作者：李军利 AI-Friend / 10th Jan. 2021

内容：主成分分析；因子分析

主成分分析(Principal Component Analysis) 、因子分析(Factor Analysis) 是常用但易混淆的`降维`方法

在很多业务场景里，变量/特征 多而且相关性强，增加了数据分析的复杂性：如果单独分析每个变量，分析结果是孤立的，而不是综合的；如果减少变量，则会损失很多信息，容易产生错误的结论。因此需要使用合理的方法，在减少特征维度的同时，尽可能的保留原特征包含的信息，以达到对数据进行全面分析的目的。由于变量间的相关性，我们有可能用较少的综合指标分别代表变量中的各类信息，主成分分析与因子分析就是这样的降维方法。

PCA ——数据的压缩：

提炼反映某种事物或现象的综合指标，即主成分，并且有可能给主成分所包含的信息以 适当的解释（每个原始变量在主成分中都占有一定的分量，很可能无法明确表述哪个主成分代表哪些原始变量，不能清晰的解释主成分的含义，而这正是 FA 的优点）。实际应用中，PCA 常用于**探索性分析**，大致了解数据，作为中间手段，而不是单独作为一种完全的分析方法。

FA ——公共因子与特殊因子：

FA 在提取公因子时，不仅考虑变量间是否相关，而且考虑了相关关系的强弱，这就使得提取的公因子不仅能起到降维的作用，而且能够被很好的解释。FA 避免 PCA 解释障碍的方法是通过 `因子轴旋转`，因子轴旋转可以使原始变量在公因子（主成分）上的载荷重新分布，从而使原始变量在公因子上的载荷两级分化，这样公因子（主成分）就能用那些载荷大的原始变量来解释。FA 与 PCA 是包含与扩展的关系。

PCA 和 FA 是 **降维和信息浓缩** 的方法， 生成的新变量均代表了原始变量的大部分信息且主成分/因子之间互相独立，主成分或者因子都可以用于后续的回归、判别、聚类等任务。

对比

	主成分分析	因子分析
发明	$pearson$	$spearman$
表示	$主成分 = 原始变量的线性组合$	$原始变量 = 公因子的线性组合$
假设	无假设	所有因子之间不相关
解释	解释原始变量的 `总方差`，强调新变量贡献了多大比例的方差，而不关心是否有明确的实际/解释 意义	解释原始变量的 `协方差`，需要构造因子模型，着重要求新变量具有`实际意义`，能解释原始变量间的内在结构
数量	$主成分个数 = 原始变量个数$	$因子个数\ = 根据数据和业务情况，人为指定$
唯一	协方差阵或相关阵的特征值唯一时，主成分也唯一	因子不唯一，并且通过旋转可以得到不同因子
用途	数据处理，降维，探索变量关系。PCA 应用广，侧重信息贡献	降维，解释，是 PCA 的推广
原理	用降维(线性变换)思想，在损失很少信息前提下，把多个指标转化为几个独立的综合指标，即每个主成分都是原始变量的**线性组合**，且互不相关，能够简化系统结构，抓住问题实质	从研究原始变量`相关矩阵内部的依赖关系`出发，把变量表示成少数的公共因子和仅对某一个变量有作用的`特殊因子`的线性组合，从数据中提取可解释的公共因子(相对 PCA，更倾向描述原始变量之间的相关关系)
求解	从协方差阵(协方差阵已知)或者相关阵出发(相关阵已知)，只能采用主成分法。(总体协方差阵与相关阵未知，须通过样本估计总体)	主成分法，主轴因子法，极大似然法，最小二乘法，阿尔法因子提取法
算法	协方差矩阵的对角元素是变量的方差	协方差阵的对角元素不再是变量的方差，而是对应变量的`共同度`（变量方差中被各因子解释的部分）

	主成分分析	因子分析
组合	$PCA+\text{判别分析}$:适用变量多样本不多时;$PCA+\text{多元回归}$:主成分独立,替代具有多重共线性的原始变量作为多元回归的变量	$FA+\text{多元回归}$:解决共线性,寻找变量间的潜在结构$FA+\text{聚类}$:借助FA寻找聚类变量;

主成分分析

通过 正交变换 将一组存在相关性的变量转换为一组线性不相关的变量,转换后的这组变量叫主成分。

不同的 `线性变换`,得到的统计特性不同,我们希望主成分之间相互独立,同时方差尽可能大,因此,只需要通过协方差矩阵或者相关稀疏矩阵 求特征值 λ_i 及特征向量 u_i,即可构成主成分分析的解。第一个主成分的方差最大,其贡献率等于其方差在全部主成分方差中的占比。

选取主成分的标准

- 保留的主成分使得方差贡献率达到80%以上
- 保留的主成分的方差(特征值)大于1
- 可借助碎石图,保留图中转折较大的主成分

分析步骤

- 分析前,先进行相关性检验,变量之间存在较强相关性,才能使用 PCA
- 将初始数据标准化,统一量纲

  当变量单位相同或者变量在同一数量等级的情况下,可以直接采用协方差阵进行计算;

  当度量单位不同或者变量的方差 差别很大时,应考虑先 `数据标准化`,再由协方差阵求主成分,

  因为PCA 对初始变量的方差非常敏感,方差大的变量占比相对大,这将导致主成分的偏差。

  在数学上,可以通过减去平均值并除以每个变量的标准偏差来防止这个问题。

  不过 标准化确实会抹杀一部分原本刻画变量之间离散程度差异的信息
- 选取初始变量
- 根据初始变量特性选择使用协方差矩阵还是 `相关矩阵` 来求主成分

  注意:从协方差阵出发和从相关阵出发,求解主成分的结果不一致时,要恰当的选取某一种方法

  协方差为正,两个变量同时增加或减少,即正相关,反正则负相关
- 计算协方差矩阵或 `相关矩阵` 的特征值和特征向量
- 确定主成分个数
- 对主成分做经济解释,主成分的经济意义由各线性组合中权重较大的几个指标来确定

  将原数据分别按第一,第二,第三主成分得分排序,观察各变量主要受哪个主成分影响

优点

- 不要求数据服从正态分布

  主成分就是按数据离散程度最大的方向对基组进行旋转,这特性扩展了其应用范围
- 通过对原始变量进行综合与简化,可以客观地确定各个指标的权重,避免主观判断的随意性

缺点

- 若原始数据相关性弱,降维作用不好

- 降维后，存在少量信息丢失，不可能包含100%原始数据
- 原始数据经过标准化处理之后，含义会发生变化，且主成分的解释含义较原始数据比较模糊
- 假设标准化后的原始变量间存在多重共线性，即原始变量之间存在不可忽视的信息重叠，

  主成分分析不能有效剔除信息重叠。注意区分相关性与多重共线性

因子分析

通过研究`众多变量之间的内部依赖关系`，探求数据的基本结构，并用少数几个假想变量（因子）表示原始数据

主成分分析相当于把原始数据的主要成分给拿了出来，而因子分析是从假设出发，假设所有的自变量 x 出现的原因是背后存在一个潜在变量 f，即因子，在这个因子的作用下，x 才表现出来/可以被观察到。比如一个学生的数理化成绩很好，那么我们认为这个学生理性思维较强，理性思维就是一个因子，在这个因子的作用下，偏理科的成绩才会那么高。因子分析最早正是由心理学家提出的。

因子分析分为探索性因子分析和验证性因子分析。探索性因子分析是不确定一堆自变量背后有几个因子，我们通过这种方法试图寻找到这几个因子。而验证性因子分析是已经假设自变量背后有若干个因子，试图通过这种方法去验证假设是否正确，相比较而言，探索性因子分析应用比较多，下面讨论探索性因子分析。

核心问题 检验是否适合因子分析; 如何构造因子变量; 如何对因子变量进行命名解释

因子特点

- 因子能反映众多原始变量的主要信息
- 因子个数远远少于原始变量个数
- 因子并非原始变量的简单取舍，而是一种新的综合
- 因子之间没有线性关系
- 因子具有明确解释性，可以最大限度地发挥专业分析的作用

步骤

- 相关性检验，选择分析变量

  用定性分析和定量分析的方法选择变量，因子分析的前提条件是变量间有较强的相关性，如果变量之间无相关性或相关性较小的话，它们不会有共享因子，一般来说，相关性小于0.3 就不适合因子分析

- 提取公因子

  取方差（特征值）大于0的因子；因子的累积方差贡献率达到80%

  这一步要确定因子求解的方法和因子的个数。需要根据研究者的设计方案或有关的经验或知识事先确定。因子个数的确定可以根据因子方差的大小。只取方差大于1(或特 征值大于1)的那些因子，因为方差小于1的因子其贡献可能很小；按照因子的累计方差 贡献率来确定，一般认为要达到60%才能符合要求。数矩阵是估计因子结构的基础。

- 因子旋转

  通过坐标变换使每个原始变量在尽可能少的因子之间有密切的关系，这样因子的实际意义更容易解释，可以为每个潜在因子赋予有实际意义的名字

- 计算因子得分

  求出各样本的因子得分，有了因子得分值，则可以在许多分析中使用这些因子，例 如以因子的得分做聚类分析的变量，做回归分析中的回归因子。如果有预期想提取的因子个数，可以主动设置输出的因子个数

Reference

[1] PCA详解

[2] 主成分原理

[3] 主成分与因子分析

[4] https://zhuanlan.zhihu.com/p/77151308

[5] 因子旋转

推荐阅读

[6] 刘建平/PCA思想

[7] 刘建平/sklearn实现PCA

[8] FA理论推导

[9] factor_analyzer 实现FA

[10] 因子分析(factor analysis)例子–Python