三大相关系数与MIC

Li Junli 李军利
Apr 4 2019

三个相关性系数(Pearson、Spearman 和Kendall) 反映的都是两个变量之间变化趋势的方向以及程度，其值范围为-1到+1，0表示两个变量不相关，正值表示正相关，负值表示负相关，值越大表示相关性越强。 计算积距pearson相关系数，连续性变量才可采用;计算Spearman秩相关系数，适合于定序变量或不满足正态分布假设的等间隔数据; 计算Kendall秩相关系数，适合于定序变量或不满足正态分布假设的等间隔数据。当资料不服从双变量正态分布或总体分布未知，或原始数据用等级表示时，宜用 spearman或kendall相关。

极强相关(0.8-1.0); 强相关(0.6-0.8); 中等程度相关(0.4-0.6); 弱相关(0.2-0.4); 极弱相关或无相关(0.0-0.2)。

最大信息系数--MIC(Maximal information coefficient)既可以衡量非线性关系，又可以衡量非线性关系。相较于互信息--Mutual Information(MI)而言有更高的准确度。

Pearson相关系数

皮尔森相关系数(Pearson correlation coefficient) 反映两个变量线性相关程度的统计量。

当对每个变量进行0均值后，相关性就与**余弦距离**相同。

适用条件：

1. 两个变量都是连续变量

2. 每个变量都应该是 正态分布，或者接近正态分布的单峰对称分布

3. 变量之间应该为线性关系

常用希腊小写字母 ρ 作为总体相关系数的代表符号，X和Y的总体相关系数：

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X\sigma_Y} = \frac{E[(X-u_X)(Y-u_Y)]}{\sigma_X\sigma_Y} = \frac{E(X,Y)-E(X)E(Y)}{\sqrt{E(X^2)-E^2(X)}\sqrt{E(Y^2)-E^2(Y)}}$$

估算样本的协方差和标准差，可得到样本相关系数(样本皮尔逊系数)，常用英文小写字母 r 代表：

$$r = \frac{\sum_{i=1}^n (X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^n (X_i-\overline{X})^2}\sqrt{\sum_{i=1}^n (Y_i-\overline{Y})^2}}$$

r 亦可由 样本点的标准分数均值估计，得到与上式等价的表达式：

$$样本标准分数：\frac{X_i-\overline{X}}{\sigma_X}，样本均值：\overline{X}，样本标准差：\sigma_X$$

$$r = \frac{1}{n-1}\sum_{i=1}^{n}(\frac{X_i - \overline{X}}{\sigma_X}\frac{Y_i - \overline{Y}}{\sigma_Y})$$

利用样本相关系数推断总体中两个变量是否相关，可以用t 统计量对总体相关系数为0的原假设进行检验。若t 检验显著，则拒绝原假设，即两个变量是线性相关的；若t 检验不显著，则不能拒绝原假设，即两个变量不是线性相关的。Pearson只能度量线性相关，相关系数小不能得出两个变量不相关的结论，因为可能存在非线性相关。

Spearman相关系数

曾经在一个关于排名的比赛中使用过Spearman相关系数作为衡量算法效果的标准,因为Spearman有这样的特性——如果真实排名是50,那么预测排名为47或者55,最后的相关系数仍然比较大,但是如果预测为100,即使其他的预测比较准,但是这一个错误的预测就会大大减少相关系数。

Spearman(斯皮尔曼等级相关)是根据等级资料研究两个变量间相关关系的方法。它是依据两列成对等级的各对等级数之差来进行计算的，所以又称为"等级差数法"。斯皮尔曼等级相关对数据条件的要求没有Pearson相关系数严格，只要两个变量的观测值是成对的等级评定资料，或者是由连续变量观测资料转化得到的等级资料，不论两个变量的总体分布形态、样本容量的大小如何，都可以用斯皮尔曼等级相关来进行研究。

总而言之，斯皮尔曼相关的计算将原始数据替代为 数据在该序列中的位置。

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

以有一个5人的视听反应时间样本，用Spearman衡量听觉和视觉时间反应是否具有一致性：

id	听觉	视觉	秩序	秩序	d
1	170	180	3	4	-1
2	150	165	1	1	0
3	210	190	5	5	0
4	180	168	4	2	2
5	160	172	2	3	-1

Kendall相关系数

肯德尔相关系数是一个用来测量两个随机变量相关性的统计值，用于反映分类变量相关性的指标，适用于两个分类变量均为有序分类的情况，连续的随机变量可以排序后用肯德尔相关系数。一个肯德尔检验是一个无参数假设检验，它使用计算而得的相关系数去检验两个随机变量的统计依赖性。肯德尔相关系数的取值范围在-1到1之间。n个同类的统计对象按特定属性排序，其他属性通常是乱序的。同序对（concordant pairs）和异序对（discordant pairs）之差与总对数（n*(n-1)/2)的比值定义为Kendall(肯德尔)系数。

假设两个随机变量分别为X、Y（也可以看做两个集合），它们的元素个数均为N，两个随即变量取的第i个值分别用Xi、Yi表示。X与Y中的对应元素组成一个元素对集合XY，其包含的元素为(Xi, Yi)（1<=i<=N）。

1) 当集合XY中任意两个元素(Xi, Yi)与(Xj, Yj)的排行相同时(也就是说当出现情况1或2时；情况1：Xi>Xj且Yi>Yj，情况2：Xi<Xj且Yi<Yj) 这两个元素就是一个同序对；

2）当出现情况3或4时(情况3：Xi>Xj且Yi<Yj，情况4：Xi<Xj且Yi>Yj）这两个元素就是一个异序对；

3）当出现情况5或6时(情况5：Xi=Xj，情况6：Yi=Yj)，这两个元素既不是一致的也不是不一致的。

以下表为例计算同序对，异序对和Kendall相关系数：

Person	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**
Rank by Height	1	2	3	4	5	6	7	8
Rank by Weight	3	4	1	2	5	7	8	6

AB和BA是相同对，只计一次，AA不算，则总对数 = 7 + 6 + 5 + 4 + 3 + 2 + 1 + 0 = 28

表中以Height为基准排序，Weight乱序。A最高，但体重排名为 3 ，贡献5个同序对，即AB，AE，AF，AG，AH。同理，我们发现B、C、D、E、F、G、H分别贡献4、5、4、3、1、0、0个同序对，因此，同序对总数 P = 5 + 4 + 5 + 4 + 3 + 1 + 0 + 0 = 22 ，则异序对 = 28 - 22 = 6。

集合中元素唯一，即不存在并列排名时，Kendall相关系数有公式：

$$R = \frac{\text{同序对数} - \text{异序对数}}{\text{总对数}} = \frac{\text{同序对数} - \text{异序对数}}{\frac{1}{2}N(N-1)}$$

由公式得 R = (22 - 6)/28 = 0.57。

当集合中的元素有重复即有并列排名时，计算复杂一些，仍然以集合X，Y为例：

$$R = \frac{\text{同序对} - \text{异序对}}{\sqrt{(N_3 - N_1)(N_3 - N_2)}}$$

$$N_3 = \frac{1}{2}N(N-1),\ N_1 = \sum_{i=1}^{s}\frac{1}{2}U_i(U_i - 1),\ N_1 = \sum_{i=1}^{s}\frac{1}{2}V_i(V_i - 1)$$

N1、N2分别是针对集合X、Y计算的，以计算N1为例，给出N1的由来（N2的计算可以类推）：

将X中的相同元素分别组合成小集合，s表示集合X中拥有的小集合数（例如X包含元素：1 2 3 4 3 3 2，那么这里得到的s则为2，因为只有2、3有相同元素），Ui表示第i个小集合所包含的元素数。

MIC(最大信息系数)

曾经在一个机器学习项目中用相关系数当做特征选择的辅助标准，但是Pearson适用于度量线性相关，而Speearman和Kendall都是用于度量等级相关的秩相关系数。项目中特征与变量间存在着比较复杂的非线性关系，不适合用三大相关系数，因此学习了MIC。2011年Science的一篇文章(参考文献)最早讲到了MIC，阅读英文文献会比直接看别人的总结 理解更深入。

MIC--最大信息系数，也叫做最大互信息系数，从名字就知道MIC是互信息的扩展，决策树ID3算法衡量两信息的相似度用到的就是互信息，具体不做阐述了，感兴趣读者可以参考互信息计算。

MIC的特性和优缺点：

MIC度量具有普适性。其不仅可以发现变量间的线性函数关系，还能发现非线性函数关系(指数的，周期的)；

不仅能发现函数关系，还能发现非函数关系(比如函数关系的叠加，或者有趣的图形模式)。

MIC度量具有均衡性。对于相同噪声水平的函数关系或者非函数关系，MIC度量具有近似的值。所以MIC度量

不仅可以用来纵向比较同一相关关系的强度，还可以用来横向比较不同关系的强度。

优点：拥有足够的统计样本时，可以捕获广泛的关系，而不限定于特定的函数(如线性、指数型、周期型等); 对不同类型的噪声程度同等的关系给予相近的分数。

缺点: MIC的统计能力遭到了一些质疑，当零假设不成立时，MIC的统计就会受到影响。在有的数据集上不存在这个问题，但有的数据集上就存在这个问题

MIC具体计算过程比较复杂，公式较多，可参考MIC计算步骤详解，Python minepy 包中的MINE模块提供MIC的计算，之前用的时候好像只支持32位的python，我在64位测试了下可以用，新版本应该是改进了。

参考文献与链接

David N. Reshef, et al.**Detecting Novel Associations in Large Data Sets**.Science 2011

百度百科

三大相关系数适用范围

相似度计算之kendall秩相关系数

声明

改论文闲暇之余，对相关系数做了简单介绍。本笔记为个人整理，仅限学习使用，转载请标明作者和来源。码字不易，如果觉得不错，git上请点个***star***吧，个人GItHub地址，谢谢配合。