raniy

微博

加好友　发纸条

写留言　加关注

主成分分析、聚类分析、因子分析的基本思想及优缺点 (2013-05-19 20:12:32)

转 载 ▼

标签： 主成分分析　因子分析　聚类分析　判别分析　对应分析/最优尺度分　分类： 机器学习

----------基本原理及优缺点----------

主成分分析：利用降维（线性变换）的思想，在损失很少信息的前提下把多个指标转化为几个综合指标（主成分），用综合指标来解释多变量的方差- 协方差结构，即每个主成分都是原始变量的线性组合，且各个主成分之间互不相关，使得主成分比原始变量具有某些更优越的性能（主成分必须保留原始变量90%以上的信息），从而达到简化系统结构，抓住问题实质的目的综合指标即为主成分。

求解主成分的方法：从协方差阵出发（协方差阵已知），从相关阵出发（相关阵R已知）。
（实际研究中，总体协方差阵与相关阵是未知的，必须通过样本数据来估计）

注意事项：1. 由协方差阵出发与由相关阵出发求解主成分所得结果不一致时，要恰当的选取某一种方法；
2. 对于度量单位或是取值范围在同量级的数据，可直接求协方差阵；对于度量单位不同的指标或是取值范围彼此差异非常大的指标，应考虑将数据标准化，再由协方差阵求主成分；
3. 主成分分析不要求数据来源于正态分布；
4. 在选取初始变量进入分析时应该特别注意原始变量是否存在多重共线性的问题（最小特征根接近于零，说明存在多重共线性问题）。

优点：首先它利用降维技术用少数几个综合变量来代替原始多个变量，这些综合变量集中了原始变量的大部分信息。其次它通过计算综合主成分函数得分，对客观经济现象进行科学评价。再次它在应用上侧重于信息贡献影响力综合评价。

缺点：当主成分的因子负荷的符号有正有负时，综合评价函数意义就不明确。命名清晰性低。

聚类分析：将个体（样品）或者对象（变量）按相似程度（距离远近）划分类别，使得同一类中的元素之间的相似性比其他类的元素的相似性更强。目的在于使类间元素的同质性最大化和类与类间元素的异质性最大化。。其主要依据是聚到同一个数据集中的样本应该彼此相似，而属于不同组的样本应该足够不相似。

常用聚类方法：系统聚类法，K-均值法，模糊聚类法，有序样品的聚类，分解法，加入法。

注意事项：1. 系统聚类法可对变量或者记录进行分类，K-均值法只能对记录进行分类；
2. K-均值法要求分析人员事先知道样品分为多少类；
3. 对变量的多元正态性，方差齐性等要求较高。

应用领域：细分市场，消费行为划分，设计抽样方案等

优点：聚类分析模型的优点就是直观，结论形式简明。

缺点：在样本量较大时，要获得聚类结论有一定困难。由于相似系数是根据被试的反映来建立反映被试间内在联系的指标，而实践中有时尽管从被试反映所得出的数据中发现他们之间有紧密的关系，但事物之间却无任何内在联系，此时，如果根据距离或相似系数得出聚类分析的结果，显然是不适当的，但是，聚类分析模型本身却无法识别这类错误。

因子分析：利用降维的思想，由研究原始变量相关矩阵内部的依赖关系出发，把一些具有错综复杂关系的变量归结为少数几个综合因子。（因子分析是主成分的推广，相对于主成分分析，更倾向于描述原始变量之间的相关关系），就是研究如何以最少的信息丢失，将众多原始变量浓缩成少数几个因子变量，以及如何使因子变量具有较强的可解释性的一种多元统计分析方法。

求解因子载荷的方法：主成分法，主轴因子法，极大似然法，最小二乘法，a因子提取法。

注意事项：5. 因子分析中各个公共因子之间不相关，特殊因子之间不相关，公共因子和特殊因子之间不相关。

应用领域：解决共线性问题，评价问卷的结构效度，寻找变量间潜在的结构，内在结构证实。

优点:第一它不是对原有变量的取舍，而是根据原始变量的信息进行重新组合，找出影响变量的共同因子，化简数据；第二，它通过旋转使得因子变量更具有可解释性，命名清晰性高。

缺点:在计算因子得分时，采用的是最小二乘法，此法有时可能会失效。

判别分析：从已知的各种分类情况中总结规律（训练出判别函数），当新样品进入时，判断其与判别函数之间的相似程度（概率最大，距离最近，离差最小等判别准则）。
常用判别方法：最大似然法，距离判别法，Fisher判别法，Bayes判别法，逐步判别法等。
注意事项：1. 判别分析的基本条件：分组类型在两组以上，解释变量必须是可测的；
2. 每个解释变量不能是其它解释变量的线性组合（比如出现多重共线性情况时，判别权重会出现问题）；
3. 各解释变量之间服从多元正态分布（不符合时，可使用Logistic回归替代），且各组解释变量的协方差矩阵相等（各组协方方差矩阵有显著差异时，判别函数不相同）。
4. 相对而言，即使判别函数违反上述适用条件，也很稳健，对结果影响不大。
应用领域：对客户进行信用预测，寻找潜在客户（是否为消费者，公司是否成功，学生是否被录用等等），临床上用于鉴别诊断。

对应分析/最优尺度分析：利用降维的思想以达到简化数据结构的目的，同时对数据表中的行与列进行处理，寻求以低维图形表示数据表中行与列之间的关系。
对应分析：用于展示变量（两个/多个分类）间的关系（变量的分类数较多时较佳）；
最优尺度分析：可同时分析多个变量间的关系，变量的类型可以是无序多分类，有序多分类或连续性变量，并对多选题的分析提供了支持。

典型相关分析：借用主成分分析降维的思想，分别对两组变量提取主成分，且使从两组变量提取的主成分之间的相关程度达到最大，而从同一组内部提取的各主成分之间互不相关。

相同点：

1.主成分分析法和因子分析法都是用少数的几个变量(因子) 来综合反映原始变量(因子) 的主要信息，变量虽然较原始变量少，但所包含的信息量却占原始信息的85 %以上，所以即使用少数的几个新变量，可信度也很高，也可以有效地解释问题。并且新的变量彼此间互不相关，消除了多重共线性。

2.这两种分析法得出的新变量，并不是原始变量筛选后剩余的变量。在主成分分析中，最终确定的新变量是原始变量的线性组合，如原始变量为x1 ，x2 ，．．．，x3 ，经过坐标变换，将原有的p个相关变量xi 作线性变换，每个主成分都是由原有p 个变量线性组合得到。在诸多主成分Zi 中，Z1 在方差中占的比重最大，说明它综合原有变量的能力最强，越往后主成分在方差中的比重也小，综合原信息的能力越弱。

因子分析是要利用少数几个公共因子去解释较多个要观测变量中存在的复杂关系，它不是对原始变量的重新组合，而是对原始变量进行分解，分解为公共因子与特殊因子两部分。公共因子是由所有变量共同具有的少数几个因子；特殊因子是每个原始变量独自具有的因子。

3.对新产生的主成分变量及因子变量计算其得分，就可以将主成分得分或因子得分代替原始变量进行进一步的分析，因为主成分变量及因子变量比原始变量少了许多，所以起到了降维的作用，为我们处理数据降低了难度。
4.聚类分析是把研究对象视作多维空间中的许多点，并合理地分成若干类，因此它是一种根据变量域之间的相似性而逐步归群成类的方法，它能客观地反映这些变量或区域之间的内在组合关系。它是通过一个大的对称矩阵来探索相关关系的一种数学分析方法，是多元统计分析方法，分析的结果为群集。对向量聚类后，我们对数据的处理难度也自然降低，所以从某种意义上说，聚类分析也起到了降维的作用。

不同之处：

1.主成分分析是研究如何通过少数几个主成分来解释多变量的方差一协方差结构的分析方法，也就是求出少数几个主成分(变量) ，使它们尽可能多地保留原始变量的信息，且彼此不相关。它是一种数学变换方法，即把给定的一组变量通过线性变换，转换为一组不相关的变量(两两相关系数为0 ，或样本向量彼此相互垂直的随机变量) ，在这种变换中，保持变量的总方差(方差之和) 不变，同时具有最大方差，称为第一主成分；具有次大方差，称为第二主成分。依次类推。若共有p 个变量，实际应用中一般不是找p 个主成分，而是找出m (m < p) 个主成分就够了，只要这m 个主成分能反映原来所有变量的绝大部分的方差。主成分分析可以作为因子分析的一种方法出现。
2.因子分析是寻找潜在的起支配作用的因子模型的方法。因子分析是根据相关性大小把变量分组，使得同组内的变量之间相关性较高，但不同的组的变量相关性较低，每组变量代表一个基本结构，这个基本结构称为公共因子。对于所研究的问题就可试图用最少个数的不可测的所谓公共因子的线性函数与特殊因子之和来描述原来观测的每一分量。通过因子分析得来的新变量是对每个原始变量进行内部剖析。因子分析不是对原始变量的重新组合，而是对原始变量进行分解，分解为公共因子和特殊因子两部分。具体地说，就是要找出某个问题中可直接测量的具有一定相关性的诸指标，如何受少数几个在专业中有意义、又不可直接测量到、且相对独立的因子支配的规律，从而可用各指标的测定来间接确定各因子的状态。因子分析只能解释部分变异，主成分分析能解释所有变异。

3.聚类分析算法是给定m 维空间R 中的n 个向量，把每个向量归属到k 个聚类中的某一个，使得每一个向量与其聚类中心的距离最小。聚类可以理解为：类内的相关性尽量大，类间相关性尽量小。聚类问题作为一种无指导的学习问题，目的在于通过把原来的对象集合分成相似的组或簇，来获得某种内在的数据规律。
从三类分析的基本思想可以看出，聚类分析中并没于产生新变量，但是主成分分析和因子分析都产生了新变量。
就数据标准化来说，区别如下：

1.主成分分析中为了消除量纲和数量级，通常需要将原始数据进行标准化，将其转化为均值为0方差为1 的无量纲数据。

2.因子分析在这方面要求不是太高，因为在因子分析中可以通过主因子法、加权最小二乘法、不加权最小二乘法、重心法等很多解法来求因子变量，并且因子变量是每一个变量的内部影响变量，它的求解与原始变量是否同量纲关系并不太大，当然在采用主成分法求因子变量时，仍需标准化。