数据科学导论第 10 讲—— 无监督学习

王小宁

中国传媒大学数据科学与智能媒体学院

2021 年 05 月 27 日

目录

无监督学习

无监督学习

- 现实生活中常常会有这样的问题：缺乏足够的先验知识，因此难以人工标注类别或进行人工类别标注的成本太高。很自然地，我们希望计算机能代我们完成这些工作，或至少提供一些帮助。
- 根据类别未知 (没有被标记) 的训练样本解决模式识别中的各种问题，称之为无监督学习。

问题的提出

例子 1

- 已知有 18 种花卉，并测得这些花卉的 8 个不同指标数据：V1（是否能过冬）、V2（是否生长在阴暗的地方）、V3（是否有块茎）、V4（花卉颜色）、V5（所生长泥土）、V6（某人对这 18 种花卉的偏好选择）、V7（花卉高度）、V8（花卉之间所需的间隔距离）
- 若要将这 18 种花卉进行聚类，那么应该聚成几类？如何划分？

数据 1

Obs	V1	V2	V3	V4	V5	V6	V7	V8
x1	0	1	1	4	3	15	25	15
x2	1	0	0	2	1	3	150	50
x3	0	1	0	3	3	1	150	50
				...	...			

例 2 主成分分析

- R 中 USArrests 数据集经常被用来做主成分分析，该数据集是 1973 年美国 50 个州的犯罪率指标，它包含 50 个观测值和 4 个变量，见下表。其中 Murder、Assault、Rape 三个变量分别为每 10 万居民中被逮捕的谋杀、暴力和强奸犯罪人数，UrbanPop 表示各州城市人口比例。
- 我们想考虑的问题是如何用综合的变量来总结这些信息，并对各州犯罪率水平进行评价。

Obs	Murder	Assault	UrbanPop	Rape
AlabaMa	13.2	236	58	21.2
Alaska	10.0	263	48	44.5
Arizona	8.1	294	80	31.0
		...	...	

例 3 因子分析

- 某班 30 名学生的数学、物理、化学、语文、历史、英语 6 门课程成绩如表 12-3 所示，该如何根据学生的成绩分析该学生较适合学文科还是理科呢?

Obs	数学	物理	化学	语文	历史	英语
x1	65	61	72	84	81	79
x2	77	77	76	64	70	55
x3	67	63	49	65	67	57
			...	...		

例 4 典型相关分析

- 测量 15 名受试者的身体形态和健康情况指标，见下表。身体形态变量有年龄、体重、胸围和日抽烟量，健康情况指标有脉搏、收缩压和舒张压三项。如果想要研究身体形态和健康状况的关系，该如何进行分析呢?

Obs	年龄	体重	抽烟量	胸围	脉搏	收缩压	舒张压
x1	25	125	30	83.5	70	130	85
x2	26	131	25	82.9	72	135	80
x3	28	128	35	88.1	75	140	90
			...	...			

聚类分析

聚类分析

- 聚类分析是数据科学的一个重要工具，在信息检索、生物学、心理学、医学、商业等领域有广泛的应用。
- 在经济研究中，为了研究不同地区城镇居民生活中的收入和消费情况，往往需要划分不同的类型去研究；
- 在商业中，把客户细分为几类，根据不同类的客户特征有针对性地营销。
- 聚类分析通常分为 R 型聚类和 Q 型聚类，R 型聚类是对样品进行聚类，Q 型聚类是对变量进行聚类。目前的研究多数是对样品进行聚类，因此本节主要以样品聚类为实例对聚类方法进行介绍。

聚类定义

- 聚类分析的目标是使得组内的对象相互之间高度相似（相关），而不同组中的对象间差异尽可能大。而衡量这种差异的指标有很多，最常用的是“距离”
- 用 $C = \{C_1, c_2, \cdots, C_k\}$ 表示在每个类中包含观测序号的集合，“不重不漏”规则：

1. $C_1 \bigcup C_2 \bigcup \cdots \bigcup C_k = \{1, 2, \cdots, n\}$，即每个观测都至少属于一类
2. $C_1 \bigcap C_j = \emptyset, i \neq j$ 即类与类之间是无重叠的，没有一个观测同时属于两个或更多类

相异度——数值型数据

- 明考夫斯基距离（Minkowski distance）

$$d_{ij} = (\sum_{k=1}^{p} |X_{ik} - X_{jk}|^q)^{1/q}, q > 0$$

- 明考夫斯基距离简称明氏距离，依据 q 的不同取值可以分成:
  - 绝对距离（$q = 1$，Absolute distance)
  - 欧氏距离（$q = 2$，Euclidean distance)
  - 切比雪夫距离（$q = \infty$，Chebyshev distance)

- 马氏距离（Mahalanobis distance）

$$d_{ij}^2 = (X_i - X_j)' \Sigma^{-1} (X_i - X_j)$$

- 马氏距离考虑了观测变量之间的变异性，且不再受各指标量纲的影响。将原始数据作线性变换后，马氏距离不变。

- 余弦距离（余弦相似度，Cosine similarity）

$$d_{ij} = \frac{X_i^{'} X_j}{||X_i||_2 ||X_j||_2}$$

- 余弦距离关注的是个体方向上的差异，对绝对数值不敏感，所以可以解决例如不同个体间存在的度量标准不统一的问题

相异度——分类型数据

- 二元变量（0 或 1）

X	1	0	Sum
1	q	r	q + r
0	s	t	s + t
Sum	q + s	r + t	p

- q：表示观测点 X_i 和 X_j 中取值都为 1 的变量数目
- r：表示仅在 X_i 中取值为 1，而在 X_j 中取值为 0 的变量数目
- s：表示在观测 X_i 中取值为 0，观测 X_j 中取值为 1 的变量数目
- t：是在观测点 X_i 和 X_j 中值都为 0 的变量数目
- 观测点 X_i 和 X_j 的距离可以定义为：$d_{ij} = \frac{r+s}{p}$

名义变量（变量取值多于两类）

- 例如，地图的颜色是一个名义变量，它可能有五个水平：红色、黄色、绿色、粉红色和蓝色。
- 名义变量的水平可以用字母、符号或者一组整数（如 1,2, , ）来表示，这些整数只是用于数据处理，并不代表任何特定的顺序对于两个取值均为名义变量的观测点 X_i 和 X_j，它们之间的相异度可以用简单匹配方法来计算，即：

$$d_{ij} = \frac{p - m}{p}$$

相异度——有序数据

- 假设一个变量 X_f 有 M_f 个状态，这些有序的状态定义了一个序列 1,M,M_f。
- 计算有序数据观测点的相异度的基本思想是将有序数据转换成 [0,1] 上的连续型数据，然后再利用计算连续型数据相异度方法计算。具体步骤如下：

1. 第 i 个观测点的第 f 个变量的取值为 $x_{if}, x_{if} \in \{1, 2, \cdots, M_f\}$
2. 令 $Z_{if} = \frac{x_{if}-1}{M_f-1}$ 即将每个变量的值域映射到 [0,1] 上，以便每个变量都有相同的权重
3. 对 Z_{if} 的相异度计算可以采用连续型变量所描述的任意一种距离度量方法。

K-means 聚类

- K-means 聚类本质上是需要最小化如下问题：

$$\min_{C_1,\cdots,C_K} \{\sum_{i=1}^{K} W(C_k)\}$$

- 式中的 $W(C_k)$ 表示第 C_k 类的类内差异，可以有很多种方法，如利用欧式距离：

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

- 其中 $|C_k|$ 表示第 k 类的观测数。

- 由上页两式可得 K-means 聚类的最优化问题（目标函数）：

$$\min_{C_1,\cdots,C_K}\{\sum_{i=1}^{K}\frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2\}$$

- 直接求解上式是非常困难的，因为有 K^n 种方法可以把 n 个观测样本分配到 K 个类中，当 K 和 n 如果较大时，这种穷举法的计算量是非常惊人的。
- 因此，我们需要寻找一种计算量相对小的算法，其中 K-means 算法是其中一种比较流行的算法。

K-means 聚类算法

1. 给定类数 K，为每个观测随机分配一个从 1 到 K 的数字，这些数字即表示这些观测的初始类
2. 重复以下步骤，直至类的分配完成为止:

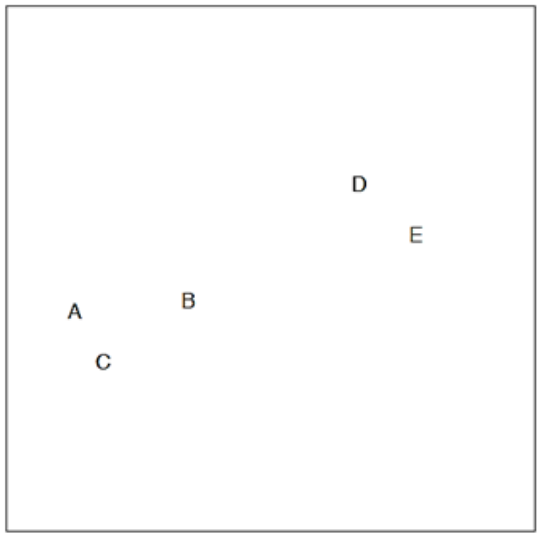
1) 分别计算 K 个类的类中心。第 k 个类的类中心是该类中所有 p 维观测向量的均值向量
2) 计算每个观测与各个类中心的相异度（距离，如欧式距离），将其重新分配到与其相异度最小的类中

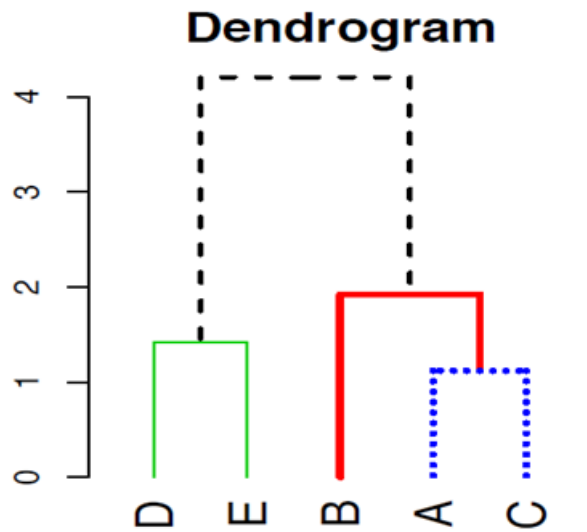
系统聚类法

- 四种常用距离形式

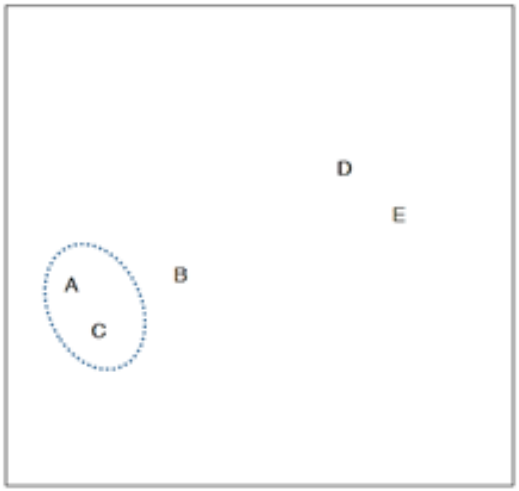
距离形式	描述
最短距离法	最小类间相异度。计算 A 类和 B 类之间的所有观测间的相异
最短距离法	最大类间相异度。计算 A 类和 B 类之间的所有观测的相异
重心法	A 类中心（长度为 p 的均值向量）和 B 类中
类平均法	平均类间相异度。计算 A 类和 B 类之前的所有观测的相异度，

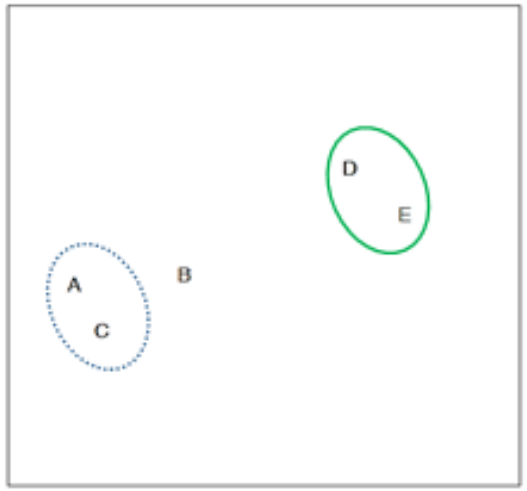
系统聚类法

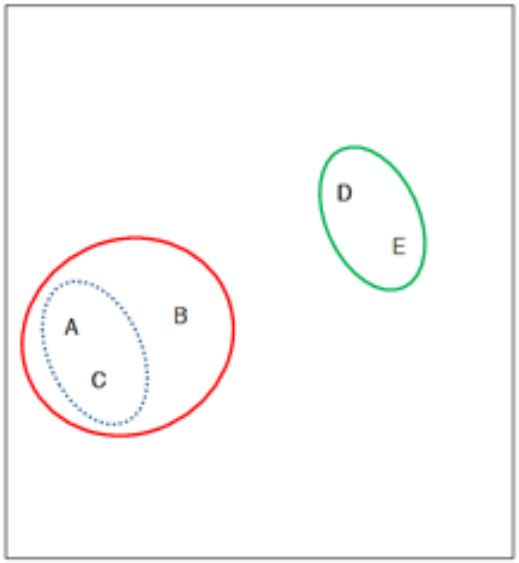


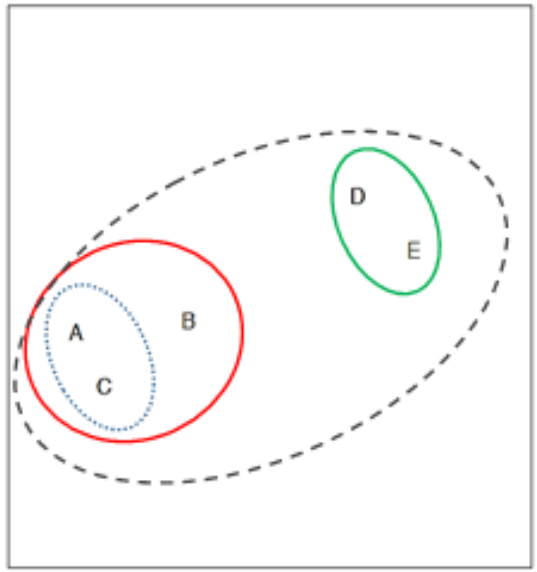


- 图左：原始数据的分布；图右：用欧式距离和最短距离法得到的谱系图









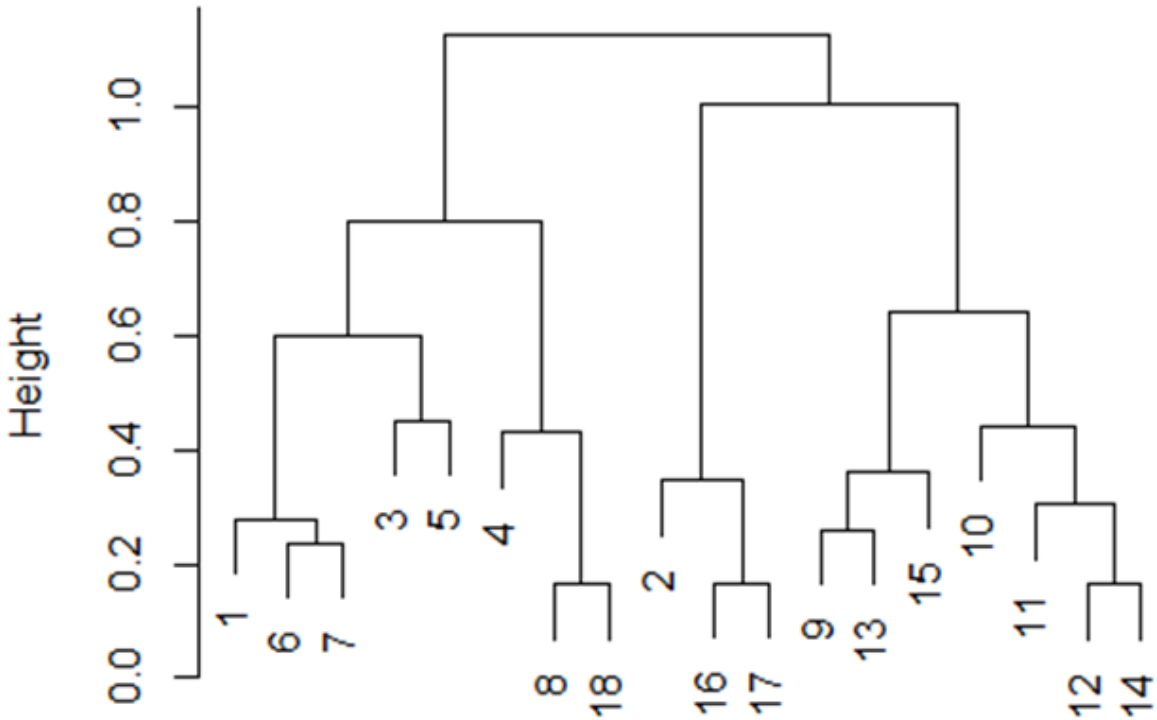
- 用欧式距离和最短距离法做系统聚类算法的图解

系统聚类算法

1. 将每个观测视为一类，共得到 n 个初始类
2. 计算 n 个观测中所有 C_n^2 对每两个观测之间的相异度（距离）
3. 对 $i = n, n-1, \cdots, 2, 1$ 重复以下步骤直至所有观测都属于一个类或者满足某个终止条件：

1). 在 i 个类中，比较任意两类间的相异度，将相异度最小的（即最相似的）那一对结合起来

2). 计算剩下的 i-1 个新类中每两个类间的相异度



- 上图给出花卉分类的谱系图。从图中可以看出，将花卉分为四类是比较合适的，即第一类为 1、6、7、3、5，第二类为 4、8、18，第三类为 2、16、17，第四类为 9、13、15、10、11、12、14

对比分析

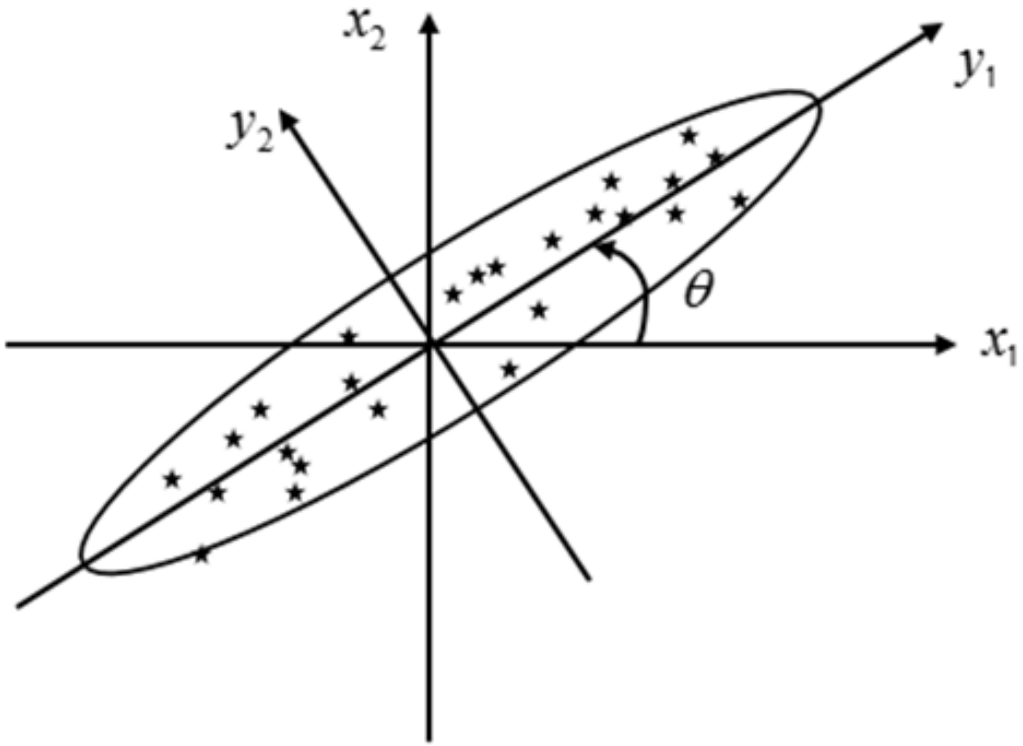
- 系统聚类法和 K 均值聚类法的算法思想都很简单，它们都是以距离的远近亲疏作为标准进行聚类的。
- K 均值聚类法只能产生指定类数的聚类结果，具体类数的确定依赖于实验的积累，而系统聚类法可直接产生一系列的聚类结果。
- 系统聚类法不具有很好的可伸缩性，它在合并类时需要检查和估算大量的对象或类，算法的复杂度为 $O(n^2)$，因此当 n 很大时并不是很适用。
- K 均值聚类法则是相对可伸缩的和高效率的，因为它的复杂度是 $O(nkt)$。
- 两种方法各有千秋，具体使用时常根据经验进行选择。有时候我们可以借助系统聚类法先将一部分样品作为对象进行聚类，再将其结果作为 K 均值聚类法确定类数的参考。

主成分分析

主成分分析

- 主成分分析的基本思想是设法将原来的指标线性组合成几个新的不相关的综合指标，同时根据实际需要从这些新的指标中提取较少的几个使其能尽可能多地反映原来的指标信息。
- 即依次选择能提取信息最多的线性组合，直至所提取的信息与原指标相差不多为止。
- 这里所说的信息是指变量的变异性，可用方差或标准差来表示。
- 当变量取单一值时，其提供的信息是非常有限的，而取一系列不同值时，我们便可从中读出最大值、最小值、平均值等信息。
- 变量的变异性越大，提供的信息就越充分，信息量也就越大。

主成分分析的几何意义



- 旋转坐标轴，旋转公式：

$$y = X_1 cos\theta + X_2 sin\theta; Y_2 = -X_1 sin\theta + X_2 cos\theta$$

数学推导

- 设 $X = (X_1, \cdots, X_p)$ 是 p 维随机向量，其均值和协方差阵分别为:$\mu = E(X), \Sigma = D(X)$, 考虑以下线性变换:

$$\begin{cases} Y_1 = t_{11}X_1 + t_{12}X_2 + \cdots + t_{1p}X_p = T_1^{'}X \\ Y_2 = t_{21}X_1 + t_{22}X_2 + \cdots + t_{2p}X_p = T_2^{'}X \\ \cdots\cdots \\ Y_p = t_{p1}X_1 + t_{p2}X_2 + \cdots + t_{pp}X_p = T_p^{'}X \end{cases}$$

- 矩阵表示为:$Y = T^{'}X$

主成分回归

- 主成分回归的主要思想是，少数的主成分足以解释大部分的数据波动和数据与因变量之间的关系。
- 也就是说，用 $Y_1, Y_2, \cdots, Y_M$ 拟合一个最小二乘模型的结果优于用 $X_1, X_2, \cdots, X_p$ 拟合的结果，因为大部分甚至全部与因变量相关的数据信息都包含在了 $Y_1, Y_2, \cdots, Y_M$ 中，估计 $M < p$ 个系数会减轻过拟合。
- 除了降维，主成分回归还有很多其他方面的作用。例如，在一个统计分析问题中，若变量之间存在多重共线性，那么统计分析的结果往往是不理想的。在这种情况下，可以通过主成分分析先提取前几个重要的主成分，再将这些主成分与因变量进行建模，这样就可以消除多重共线性的影响。
- 不止在回归问题当中，如分类和聚类中，也可以通过主成分分析提取前几个重要的主成分进行建模，而不是将所有的原始变量用于建模。
- 首先，依靠某几个重要的主成分我们已经可以得到一个噪声较小的结果，因为数据集中的主要信号（而不是噪声）通常集中在少数几个主成分中；其次，在原始数据维度较大时，这样做也起到了降

因子分析

因子分析

- 假设 $X = (X_1, X_2, \cdots, X_p)'$ 是 p 维随机向量，则在因子分析模型中，每一个变量都可以表示成公共因子的线性函数和特殊因子之和，即

$$X_i = a_{i1}F_1 + a_{i2}F_2 + \cdots + a_{im}F_m + \varepsilon_i$$

- 其中 $F_1, F_2, \cdots, F_m$ 称为公共因子，是不可直接观测但又客观存在的共同影响因素；
- ε_i 称为特殊因子，因 X_i 而异；
- 公共因子的系数 $a_{ij}, i = 1, \cdots, p; j = 1, \cdots, m$ 称为因子载荷，是第 i 个变量在第 j 个因子上的负荷。

因子分析的数学模型

- 该模型可用矩阵表示为：$X = AF + \varepsilon$
- 为了构建因子分析模型，我们需要估计因子载荷阵 A 和特殊因子的方差 σ^2，常见的方法有主成分法、主轴因子法和极大似然法。
  1. 主成分法是从原始变量的总体方差变异出发，尽可能使其能够被公因子（主成分）所解释，并且使得各公因子对原始变量的方差变异的解释比例依次降低。
  2. 主轴因子法是从原始变量的相关系数矩阵出发，使原始变量的相关程度尽可能地被公因子所解释。
  3. 极大似然法则是建立在公共因子和特殊因子服从正态分布的假设之下，如果满足这个假设条件，那么就可以得到因子载荷和特殊因子方差的极大似然估计。

典型相关分析

典型相关分析

- 一般地，设 $X^{(1)}=(X_1^{(1)},X_2^{(1)},\cdots,X_p^{(1)})$ 和 $X^{(2)}=(X_1^{(2)},X_2^{(2)},\cdots,X_p^{(2)})$ 是两组相互关联的随机向量，且 $Cov(X^{(1)},X^{(1)})=\Sigma_{11},Cov(X^{(2)},X^{(2)})=\Sigma_{22},Cov(X^{(1)},X^{(2)})=\Sigma_{12}=\Sigma_{21}^{'}$，在研究它们的相关关系时，可以采用类似于主成分分析的方法找出两组变量的一个线性组合：

$$U=a^{'}X^{(1)}=a_1X_1^{(1)}+\cdots+a_pX_p^{(1)}$$

$$V=b^{'}X^{(2)}=b_1X_1^{(2)}+\cdots+b_qX_q^{(2)}$$

- 我们希望寻找 a 和 b 使 U 和 V 之间的相关系数达到最大

典型相关步骤

- 在进行典型相关分析之前，一般也建议将数据进行标准化处理，所以，可以将典型相关分析的步骤概括为：
  1）原始数据标准化
  2）典型相关分析
  3）相关系数显著性检验

本周推荐

1. 一本书：《数字情种-埃尔德什传》，保罗霍夫曼著，上海科技教育出版社，2000.8
2. 一部电影：《逻辑的乐趣（The joy of Logic)》，BBC
3. 练习：《R 语言实战（第 2 版)》，第 14 章和 16 章代码实现

谢　谢!