## 大数据分析实验报告——数据降维

### 苏致成 201250104

### 注明：

本文档采用markdown书写，因此，转为word部分公式可能出现显示异常等问题。

### 使用方法

PCA（主成分分析法），主要思想是将n维特征映射到k维上，生成的k维的正交特征被称为主成分。

数据集为 。

注：本例数据集 中，因为分类变量较多，应使用 等工具更为合适，但是因为无法在 中找到对应的工具包，因此此处仍使用经典的 。

#### 算法流程

1. 输入数据集。
2. 去中心化，将每一特征减去各自的均值。
3. 计算协方差矩阵。
4. 求出上述协方差矩阵的特征值和特征向量。
5. 对特征值从大到小进行排序，选取最大的 个特征值。将其对应的 个特征向量作为行向量组成特征向量矩阵 。
6. 利用上述矩阵 ,计算出降维后的向量表示，即 。

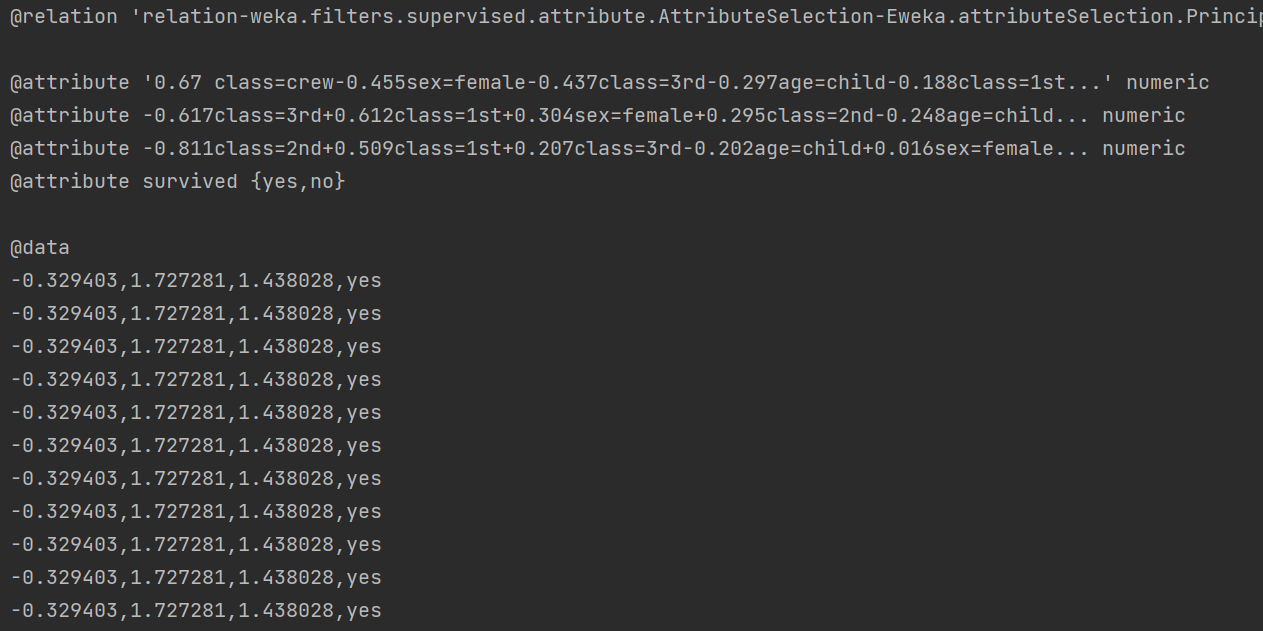
### 数据集处理思路

1. 导入 数据集。
2. 处理缺失值，打印输出可知实际上并无缺失值。规范化和标准化等过程在调用库中进行处理。
3. 设置ClassIndex的数目，使得最后一列不作为降维的标准（最后一列为生存与否，与前三列属性意义不一致）。
4. 利用 类调用 设置降维后选择的主成分为三维。
5. 调用 将过滤标准导入。输出结果。

注：该库（）已完成了去均值化等操作，故此处不予显式处理。

### 实验结果

部分降维结果如下：



Correlation matrix：相关系数矩阵

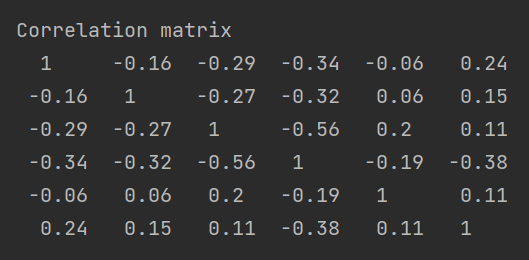


image-20221010103133361

Eigenvalue：特征值

Proportion：方差贡献率

Cumulative：累计方差贡献率

Eigenvectors：特征向量

