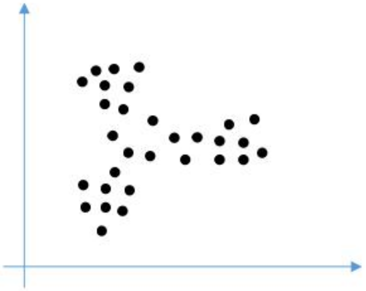
1、有监督学习与无监督学习

- **无监督学习：**是从无标注的数据样本中，学习数据中蕴含的模式，完成如**聚类**或**降维**等特定任务。

无监督学习

Unsupervised learning

训练数据无标签，如何仅利用数据本身的信息，发掘数据样本中蕴含的模式？



人类学习中，无监督学习占据主导地位：通过观察客观事务去洞悉其内在的规律。

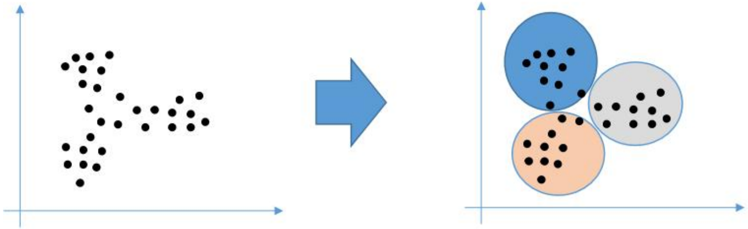
1、无监督学习的概念

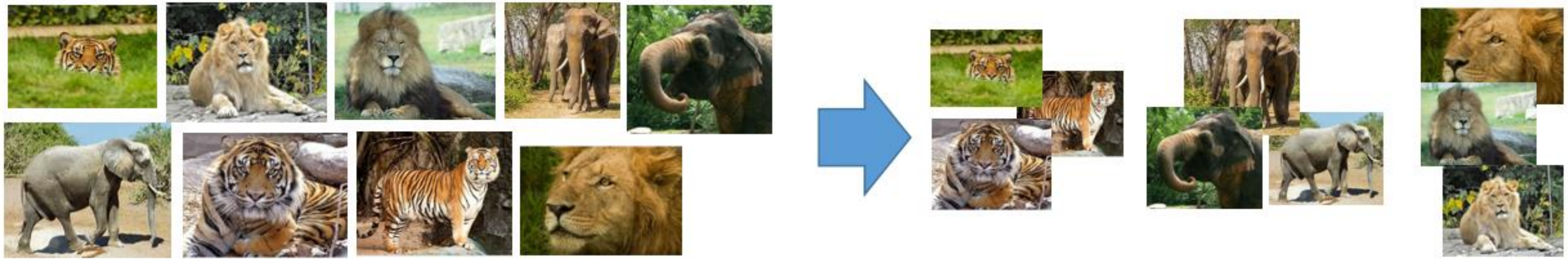
1.1概念——聚类、降维

- 1. **聚类：**根据事物的相似程度，将他们分成不同类别。

  物以类聚、人以群分。

  - 目标用户的群体分类
  - 图像切割
  - 基因聚类

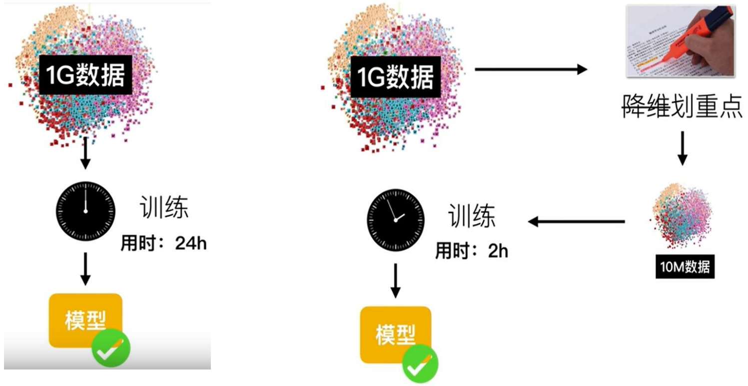


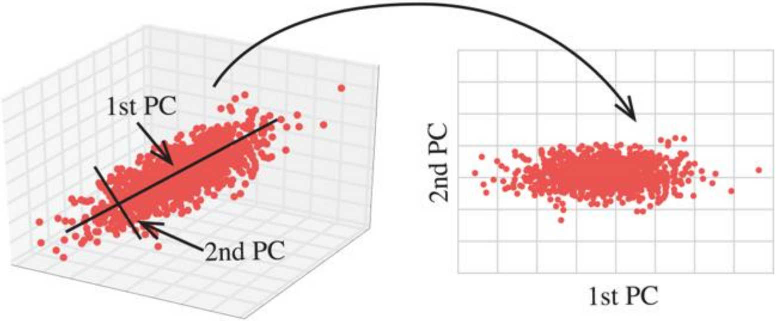


1、无监督学习的概念

1.1概念——聚类、降维

- 2. **降维：**实际上就是划重点、提精华。





Principal Components Analysis (PCA):主成份分析

提纲

1、无监督学习的概念

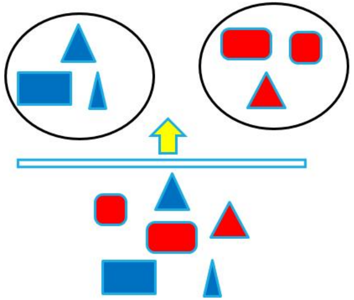
2、K均值聚类

3、主成份分析

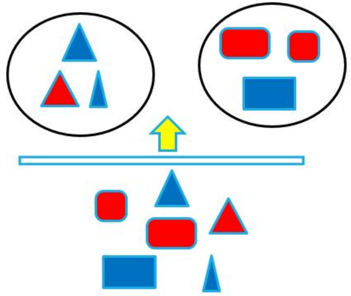
2、K均值聚类

2.1概念

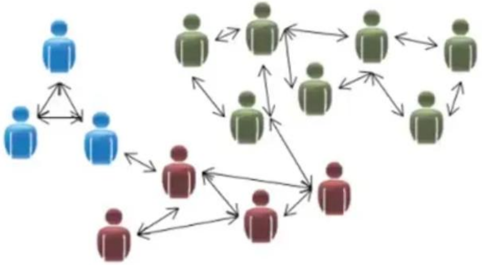
- 物以类聚，人以群分(《战国策·齐策三》)
- 目的：将n个数据聚类到k个集合（也称为类簇）
- 输入：n个数据（无任何标注信息），聚类数k
- 输出：k个聚类结果



- 换个说法：如何将n个数据依据**相似程度**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。



对人群进行社交关系分类，自动推荐可能认识的人



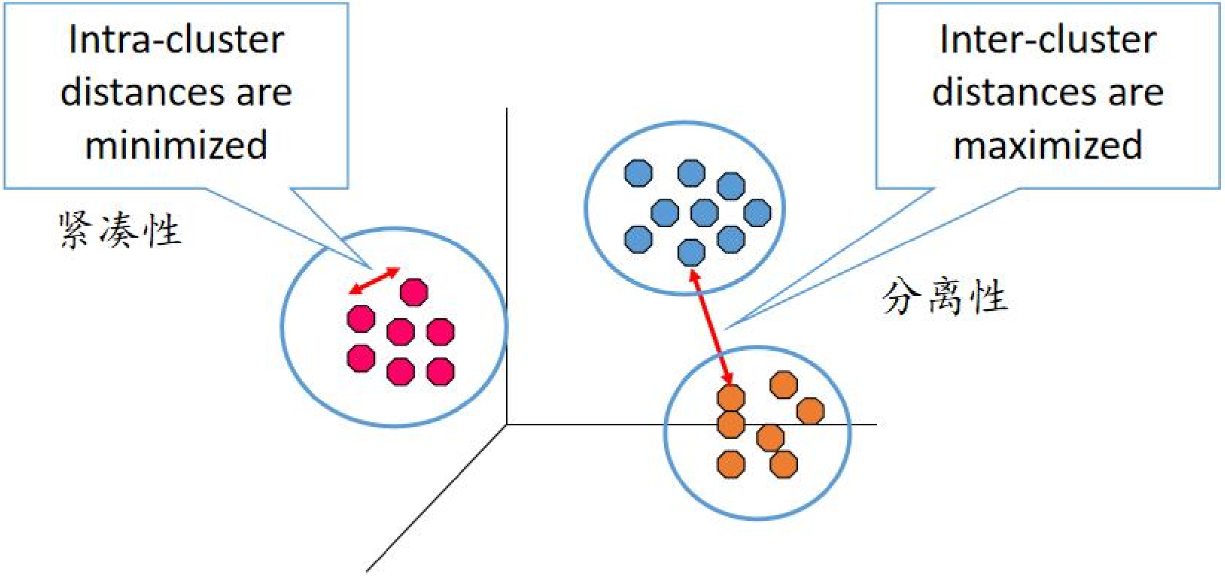
社交关系分组

2、K均值聚类

2.1概念

聚类的特点

- 同一个聚类中的样本相似（或距离近）
- 不同聚类间的样本不相似（或距离远）

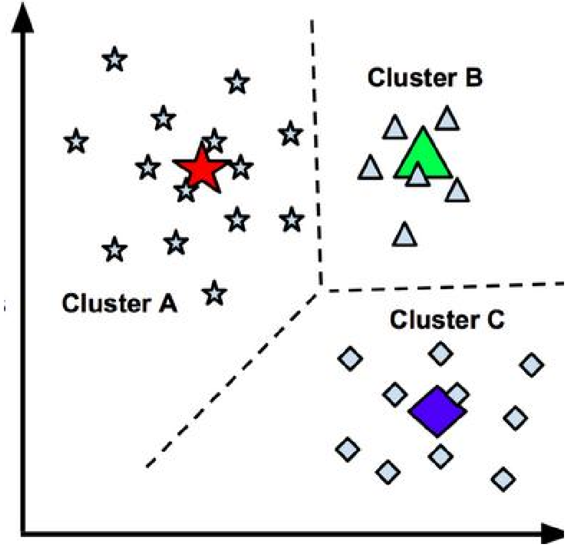


2、K均值聚类

2.1概念

是一种基于中心的分割型聚类方法

- 聚类个数k需指定
- 每个聚类有一个对应的中心点（Centroid）
- 每个样本被划分到聚类最近的中心点

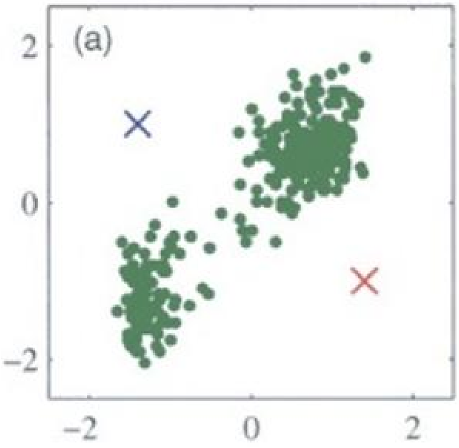


算法基本过程

1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

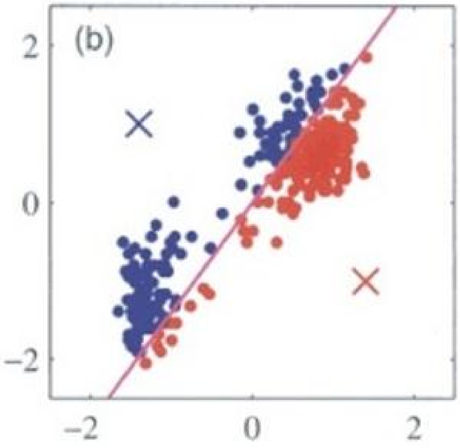
2、K均值聚类

2. 2算法



Initial centroids

初始中心

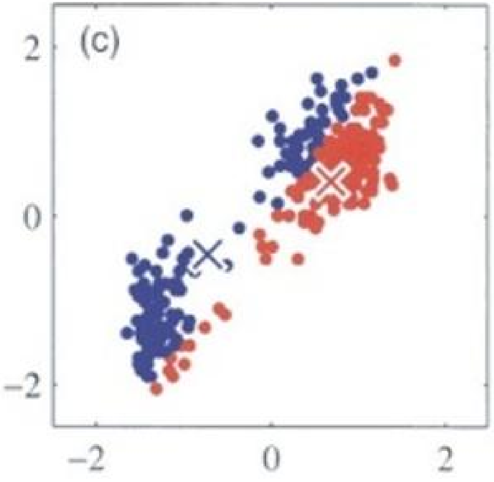


Initial clusters

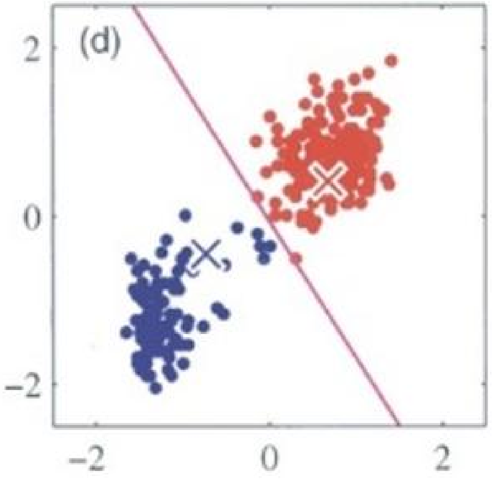
初始聚类

2、K均值聚类

2. 2算法



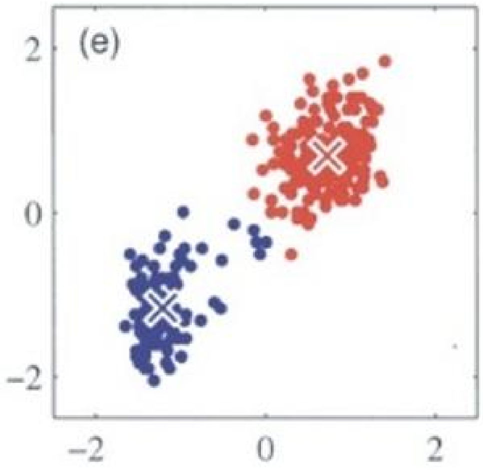
更新中心



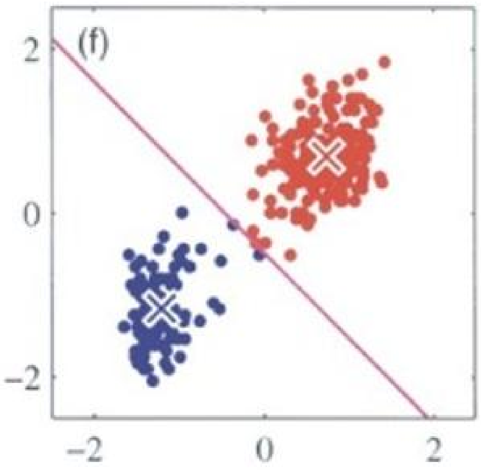
更新聚类

2、K均值聚类

2. 2算法



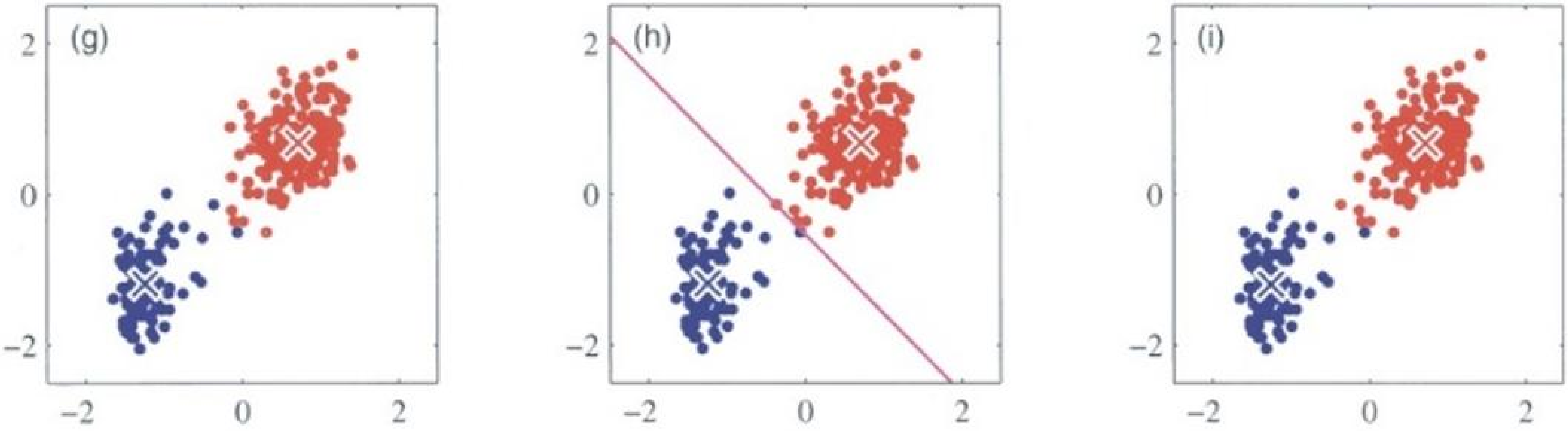
更新中心



更新聚类

2、K均值聚类

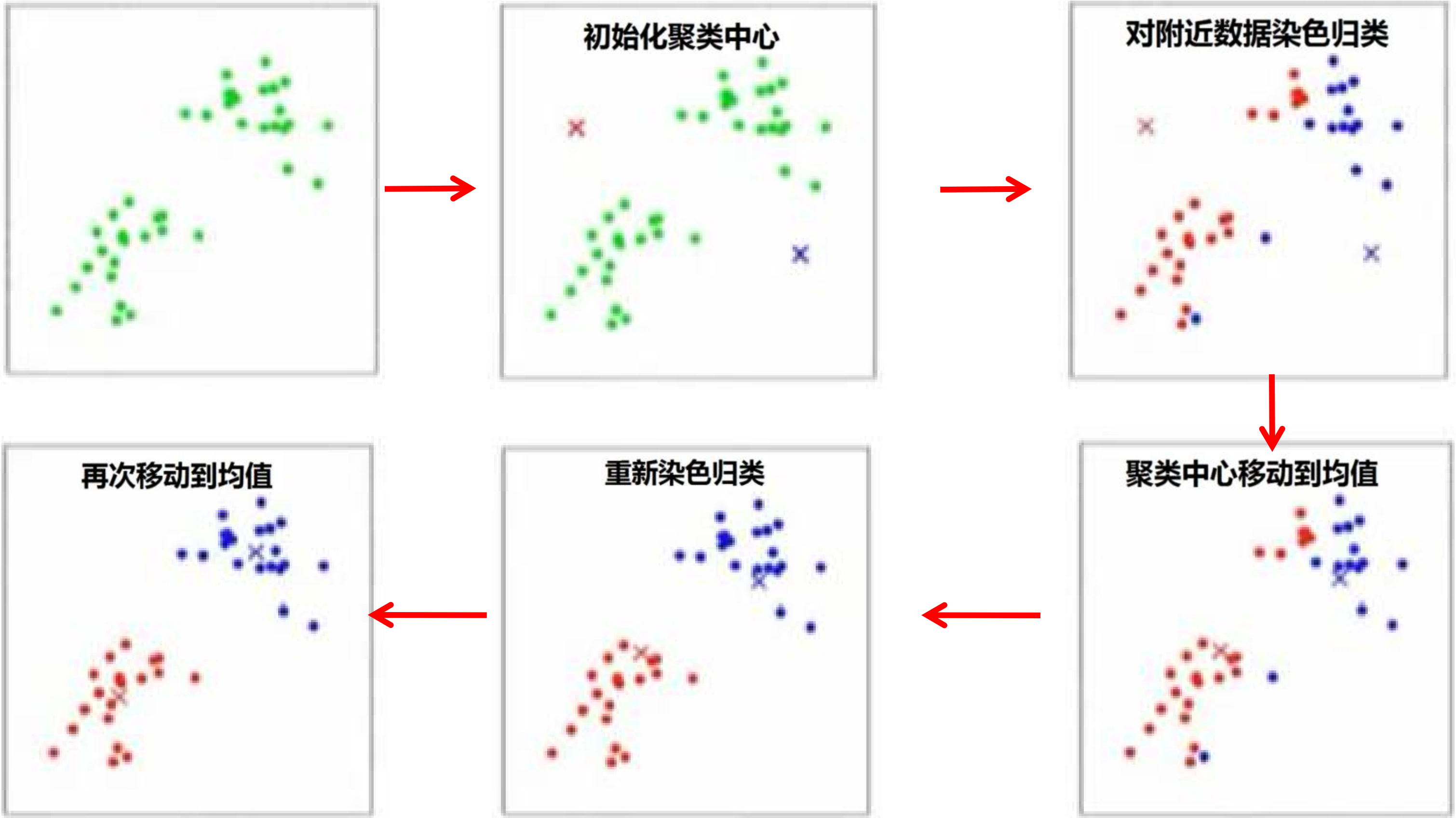
2. 2算法



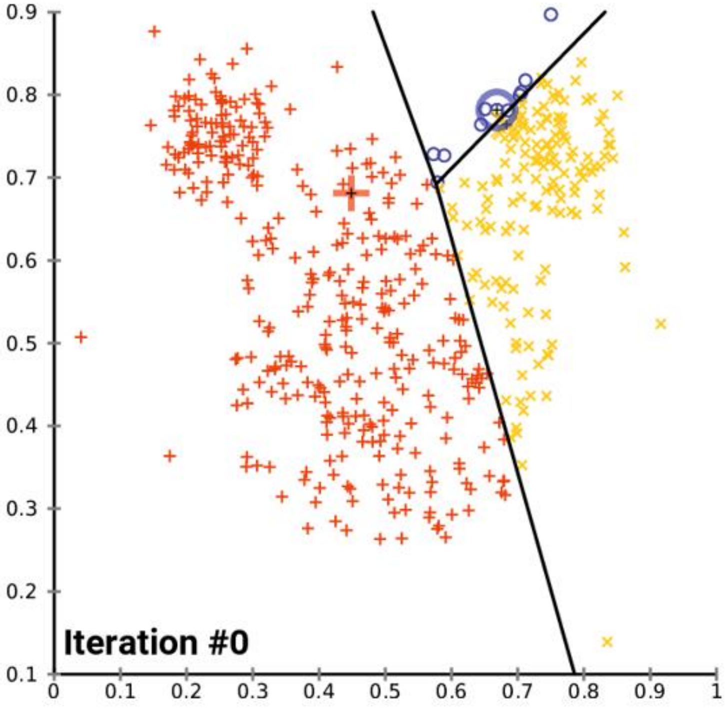
更新中心　　　　更新聚类　　　　最终结果

2、K均值聚类

2. 2算法



2.2算法



输入：聚类个数K

Step1: 设有K个聚类$C=\{C_1, C_2, ..., C_K\}$, 随机选择K个样本分别作为K个聚类的中心点$\{m_1, m_2, ..., m_K\}$

Step2:对每个样本x_j，将其分配到距离最近的聚类中心所对应的聚类

Step3:对第i个聚类，重新计算其中心

$$m'_i = \frac{1}{n_i}\sum_{x_j \in C_i} x_j$$

其中n_i为聚类C_i中的样本个数

Step4: 重复step2和step3，直到满足收敛条件（达到指定的迭代次数，聚类中心不再变化等）

输出：聚类结果

2、K均值聚类

2.2算法

- 相似度定义：用**欧式距离**描述数据之间的相似性。
- 有n个数据$\{x_1, x_2, \cdots, x_n\}$，（$1 \leqslant i \leqslant n$）
- 每个数据是m维的，$x_i = [x_{i1} \quad x_{i2} \quad \cdots \quad x_{im}]^T$
- 两个m维数据之间的**欧氏距离**定义为：

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

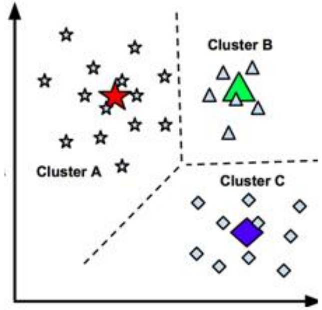
2、K均值聚类

2.2算法

- 需要事先确定聚类数目K，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

2.2算法 需要事先确定聚类数目K

- 从目标优化的角度看待：误差平方和（Sum of Squared Error, SSE）

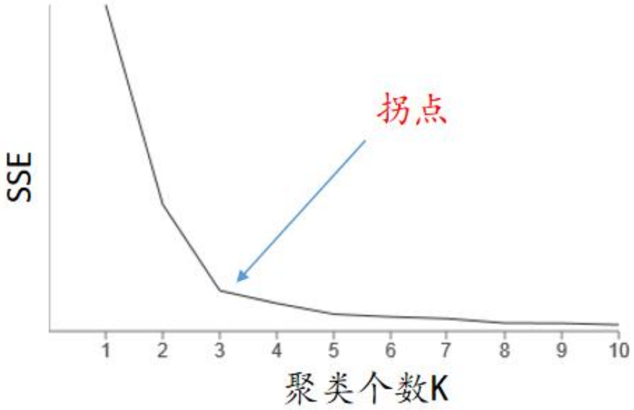


$$SSE = \sum_{i=1}^{K}\sum_{x\in C_i} dist^2(m_i, x)$$

x是聚类C_i中的一个样本

m_i是聚类C_i的中心

- 聚类的目标是实现SSE的最小化
  - 前提是K需要事先给定
  - 如何自动确定K的值？
    - 根据SSE的拐点



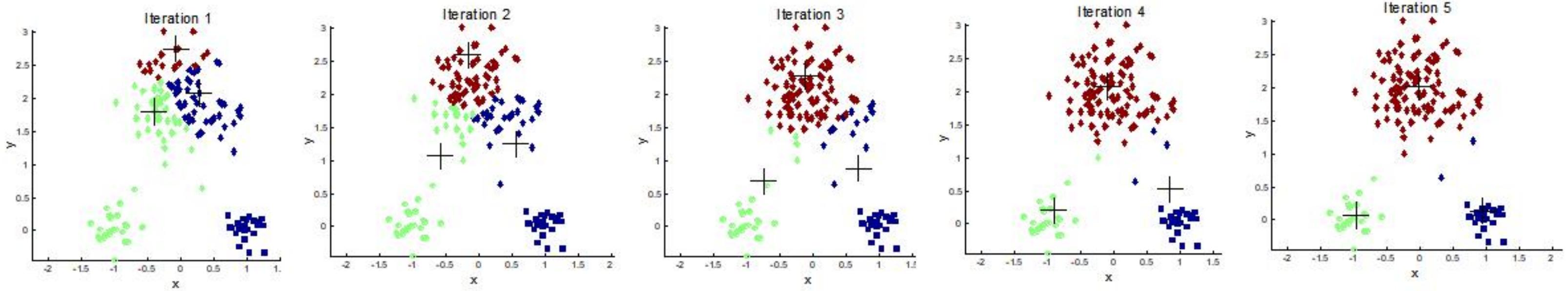
2、K均值聚类

2.2算法

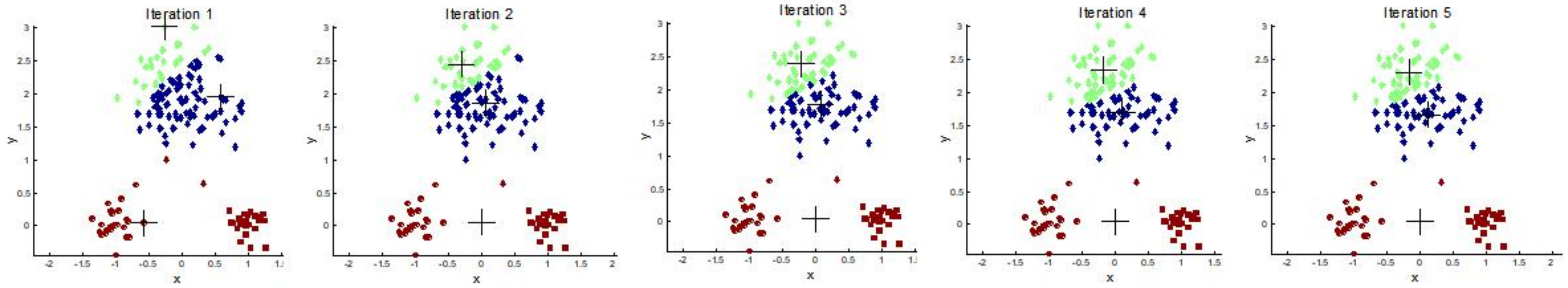
需要初始化聚类质心

- 不是全局最优（对初始化敏感）
- K需要事先指定
- 聚类的边界是线性的
- 计算开销较大

好的初始化



不好的初始化



2、K均值聚类

2.2算法

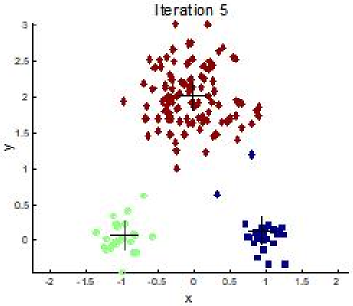
如何解决对初始化敏感的问题?

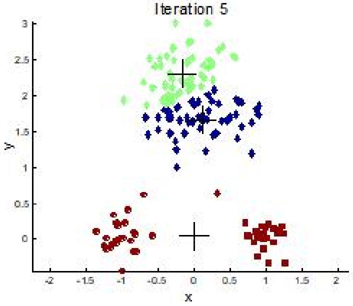
1）多次运行，选择SSE最小的一次结果

2）通过采样或者分层次聚类得到更稳定的初始化

3）后处理

- 去除小的聚类（有可能是离群点Outlier）
- 将松散（较高的SSE值）的聚类分解
- 将两个邻近的小聚类融合





2、K均值聚类

2. 3举例和应用

将1、2、3、4、5、6、7、8、9九个整数，用K均值分为3类。

第0步：初始化三个中心点： $E(c_1^0)=1$，$E(c_2^0)=2$，$E(c_3^0)=3$

第1a步：分别计算各样本距离中心点的距离，将样本进行分类；

$$c_1^0=\{1\} \qquad c_2^0=\{2\} \qquad c_3^0=\{3,4,5,6,7,8,9\}$$

第1b步：计算分别分类的均值：

$$\mathrm{E}(c_1^1)=1 \qquad \mathrm{E}(c_2^1)=2 \qquad \mathrm{E}(c_3^1)=6$$

2、K均值聚类

2.3举例和应用

将1、2、3、4、5、6、7、8、9九个整数，用K均值分为3类。

第2a步：分别计算各样本距离中心点的距离，将样本进行分类；

$$\mathrm{E}(c_1^1)=1 \qquad \mathrm{E}(c_2^1)=2 \qquad \mathrm{E}(c_3^1)=6$$

$$c_1^1=\{1\} \qquad c_2^1=\{2,3,4\} \qquad c_3^1=\{5,6,7,8,9\}$$

第2b步：计算分别分类的均值：

$$\mathrm{E}(c_1^2)=1 \qquad \mathrm{E}(c_2^2)=3 \qquad \mathrm{E}(c_3^2)=7$$

将1、2、3、4、5、6、7、8、9九个整数，用K均值分为3类。

第3a步：分别计算各样本距离中心点的距离，将样本进行分类；

$$\mathrm{E}(c_1^2)=1 \quad \mathrm{E}(c_2^2)=3 \quad \mathrm{E}(c_3^2)=7$$

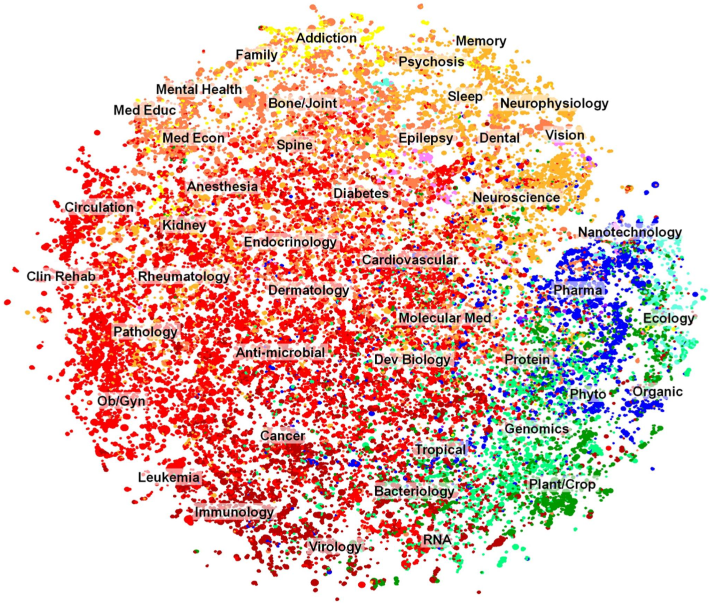
$$c_1^2=\{1,2\} \quad c_2^2=\{3,4,5\} \quad c_3^2=\{6,7,8,9\}$$

第3b步：计算分别分类的均值：

$$\mathrm{E}(c_1^3)=2 \quad \mathrm{E}(c_2^3)=4 \quad \mathrm{E}(c_3^3)=8$$

2、K均值聚类

2.3应用



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

提纲

1、无监督学习的概念

2、K均值聚类

3、主成份分析

3、主成份分析

3.1概念

- 主成份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”
- Principle Component Analysis （PCA）
- 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”

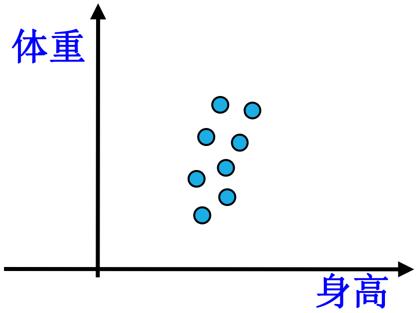
3、主成份分析

3.1概念

- 主成份分析主要用于发现数据中的基本结构(数据之间的关联)，是数据分析的主要工具；
- 主成份分析，可把数据用主要成份来表示，这可理解为对数据降维。
- 主成份分析属于多元统计分析的经典方法，最早由Pearson于1901年提出，当时只针对非随机变量，1933年由Hotelling推广到多维随机变量；

1）有一群女神参加选美，身高、体重二维数据如右图；
2）降维：现在需用一个指标(数据)来对女神排序；
　　选身高还是体重作为排序的依据？

选方差大的数据维度
因为对应的信息量大



3、主成份分析

3. 1概念

- 降维后的结果要保持原始数据固有结构

- 原始数据中的结构
  - 图像数据中结构：视觉对象区域构成的空间分布
  - 文本数据中结构：单词之间的(共现)相似或不相似

200万像素点

约减

60个像素点

3、主成份分析

3.2数据基础：均值、方差、协方差

- **均值**：假设有n个数据，记为X={$x_{i,}$ i=1...n}，均值u为：

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

- **方差**：描述了样本数据的波动程度，等于各个数据与样本均值之差的平方和之平均数：

$$\text{var}(X) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - u)^2$$

- **协方差**：衡量两个变量之间的相关程度。假设样本由两个变量X、Y描述，观察到了n组值，记为(X, Y)={(x_1, y_1), (x_2, y_2),......(x_n, y_n)}

$$\text{cov}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - E(X))(y_i - E(Y))$$

cov(X,Y)>0，正相关；cov(X,Y)<0负相关；cov(X,Y)=0，不相关

3、主成份分析

3.2数据基础：均值、方差、协方差

编号	x_i	y_i	$x_i - E(X)$	$y_i - E(Y)$	$[x_i - E(X)[y_i - E(Y)]$
1	1	7	-8.33	-16.67	138.89
2	3	11	-6.33	-12.67	80.22
3	6	17	-3.33	-6.67	22.22
4	10	25	0.67	1.33	0.89
5	15	35	5.67	11.33	64.22
6	21	47	11.67	23.33	272.22
	$E(X)$ =9.33	$E(Y)$ =23.67	$Var(X)$ =48.22	$Var(Y)$ =192.89	$E([x_i - E(X)][y_i - E(Y)])$ =96.44

$$u = E(X) = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \mathrm{var}(X) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - u)^2 \qquad \mathrm{cov}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - E(X))(y_i - E(Y))$$

3、主成份分析

3.2数据基础：协方差矩阵

- **m维随机变量：** 假设m维变量，记为$x=[x_1, x_2, ..., x_m]^T$，观测到了n组值，X可写为矩阵形式：

$$X=\begin{bmatrix}\mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_n\end{bmatrix}=\begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix}$$

- **协方差矩阵：** 定义如下：

$$\Sigma=\operatorname{cov}(X,X)=\begin{bmatrix} \operatorname{cov}(\mathbf{x}_1,\mathbf{x}_1) & \operatorname{cov}(\mathbf{x}_1,\mathbf{x}_2) & \dots & \operatorname{cov}(\mathbf{x}_1,\mathbf{x}_m) \\ \operatorname{cov}(\mathbf{x}_2,\mathbf{x}_1) & \operatorname{cov}(\mathbf{x}_2,\mathbf{x}_2) & \dots & \operatorname{cov}(\mathbf{x}_2,\mathbf{x}_m) \\ \dots & \dots & \dots & \dots \\ \operatorname{cov}(\mathbf{x}_m,\mathbf{x}_1) & \operatorname{cov}(\mathbf{x}_m,\mathbf{x}_2) & \dots & \operatorname{cov}(\mathbf{x}_m,\mathbf{x}_m) \end{bmatrix}$$

3、主成份分析

3.3主成份定义

- **线性变换：** 考虑m维变量$x=[x_1, x_2, ..., x_m]^T$，到m维随机变量$y=[y_1, y_2, ..., y_m]^T$的线性变换

$$\mathbf{y}=\begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_m \end{bmatrix}=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}\cdot\begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_m \end{bmatrix}=\begin{bmatrix} a_1 \\ a_2 \\ \cdots \\ a_m \end{bmatrix}\cdot\mathbf{x}=Ax$$

- 如上式的线性变换，**主成份**定义如下：

1)未知变换的系数向量a_i**为单位向量**，$\|a_i\|=1 \quad i=1,2\ldots m$

2)y_i与y_j互不相关，即 $\mathrm{cov}(y_i, y_j)=0 \quad i\neq j$

3)**y1**是所有x的线性变换中**方差最大**的，**y2**是x的所有线性变换中与y1不相关中方差最大的……；称y1, y2.....为x的**第一主成份**、**第二主成份**.....；

3、主成份分析

3.4主成份分析

- **主成份分析：** 实际上求解一组标准正交基，a1, a2. . . . . . am, 使变换后的随机向量y1, y2, ... ym的方差取极值，即：

$$\max_{a_1} \operatorname{var}(y_1) = \max_{a_1} \operatorname{var}(a_1 x) \qquad s.t.\quad a_1 a_1^{\mathrm{T}} = 1$$

推导如下：

$$\operatorname{var}(y_1) = \operatorname{var}(a_1 x) = \frac{1}{n-1}\sum_{i=1}^{n}(a_1 x^{(i)} - E(a_1 x^{(i)}))^T (a_1 x^{(i)} - E(a_1 x^{(i)}))^T$$

$$= \frac{1}{n-1} a_1 \sum_{i=1}^{n}(x^{(i)} - E(x^{(i)}))^T (x^{(i)} - E(x^{(i)}))^T a_1^T$$

$$= a_1 \operatorname{cov}(\mathrm{X},\mathrm{X}) a_1^T \quad = a_1 \Sigma a_1^T$$

$$\operatorname{var}(X) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - u)^2$$

$$\operatorname{cov}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - E(X))(y_i - E(Y))$$

原问题等价为：

$$\max_{a_1} \operatorname{var}(y_1) = \max_{a_1} a_1 \Sigma a_1^T \qquad s.t.\quad a_1 a_1^{\mathrm{T}} = 1$$

3.4主成份分析

- 等式约束的拉格朗日极值问题：

$$\max_{a_1} \mathrm{var}(y_1) = \max_{a_1} a_1 \Sigma a_1^T \qquad s.t. \quad a_1 a_1^{\mathrm{T}} = 1$$

- 用拉格朗日乘子法，构建拉格朗日函数：

$$L(a_1, \lambda) = a_1 \Sigma a_1^T - \lambda(a_1 a_1^T - 1)$$

对拉格朗日函数取偏导，并令其为0，

$$\frac{\partial L(a_1, \lambda)}{\partial a_1} = \Sigma a_1^T - \lambda a_1^T = 0$$

$\Sigma = \mathrm{cov(X, X)}$

$\Rightarrow \Sigma a_1^T = \lambda a_1^T$

a_1为协方差矩阵的单位特征向量；

λ 为特征值。

$\Rightarrow \mathrm{var}(y_1) = a_1 \Sigma a_1^T = \lambda a_1 a_1^T = \lambda$

3、主成份分析

3.4主成份分析

- 主成份的性质：

设x是m维随机变量，Σ是x的协方差矩阵，Σ的特征值分别为 $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_k$ 对应的特征向量分别为a1、a2. . . . ak, 那么x对应的第k主成份是：

$$y^{(j)}{}_k = a_k x^{(j)} \quad j = 1,2 \ldots n$$

那么x对应的第k主成份yk的协方差是协方差矩阵的第k个特征值：

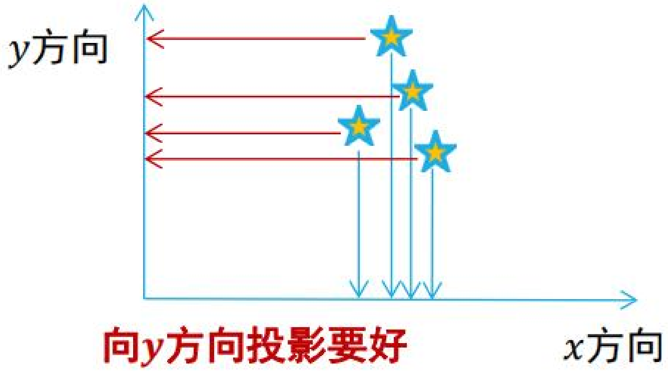
$$\mathrm{var}(y_k) = \lambda_k$$

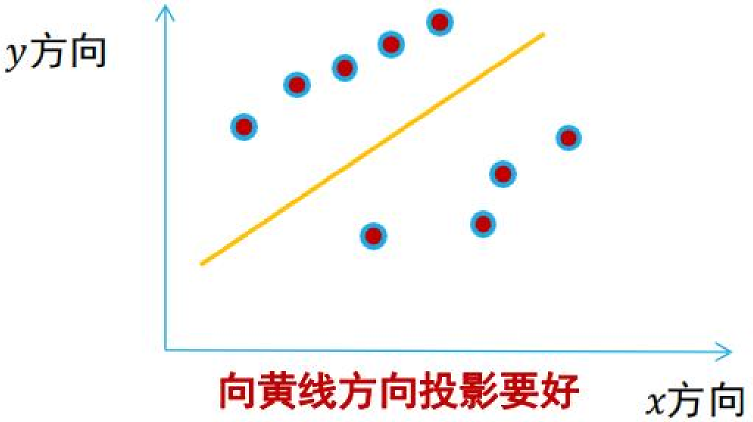
3、主成份分析

3.4主成份分析：算法动机

保证样本
投影后方差最大

- 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。
- 如左下图所示，向　方向投影（使得二维数据映射为一维）就比向　方向投影结果在降维这个意义上而言要好；右下图则是黄线方向投影要好。





3.4主成份分析：算法动机

- 主成分分析思想是将m维特征数据映射到k维空间（ k ≫ m），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m维高维空间映射到k维低维空间，每个数据所得到最好的特征就是使得每一维上样本方差都尽可能大。
- 主成份维数k的选择, 可由线性变换后期望保留的信息量比例来决定：

$$\frac{\lambda_1+\lambda_2+\ldots\lambda_k}{\lambda_1+\lambda_2+\ldots\lambda_m}\geq\sigma\quad（\sigma=0.95,\ldots）$$

- 不足及改进：当数据量比较大时，协方差矩阵计算费时，特征值和特征向量求解困难。可采用矩阵的**奇异值分解**来求解特征值和特征向量(参考《矩阵分析》）

3、主成份分析

3.4主成份分析：算法动机

- 假设有n个d维样本数据所构成的集合$D=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_n\}$，其中$\boldsymbol{x}_i(1\leq i\leq n)\in R^d$。
- 集合D可以表示成一个$n\times d$的矩阵$\mathbf{X}$。
- 假定每一维度的特征均值均为零（已经标准化）。
- 主成分分析的目的是求取一个且使用一个$d\times l$的映射矩阵$\mathbf{W}$。
- 给定一个样本$\boldsymbol{x}_i$，可将$\boldsymbol{x}_i$从d维空间如下映射到l维空间：$(\boldsymbol{x}_i)_{1\times d}(\mathbf{W})_{d\times l}$
- 将所有降维后数据用$\mathbf{Y}$表示，有$\mathbf{Y}=\mathbf{X}\,\mathbf{W}$

? 如何求取映射矩阵$\mathbf{W}$

降维结果 原始数据 映射矩阵

- $\mathbf{Y}=n\times l$
- $\mathbf{X}=n\times d$
- $\mathbf{W}=d\times l$

3.4主成份分析：算法动机

- 输入： n个d维样本数据所构成的矩阵$\mathbf{X}$，降维后的维数l
- 输出：映射矩阵$\mathbf{W}=\{\boldsymbol{w}_1,\boldsymbol{w}_2,\dots,\boldsymbol{w}_l\}$
- 算法步骤：

1：对于每个样本数据$\boldsymbol{x}_i$进行中心化处理： $\boldsymbol{x}_i=\boldsymbol{x}_i-\mu,\mu=\frac{1}{n}\sum_{j=1}^{n}x_j$

2: 计算原始样本数据的协方差矩阵：$\mathbf{\Sigma}=\frac{1}{n-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{X}$

3: 对协方差矩阵$\mathbf{\Sigma}$进行特征值分解，对所得特征根按其值大到小排序$\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_l$

4: 取前l个最大特征根所对应特征向量$\boldsymbol{w}_1,\boldsymbol{w}_2,\dots,\boldsymbol{w}_l$组成映射矩阵$\mathbf{W}$

5：将每个样本数据$\boldsymbol{x}_i$按照如下方法降维：$(\boldsymbol{x}_i)_{1\times d}(\mathbf{W})_{d\times l}=1\times l$

谢谢