内积矩阵与协方差矩阵

一、内积矩阵

假设n维空间单个样本$x_i = (x_{i1}, x_{i2}, \cdots x_{in})$,用列向量表示，即

$$x_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \cdots \\ x_{in} \end{bmatrix}$$

由m个样本组成的样本矩阵表示为$X_{m\times n}$(因为编程时一般习惯 array 的第一个维度为样本数)，即

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{bmatrix}_{m\times n} = \begin{bmatrix} -x_1^T - \\ \cdots \\ -x_m^T - \end{bmatrix}_{m\times n}$$

这里有如下两种观点，从行的角度，每一行代表了 1 个样本，其有 n 个特征，总共有 m 个这样的样本。从列的角度，每一列代表了 1 个特征的 m 份抽样，总共有 n 个这样的特征。从这两个角度出发，分别引出了样本的内积矩阵和协方差矩阵。

下面，先介绍样本的内积矩阵。内积矩阵是从行的角度，研究两两样本的相似性。由于有 m 个样本，不难得知这样的相似度结果共有m^2个。如果将其用矩阵表示，那应该是$m\times m$维。

如何表征样本的相似性呢？我们知道单个样本是由 n 个特征表示的 n 维列向量，而向量的内积正是刻画两个向量相似度的一种手段。

$$x_1^T x_2 = \sum_{i=1}^{n} x_{1i} x_{2i}$$

也就是说，最终的结果应该应该是下面这个样子。

$$P_{m\times m} = \begin{bmatrix} x_1^T x_1 & \cdots & x_1^T x_n \\ \vdots & \ddots & \vdots \\ x_n^T x_1 & \cdots & x_n^T x_n \end{bmatrix}_{m\times m} = \begin{bmatrix} -x_1^T - \\ \cdots \\ -x_m^T - \end{bmatrix}_{m\times n} \begin{bmatrix} x_1 & \cdots & x_m \end{bmatrix}_{n\times m}$$

即样本的内积矩阵P可以表示为，

$$\boldsymbol{P = XX^T}$$

内积矩阵P的每一个元素p_{ij}表示了样本x_i和样本x_j的相似度(内积)。

二、协方差矩阵

接下来，从列的角度(特征)研究协方差矩阵。用$z_i = (z_{i1}, z_{i2}, \cdots z_{im})$表示样本矩阵$X$的第$i$列。此时样本矩阵$X$表示如下，

$$X = [z_1 \quad \cdots \quad z_n]_{m\times n} = \begin{bmatrix} z_{11} & \cdots & z_{n1} \\ \vdots & \ddots & \vdots \\ z_{1m} & \cdots & z_{nm} \end{bmatrix}_{m\times n}$$

不失一般性，假设样本的每一个特征都已经被中心化，即样本矩阵X的每一列中的元素都减去这一列元素的均值。以第一列为例子，

$$\begin{bmatrix} z_{11} \\ \vdots \\ z_{m1} \end{bmatrix} \rightarrow \begin{bmatrix} z_{11} - \mu_1 \\ \vdots \\ z_{m1} - u_1 \end{bmatrix}$$

其中

$$\mu_1 = \frac{1}{m}(z_{11} + z_{21} + \cdots + z_{m1})$$

根据数理统计可知，两个**随机变量**(在这里随机变量就是样本的特征)的协方差为：

$$Cov(z_1, z_2) = E[(z_1 - E[z_1])(z_2 - E[z_2])]$$

由于，我们已经将特征中心化(即$E[z_1] = 0, E[z_2] = 0$),

$$Cov(z_1, z_2) = E[z_1 z_2]$$

对其进行m次抽样，可以得到**样本**的协方差

$$Cov(z_1, z_2) = z_1^T z_2 = z_{11}z_{21} + z_{12}z_{22} + \cdots + z_{1m}z_{2m}$$

下面将对样本的n个特征，分析其两两的协方差。不难得知，样本的协方差矩阵应该为$n\times n$维。

为了方便，我们将样本矩阵X表示如下：

$$X = [z_1 \quad \cdots \quad z_n]_{m\times n} = \begin{bmatrix} z_{11} & \cdots & z_{n1} \\ \vdots & \ddots & \vdots \\ z_{1m} & \cdots & z_{nm} \end{bmatrix}_{m\times n}$$

仿照内积矩阵的推导，可以得知样本的协方差矩阵C为

$$C = \begin{bmatrix} cov(z_1, z_1) & \cdots & cov(z_1, z_n) \\ \vdots & \ddots & \vdots \\ cov(z_n, z_1) & \cdots & cov(z_n, z_n) \end{bmatrix} = \begin{bmatrix} z_1^T z_1 & \cdots & z_1^T z_n \\ \vdots & \ddots & \vdots \\ z_n^T z_1 & \cdots & z_n^T z_n \end{bmatrix} = \begin{bmatrix} -z_1^T - \\ \cdots \\ -z_1^T - \end{bmatrix} \begin{bmatrix} z_1 & \cdots & z_n \end{bmatrix}$$

也就是说，

$$\boldsymbol{C = X^T X}$$

三、总结

总结一下，样本矩阵X有两种理解角度。从行的角度，也就是单个样本的角度，可以研究样本的相似性，推导出样本的内积矩阵。

$$X = \begin{bmatrix} -x_1^T - \\ \cdots \\ -x_m^T - \end{bmatrix}_{m \times n}$$

$$P = XX^T$$

从列的角度，也就是单个特征的角度，可以研究特征之间的相关性，推导出样本的协方差矩阵。

$$X = \begin{bmatrix} z_1 & \cdots & z_n \end{bmatrix}_{m \times n}$$

$$C = X^T X$$

注意，不要死记硬背，具体使用公式时可以从维度一致的角度推导。