

INT104W03_主成分分析(PCA)

在这之前不可不知的统计学知识

如果我们有一组数据 X ；其中第 i 个数据= X_i ；数据个数= n ；其平均数= X_{mean} ；期望值= $E[X]$ ；

方差 (Variance)

那么该如何表示数据集 X 的离散程度呢？

离散离散，就是看每个数据对于本组数据的平均值的偏差程度有多大。自然地，对数据 X_i ，作差 $(X_i - X_{mean})$ 就可以表示偏差程度的大小，但我们并不关心每个数据是偏大还是偏小（因为在和平均值比），故我们为此式开平方取正 $(X_i - X_{mean})^2$ 。对每个数据进行此操作后求和再平均，我们就得到了方差：一组数据中各数据与平均数的差的平方的平均数。

$$\sigma^2 = \frac{\sum_{i=1}^n (X_i - X_{mean})^2}{n}$$

总体方差代表着随机变量对数学期望的偏离程度（*当随机事件中各结果的可能性相等时，期望 $E[X]$ 退化为平均值 X_{mean} ）

* X 的方差亦可写作 $D(X)$ ；样本方差写作 S^2 ，此时因无偏估计需 $n-1$

*拓展了解：（概率分布函数的）矩（moment），变量的一阶原始矩等价于数学期望（expectation）、二至四阶中心矩被定义为方差（variance）、偏度（skewness）和峰度（kurtosis）。

标准差 (Standard Deviation)

方差固然好，告诉了我们数据偏离程度的大小，但在实际应用中却有局限。假设我加工3根10cm的木棍，但因加工误差导致成品实际长度为{5,10,15}。我们发现样本方差为 25cm^2 ，但这并没有什么意义，因为方差的量纲总是原数据单位的平方。如何解决？聪明的你一定想到了，开根。这就是标准差，由于与原数据单位相同，它更富有实际意义。

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (X_i - X_{mean})^2}{n}}$$

对于正态分布的数据，约68%的数据处于（平均值 $\pm\sigma$ ），约95%的数据处于（平均值 $\pm 2\sigma$ ）。

*样本标准差写作 S ，此时因无偏估计需 $n-1$

$$\text{计算总体方差: } \sigma^2 = \frac{\sum (X_i - \bar{X})^2}{N}$$

$$\text{总体标准差: } \sigma = \sqrt{\frac{\sum (X_i - \bar{X})^2}{N}}$$

$$\text{计算样本方差: } S^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$$

$$\text{样本标准差: } S = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n-1}}$$

式中 X_i 是数值序列中的单个数值, \bar{X} 是这组数值的平均值, N 是总体数值的个数, n 是样本数值的个数。

▮ 协方差 (Covariance)

我们已经可以描述清楚单个变量的偏离程度了! 但如果要统计两个随机变量(X, Y)间的偏离/变化趋势, 该怎么做呢? 协方差Cov(X, Y)可用来描述随机变量X与Y间相对期望的偏差的关联程度。当两个变量变化的趋势越一致 ($X > E[X]$ 时 $Y > E[Y]$ 反之亦然) 则 $Cov(X, Y) > 0$, 且越相关越大; 当两个变量变化的趋势越相反 ($X > E[X]$ 时 $Y < E[Y]$ 反之亦然) 则 $Cov(X, Y) < 0$, 且越负相关越小; 若协方差为0, 代表两组变量线性无关。协方差使我们可以计算两组变量间的线性关系, 它们有多么的线性相关。

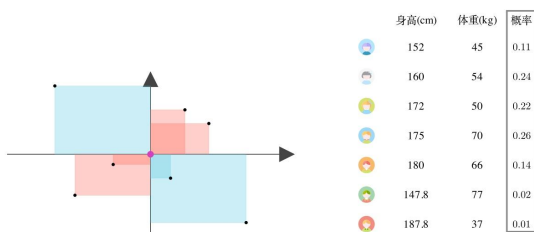
$$Cov(X, Y) = E[(X - E[X])(Y - E[Y])] = E(XY) - E[X]E[Y]$$

$$\text{其中 } E(XY) = \begin{cases} \sum_1^n \sum_1^n X_i Y_j P & (\text{离散情况}) \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf(x, y) & (\text{连续情况}) \end{cases}$$

即:

$$Cov(X, Y) = \frac{\sum_1^n (X_i - X_{mean})(Y_i - Y_{mean})}{n}$$

*可以发现, 方差就是一种特殊的协方差, 只是两个变量均为X, $D(X) = Cov(X, X)$ 。



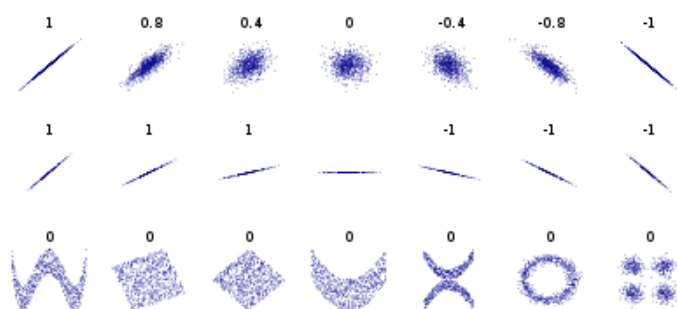
知乎 @马同学

若以各样本的重心为原点, 不难发现, 处在一三象限的样本 (正相关) 使协方差变大, 二四象限的样本 (负相关) 使协方差变小。

▮ 皮尔森相关系数 (Pearson correlation coefficient)

协方差很好，但如同方差一般，他包含原始数据的量纲（原数据单位的平方），取值范围并不确定，并不能方便地比较不同类型数据的相关性大小。如果能归一化，将取值范围约束进[-1,1]就更有可比性了。这里不除以数据个数，而是两个变量的标准差即可。

$$\rho_{X,Y} = \frac{\sum_1^n (X_i - X_{mean})(Y_i - Y_{mean})}{\sigma_x \sigma_y}$$



*在几何上，相关系数的绝对值等于数据集两条可能的回归线 $y=g_x(x)$ 与 $x=g_y(y)$ 夹角的余弦

但要注意，连续，正态分布，线性的数据才能用pearson相关系数**，皮尔森相关系数的前提是数据成正态分布；对于非正态分布的数据，要检验其**单调相关性，使用spearman相关系数**。

在这之前不可不知的线代知识

$$\text{若数据集 } D = \begin{bmatrix} X_1 & X_2 & X_3 & X_4 \\ Y_1 & Y_2 & Y_3 & Y_4 \end{bmatrix}$$

拉伸

对矩阵D，若要将其在X，Y轴方向上拉伸，可构造对角阵S左乘D，对角线上的每个元素就是对应维度的拉伸倍数，很容易扩展到高维。

$$S = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$$

这里在X方向拉伸为原长的两倍，Y方向不变。

$$SD = \begin{bmatrix} 2X_1 & 2X_2 & 2X_3 & 2X_4 \\ Y_1 & Y_2 & Y_3 & Y_4 \end{bmatrix}$$

旋转

对矩阵D，若要将其对原点处旋转 θ 角度，可构造旋转矩阵R。（可以试着写出R左乘对基向量的影响）（还有很多其他的方法旋转空间）

$$R = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}$$

$$RD = \begin{bmatrix} \cos(\theta)X_1 - \sin(\theta)Y_1 & .. & .. & .. \\ \sin(\theta)X_1 + \cos(\theta)Y_1 & .. & .. & .. \end{bmatrix}$$

■ 协方差矩阵

可以看到主对角线上分别是两个变量的方差，其余是两个变量的协方差。

$$C = \begin{bmatrix} Cov(x, x) & Cov(x, y) \\ Cov(x, y) & Cov(y, y) \end{bmatrix}$$

*对高维情况， $c_{ij} = Cov(F_i, F_j)$

■ 主成分分析

特征向量=新坐标轴方向（若有多个，必正交）

特征值=对应新坐标轴方向上的方差

基本步骤：原始数据→数据矩阵→Z-Score（零均值中心化）→每个点与数据集的协方差矩阵（等式）→解方程求得特征向量与对应特征值

■ 1.原始数据

$$\text{假设我们有数据集 } F = \begin{bmatrix} 2 & 5 & 6 & 8 \\ 3 & 5 & 6 & 9 \end{bmatrix}$$

$$\text{容易得到 } \begin{cases} F_{1mean} = 5.25 \\ F_{2mean} = 5.75 \end{cases} \quad \begin{cases} F_{1std} = 2.5 \\ F_{2std} = 2.5 \end{cases} \text{ (为方便计算标准差假设为2.5)}$$

■ 2.Z-Score中心化（零均值化）

我们在这一步将坐标原点移动至数据集重心，将平均数归零，简便了协方差的计算

$$X'_i = \frac{X_i - X_{mean}}{X_{std}}$$

$$A' = \left(\frac{2 - 5.25}{2.5}, \frac{3 - 5.75}{2.5} \right) = (-1.3, -1.1) \quad (1)$$

$$B' = (-0.1, -0.3) \quad (2)$$

$$C' = (0.3, 0.1) \quad (3)$$

$$D' = (1.1, 1.3) \quad (4)$$

3. 协方差矩阵

$$\text{数据集的协方差矩阵} : S = \frac{1}{n} \sum_1^n X_i \cdot X_i^T$$

我们先求出每组数据的协方差矩阵，仅需自己（列向量）乘以自己的转置

$$A' = \begin{bmatrix} -1.3 \\ -1.1 \end{bmatrix} \cdot \begin{bmatrix} -1.3 & -1.1 \end{bmatrix} = \begin{bmatrix} 1.69 & 1.43 \\ 1.43 & 1.21 \end{bmatrix} \quad (5)$$

$$B' = \begin{bmatrix} 0.01 & 0.03 \\ 0.03 & 0.09 \end{bmatrix} \quad (6)$$

$$C' = \begin{bmatrix} 0.09 & 0.03 \\ 0.03 & 0.01 \end{bmatrix} \quad (7)$$

$$D' = \begin{bmatrix} 1.21 & 1.43 \\ 1.43 & 1.69 \end{bmatrix} \quad (8)$$

然后求出数据集的协方差矩阵S

$$S = \frac{A' + B' + C' + D'}{4} = \begin{bmatrix} 0.75 & 0.73 \\ 0.73 & 0.75 \end{bmatrix}$$

4. 求出协方差矩阵的特征值及对应的特征向量

数据集在某方向C₁上投影的方差（我就是特征值！）V₁：

$$V_1 = \frac{1}{n} \sum_1^n (C_1^T \cdot X_i)^2 \quad (9)$$

$$= \frac{1}{n} \sum_1^n C_1^T \cdot X_i \cdot C_1^T \cdot X_i \quad (10)$$

$$= \frac{1}{n} \sum_1^n C_1^T \cdot (X_i \cdot X_i^T) \cdot C_1 \quad (11)$$

$$= C_1^T \left(\frac{1}{n} \sum_1^n X_i \cdot X_i^T \right) C_1 \quad (12)$$

$$= C_1^T \cdot S \cdot C_1 = V_1 \quad (13)$$

*注：记最后一行就行

由 (13) 可得：

$$C_1^T \cdot S \cdot C_1 = V_1 = \lambda \quad (\text{特征值}) \quad (14)$$

$$S \cdot C_1 = \lambda \cdot C_1 \quad (15)$$

$$(S - \lambda I)C_1 = 0 \quad (16)$$

由于方向C1不为零，只可能是括号内部分=0

$$S - \lambda I = \begin{bmatrix} 0.75 - \lambda & 0.73 \\ 0.73 & 0.75 - \lambda \end{bmatrix} = 0 \quad (17)$$

$$\text{即行列式为零：} (0.75 - \lambda)^2 - 0.73^2 = 0 \quad (18)$$

$$\lambda^2 - 1.5\lambda + 0.0296 = 0 \quad (19)$$

$$\text{解得：} \begin{cases} \lambda_1 = 1.48 \\ \lambda_2 = 0.02 \end{cases} \quad (20)$$

我们就求得了特征值，接下来求对应的特征向量

$$\text{由此式: } (S - \lambda I)C_1 = 0 \quad (21)$$

$$\text{特征值 } \lambda_1 \text{ 得: } \begin{bmatrix} 0.75 - \lambda_1 & 0.73 \\ 0.73 & 0.75 - \lambda_1 \end{bmatrix} \cdot \begin{bmatrix} C_{11} \\ C_{12} \end{bmatrix} = 0 \quad (22)$$

$$\begin{bmatrix} -0.73C_{11} + 0.73C_{12} \\ 0.73C_{11} - 0.73C_{12} \end{bmatrix} = 0 \quad (23)$$

$$\Rightarrow C_{11} = C_{12} \quad (24)$$

$$\text{类似地, 特征值 } \lambda_2 \text{ 得: } C_{21} = -C_{22} \quad (25)$$

5.归一化特征向量

$$\because C_{11} = C_{12} \quad (26)$$

$$\therefore C_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad (27)$$

$$\because |C_1| = \sqrt{2}, \text{ 我们需要将模长归一} \quad (28)$$

$$\therefore C_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \times \frac{1}{\sqrt{2}} \quad (29)$$

$$\text{类似地, 对特征值 } \lambda_2: C_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \times \frac{1}{\sqrt{2}} \quad (30)$$

*注: 有些答案中需要对特征向量加上 \pm