大数据分析

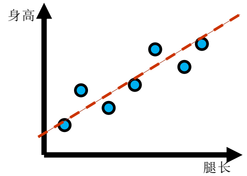
Statistics
Correlations

刘盛华

相关性分析 ——什么是相关

问题： 如果在一座古墓中发现一根腿骨，那么通过这根腿骨的长度可以判断墓主的身高吗？

条件： 提供348个成年男子的身高及腿长数据。



回答： 观察发现，身高会随着腿长的变化而变化，利用身高和腿长的关系可以解决此问题。

思考： 上述提到的关系就叫做相关关系

相关性分析 ——什么是相关

变量 ⟷ 变量

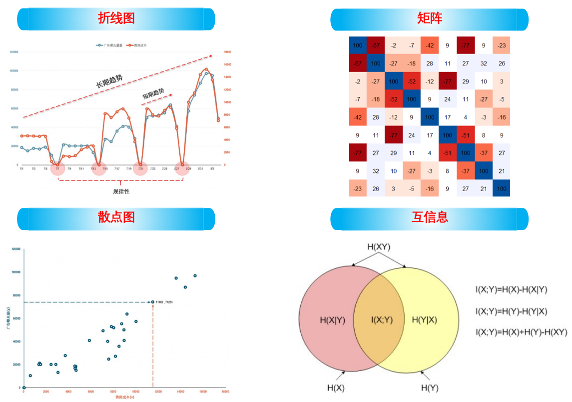
关系

一个变量随着另一个变量变化。 **相关关系**

一个变量随着另一个变量增加而增加。 **正相关**

一个变量随着另一个变量增加而减少。 **负相关**

相关性分析 ——相关的刻画方法



相关性分析 ——相关性的量化

- **传统统计相关性分析**
  - **皮尔森相关系数** PEARSON CORRELATION COEFFICIENT

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

对测量值的好坏要求比较高

形式一样
一个关注值
一个关注序

  - **斯皮尔曼相关系数** Spearman's rank correlation coefficient

$$\rho = \frac{\sum_{i=1}^{n}(r_i - \bar{r})(s_i - \bar{s})}{\sqrt{\sum_{i=1}^{n}(r_i - \bar{r})^2}\sqrt{\sum_{i=1}^{n}(s_i - \bar{s})^2}}$$

其中，r_i和s_i分别是X_i和Y_i的秩 ← 序

因为用了序，异常值影响小，使用范围广

相关性分析 ——相关性的量化

- **传统统计相关性分析**
  - **肯德尔相关系数** Kendall correlation coefficient
    - 计算的变量可以是分类变量
    - Pearson和Spearman必须是有序变量

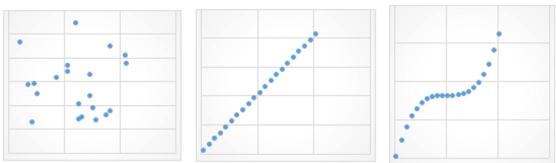
$$\tau = \frac{2}{n(n-1)} \sum_{i<j} \operatorname{sgn}(x_i - x_j)\operatorname{sgn}(y_i - y_j)$$

一致的序对个数 － 相反序对个数

（x1，y1），（x2，y2），...，（xn，yn）分别是联合随机变量*X*和*Y*的一组观察值，
(xi, yi)和 (xj, yj) 其中i<j

相关性分析 ——相关性的量化

- **对比**



(a)	(b)	(c)
皮尔森相关系数：-0.0189	皮尔森相关系数：1.0000	皮尔森相关系数：0.9179
斯皮尔曼相关系数：-0.0208	斯皮尔曼相关系数：1.0000	斯皮尔曼相关系数：1.0000
肯德尔相关系数：-0.0095	肯德尔相关系数：1.0000	肯德尔相关系数：1.0000

相关性分析 ——新的挑战

- **传统统计相关性分析**
  - PEARSON相关系数
  - Spearman相关系数
  - KENDALL相关系数

非线性？
高维性？
海量性？

- **大数据中的统计相关性分析**
  - 基于互信息的相关系数
  - 基于协方差矩阵的相关系数
  - 基于距离的相关系数

大数据中的统计相关性分析

- **互信息**
  - **正式地，两个离散随机变量 X 和 Y 的互信息可以定义为**

$$I(X;Y)=\sum_{y\in Y}\sum_{x\in X}p(x,y)\log\left(\frac{p(x,y)}{p(x)\,p(y)}\right)$$

    - 其中 p(x,y) 是 X 和 Y 的联合概率分布函数，而p(x)和p(y)分别是 X 和 Y 的边缘概率分布函数。
  - 在连续随机变量的情形下，求和被替换成了二重定积分

$$I(X;Y)=\int_Y\int_X p(x,y)\log\left(\frac{p(x,y)}{p(x)\,p(y)}\right)\,dx\,dy,$$

    - 其中 p(x,y) 当前是 X 和 Y 的联合概率密度函数，而p(x)和p(y)分别是 X 和 Y 的边缘概率密度函数。

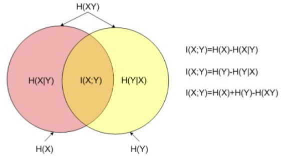
大数据中的统计相关性分析

- **互信息与其他量的等价关系**

$$\begin{aligned}I(X;Y)&=H(X)-H(X|Y)\\&=H(Y)-H(Y|X)\\&=H(X)+H(Y)-H(X,Y)\\&=H(X,Y)-H(X|Y)-H(Y|X)\end{aligned}$$

信息熵的增益

  - 其中H(X)和H(Y) 是边缘熵，H(X|Y)和H(Y|X)是条件熵，而H(X,Y)是X和Y的联合熵。注意到这组关系和并集、差集和交集的关系类似，用Venn图表示：



大数据中的统计相关性分析

- **基于矩阵计算的相关系数**
  - **RV系数（ROBERT P, 1976）**

两个矩阵A和B的协方差定义为$tr(AA'\ BB')$，方差分别为$tr(AA')^2$，$tr(BB')^2$（其中$tr(\cdot)$是矩阵的迹，定义为矩阵主对角线元素的和）。鉴于上述定义，RV 系数以皮尔森相关系数的方式重新构造，即得

$$RV(A,B)=\frac{tr(AA'\ BB')}{\sqrt{tr(AA')^2\,tr(BB')^2}}$$

RV是测量A和B协方差矩阵紧密程度的测度, 取值范围为[0,1]. 当RV约接近1, 说明用A（B）代替B（A）越合理

相关性分析 ——相关性的量化

- **大数据中的统计相关性分析**
  - **基于距离的相关系数（两个不同维度向量X,Y）**

对于实数向量$s=(s_1,s_2,\cdots,s_p)\in R^p$，它的欧氏范数为$\|s\|=(s_1^2+s_2^2+\cdots+s_p^2)^{1/2}$. 进一步定义$\langle s,X\rangle=s_1X_1+s_2X_2+\cdots+s_pX_p$为$s$与$X$的内积。同理，可以$i$定义$t=(t_1,t_2,\cdots,t_q)\in R^q$，$\|t\|$，$\langle t,Y\rangle$。在此基础上，随机向量$(X,Y)$的联合特征函数定义为

$$f_{XY}(s,t)=Eexp[i\langle s,X\rangle+i\langle t,Y\rangle],$$

其中，i为虚数单位，X,Y各自的特征函数为

$$f_X(s)=f_{XY}(s,0)=Eexp[i\langle s,X\rangle],$$

$$f_Y(t)=f_{XY}(0,t)=Eexp[i\langle t,Y\rangle].$$

相关性分析 ——相关性的量化

- 大数据中的统计相关性分析
  - 基于距离的相关系数（两个不同维度向量X,Y）

定义随机向量X与Y的距离协方差$V(X,Y)$，方差$V^2(X)$，$V^2(Y)$. 公式如下：

$$V^2(X,Y) = \|f_{XY}(s,t) - f_X(s)f_Y(t)\|_\omega^2$$
$$= \int_{R^{p+q}} |f_{XY}(s,t) - f_X(s)f_Y(t)|^2 \omega(s,t)dsdt$$
$$V^2(X) = V^2(X,X) = \|f_{XX}(s,t) - f_X(s)f_X(t)\|_\omega^2$$
$$V^2(Y) = V^2(Y,Y) = \|f_{YY}(s,t) - f_Y(s)f_Y(t)\|_\omega^2$$

其中，$\omega(s,t)$是权重函数，它的选择需要满足三个条件，即保证被积函数可积性；X与Y独立时，相关系数为零；X与Y同比例变化时，相关系数不变。在此基础上定义距离相关系数

$$R^2(X,Y) = \begin{cases} \dfrac{V^2(X,Y)}{\sqrt{V^2(X)V^2(Y)}}, & V^2(X)V^2(Y) > 0 \\ 0 & V^2(X)V^2(Y) = 0 \end{cases}$$

此方法可以度量非线性相关性，且适用于度量任意两个不同维数的随机向量

THANK YOU