Q2 实验报告

1854116 朱明志

目录

1 数据描述

本实验的数据集来源于 R 语言包 NetworkRiskMeasures 模拟的结果。数据集有 125 行，每一行包含 4 个属性，每一条数据代表一家银行的 4 种数据，即资产、负债率、准备金和权值。整个数据集的分布符合长尾分布，生成随机数的随机时间种子是 1100，具体数据已附在 data 子文件夹里。

2 代码运行结果屏幕拷贝

2.1 距离和相似度计算结果

```
[32] print('Euclidean Distances:', pdist
     ([npy[0], npy[1]]),'')...

     Euclidean Distances: [14.5229349]
     Standardized Euclidean Distances: [2.82
     842712]
     Manhattan Distances: [23.10850619]
     Chebyshev Distances: [11.11507591]
     Canberra Distances: [1.9521123]

     Cosine Similarity: [0.91742122]
     Dice Similarity: [1.]
     Jaccard-Needham Similarity: [0.]
     Kulsinski Similarity: [1.]
     Sokal-Michener Similarity: [1.]
```

2.2 相关系数和基于信息论的信息度量计算结果

```
Pearson Correlation: -0.008496238018387444
Spearman Correlation: 0.004958525345622119
Mutual Information: 4.828313737302302
```

3 实验内容讨论

3.1 距离和相似度

3.1.1 实验原理

- 欧几里得距离（Euclidean Distance）

在数学中，欧几里得距离或欧几里得度量是欧几里得空间中两点间“普通”（即直线）距离。使用这个距离，欧氏空间成为度量空间。相关联的范数称为欧几里得范数。

公式：$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$

- 标准化的欧式距离（Standardized Euclidean distance）

现将各个维度的数据进行标准化：标准化后的值 = (标准化前的值 － 分量的均值) /分量的标准差，然后计算欧式距离。

公式: $\sqrt{\sum_{k=1}^{n}\left(\frac{x_{1k}-x_{2k}}{s_k}\right)^2}$

- 曼哈顿距离（Manhattan Distance）

曼哈顿距离（Manhattan Distance）是由十九世纪的赫尔曼·闵可夫斯基所创词汇，是种使用在几何度量空间的几何学用语，用以标明两个点在标准坐标系上的绝对轴距总和。曼哈顿距离的命名原因是从规划为方型建筑区块的城市（如曼哈顿）间，最短的行车路径而来（忽略曼哈顿的单向车道以及只存在于3、14 大道的斜向车道）。任何往东三区块、往北六区块的的路径一定最少要走九区块，没有其他捷径。

公式:: $\sum_{k=1}^{n}|x_{1k}-x_{2k}|$

- 切比雪夫距离（Chebyshev Distance）

切比雪夫距离（Chebyshev distance）是向量空间中的一种度量，二个点之间的距离定义是其各坐标数值差绝对值的最大值。以数学的观点来看，切比雪夫距离是由一致范数（uniform norm）（或称为上确界范数）所衍生的度量，也是超凸度量（injective metric space）的一种。

公式: $\lim_{p\to\infty}\left(\sum_{i=1}^{n}|x_i-y_i|^p\right)^{\frac{1}{p}}=max\,|x_i-y_i|$

- 坎贝拉距离（Canberra distance）

Canberra distance 是用来衡量两个向量空间的居间，1966 年被提出，1977 年被 G. N. Lance 和 W. T. Williams 重新提出。是 Manhattan distance 的加权版本，Canberra distance 已被用作比较排名列表和计算机安全中的入侵检测的测量。

公式: $\sum_{i=1}^{n}\frac{|x_i-y_i|}{|x_i|+|y_i|}$

- 余弦相似度（Cosine Similarity）

余弦相似性通过测量两个向量的夹角的余弦值来度量它们之间的相似性。0 度角的余弦值是 1，而其他任何角度的余弦值都不大于 1；并且其最小值是-1。从而两个向量之间的角度的余弦值确定两个向量是否大致指向相同的方向。两个向量有相同的指向时，余弦相似度的值为 1；两个向量夹角为 90°时，余弦相似度的值为 0；两个向量指向完全相反的方向时，余弦相似度的值为-1。这结果

是与向量的长度无关的，仅仅与向量的指向方向相关。余弦相似度通常用于正空间，因此给出的值为 0 到 1 之间。

公式: $\cos(\theta) = \frac{X \cdot Y}{|X||Y|}$

- Dice 相似度（Dice similarity）

Dice 系数是一种集合相似度度量函数，通常用于计算两个样本的相似度

公式: $\frac{2|X \cap Y|}{|X|+|Y|}$

- 杰卡德相似度（Jaccard-Needham similarity）

两个集合 A 和 B 的交集元素在 A，B 的并集中所占的比例，称为两个集合的杰卡德相似系数。

公式: $\frac{|X \cap Y|}{|X \cup Y|}$

- 库尔辛斯基相似度（Kulsinski similarity）

公式: $\frac{C_{TT}}{C_{TF}+C_{FT}+n}$

- 索卡尔米切纳相似度（Sokal-Michener similarity）

公式: $\frac{S}{S+R}$

其中有:

$$R = 2(C_{TF} + C_{FT})$$

$$S = C_{FF} + C_{TT}$$

3.1.2 实验过程

本次实验使用了数据集的前两条数据。使用 Python 语言进行实现，具体实现见代码。

3.2 相关系数和基于信息论的信息度量

3.2.1 实验原理

- 皮尔森相关系数（Pearson Correlation Coefficient）

两个变量之间的皮尔逊相关系数定义为两个变量之间的协方差和标准差的商。

公式一：

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_y} = \frac{E\big((X-\mu_X)(Y-\mu_Y)\big)}{\sigma_X \sigma_y} = \frac{E(XY)-E(X)E(Y)}{\sqrt{E(X^2)-E^2(X)}\sqrt{E(Y^2)-E^2(Y)}}$$

公式二：

$$\rho_{X,Y} = \frac{N\sum XY - \sum X \sum Y}{\sqrt{N\sum X^2 - (\sum X)^2}\sqrt{N\sum Y^2 - (\sum Y)^2}}$$

公式三：

$$\rho_{X,Y} = \frac{\sum (X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum (X-\bar{X})^2 \sum (Y-\bar{Y})^2}}$$

公式四：

$$\rho_{X,Y} = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\left(\sum X^2 - \frac{(\sum X)^2}{N}\right)\left(\sum Y^2 - \frac{(\sum Y)^2}{N}\right)}}$$

以上列出的四个公式等价，其中 E 是数学期望，cov 表示协方差，N 表示变量取值的个数。

- 斯皮尔曼相关性系数（Spearman Rank Correlation）

斯皮尔曼相关性系数，通常也叫斯皮尔曼秩相关系数。“秩”，可以理解成就是一种顺序或者排序，那么它就是根据原始数据的排序位置进行求解，这种表征形式就没有了求皮尔森相关性系数时那些限制。

公式：$r = \rho_{rgx,rg_r} = \frac{cov(rg_x, rg_y)}{\sigma_{rgx}\sigma_{rgy}}$

- 互信息/信息增益

互信息/信息增益：信息论中两个随机变量的相关性程度

公式：$I(X,Y) = \sum_{x\in X, y\in Y} P(x,y)\log\frac{p(x,y)}{p(x)p(y)}$

$I(X,Y) = H(X) - H(X \mid Y)$

H(X)是信息熵，H(X | Y)是条件熵

3.2.2 实验过程

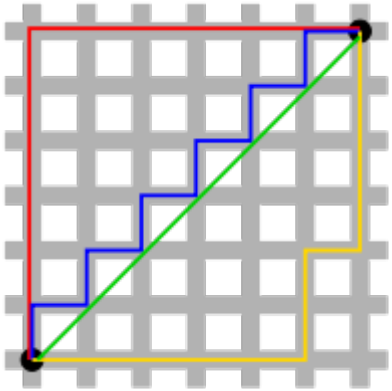
本次实验选择了数据集的 assets 和 liabilities 两列数据属性，即探究 125 家银行的资产与负债率之间的联系。使用 Python 语言进行实现，具体实现见代码。

4 实验结果图表与分析

4.1 距离和相似度计算实验

两行数据间距离种类	距离值
Euclidean Distances	14.5229349
Standardized Euclidean Distances	2.82842712
Manhattan Distances	23.10850619
Chebyshev Distances	11.11507591
Canberra Distances	1.9521123

标准欧氏距离的思路是现将各个维度的数据进行标准化：标准化后的值 = (标准化前的值 － 分量的均值) /分量的标准差，然后计算欧式距离。这样算出的距离显然要比不做标准化的欧式距离小很多。欧氏距离虽然很有用，但也有明显的缺点。它将样品的不同属性（即各指标或各变量量纲）之间的差别等同看待，这一点有时不能满足实际要求。例如，在教育研究中，经常遇到对人的分析和判别，个体的不同属性对于区分个体有着不同的重要性。因此，欧氏距离适用于向量各分量的度量标准统一的情况，标准化欧氏距离正是针对简单欧氏距离的缺点而作的一种改进方案。



上图中绿线代表的便是两点之间的欧氏距离，而其它颜色线都可以看作两点之间的曼哈顿距离。由三角形两边之和大于第三边可知，曼哈顿距离显然要大于同起终点的欧式距离，这与本实验中的结论相同。曼哈顿距离中的距离计算公式比欧氏距离的计算公式看起来简洁很多，只需要把两个点坐标的 x 坐标相减取绝对值，y 坐标相减取绝对值，再加和。

从公式定义上看，曼哈顿距离一定是一个非负数，距离最小的情况就是两个点重合，距离为 0，这一点和欧氏距离一样。曼哈顿距离和欧氏距离的意义相近，也是为了描述两个点之间的距离，不同的是曼哈顿距离只需要做加减法，这使得计算机在大

量的计算过程中代价更低，而且会消除在开平方过程中取近似值而带来的误差。不仅如此，曼哈顿距离在人脱离计算机做计算的时候也会很方便。

当 p 趋近于无穷大时，闵可夫斯基距离转化成切比雪夫距离。对于 2 维数据，切比雪夫距离可以形象地想象为国际象棋国王的最短移动步数。

坎贝拉距离是曼哈顿距离的加权结果，多用于计算机安全领域。

两行数据间相似度种类	相似程度
Cosine Similarity	0.91742122
Dice Similarity	1
Jaccard-Needham Similarity	0
Kulsinski Similarity	1
Sokal-Michener Similarity	1

下表是计算这些相似度使用的数据，可以看出，因为这两行数据没有一个属性的值是完全相同的，所以除了 cosine 相似度，其余基于两行数据属性交并集的相似度计算方法都得到了非 0 即 1 的结果，这显然是与实际情况不相符合的。由此可见，在数据集中数据的属性值交集很少甚至没有的时候，使用后面四种基于值交并集的相似度计算方法是不合理的，这时候 cosine 相似度就可以体现数据的相似程度了。相反，如果数据集中数据的属性值由一些固定值产生的时候，后面四种相似度也就能派上用处了。

bank	assets	liabilities	buffer	weights
b1	0.374909	9.631713	5.628295	17.11955
b2	0.668059	0.712655	2.847072	6.004475

4.2 相关系数和基于信息论的信息度量计算

两列数据间的相关系数和基于信息论的信息度量	计算值
Pearson Correlation	-0.008496238018387444
Spearman Correlation	0.004958525345622119
Mutual Information	4.828313737302302