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 $titanic.arff$ 。

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

算法流程

1. 输入数据集$X=\{x_1,x_2,x_3,\ldots,x_n\}$。
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$\frac{1}{n}XX^T$。
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 k 个特征值。将其对应的 k 个特征向量作为行向量组成特征向量矩阵 P。
6. 利用上述矩阵 P,计算出降维后的向量表示，即 $Y=PX$。

数据集处理思路

1. 导入 $titanic.arff$ 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 $Ranker$ 类调用 $setNumToSelect$ 设置降维后选择的主成分为三维。
5. 调用$Filter$ 将过滤标准导入。输出结果。

注：该库（$PrincipalComponents$）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：

```
@relation 'relation-weka.filters.supervised.attribute.AttributeSelection-Eweka.attributeSelection.Princi

@attribute '0.67 class=crew-0.455sex=female-0.437class=3rd-0.297age=child-0.188class=1st...' numeric
@attribute -0.617class=3rd+0.612class=1st+0.304sex=female+0.295class=2nd-0.248age=child... numeric
@attribute -0.811class=2nd+0.509class=1st+0.207class=3rd-0.202age=child+0.016sex=female... numeric
@attribute survived {yes,no}

@data
-0.329403,1.727281,1.438028,yes
-0.329403,1.727281,1.438028,yes
-0.329403,1.727281,1.438028,yes
-0.329403,1.727281,1.438028,yes
-0.329403,1.727281,1.438028,yes
-0.329403,1.727281,1.438028,yes
-0.329403,1.727281,1.438028,yes
-0.329403,1.727281,1.438028,yes
-0.329403,1.727281,1.438028,yes
-0.329403,1.727281,1.438028,yes
-0.329403,1.727281,1.438028,yes
```

Correlation matrix：相关系数矩阵

```
Correlation matrix
  1      -0.16   -0.29   -0.34   -0.06    0.24
 -0.16    1      -0.27   -0.32    0.06    0.15
 -0.29   -0.27    1      -0.56    0.2     0.11
 -0.34   -0.32   -0.56    1      -0.19   -0.38
 -0.06    0.06    0.2    -0.19    1       0.11
  0.24    0.15    0.11   -0.38    0.11    1
```

Eigenvalue：特征值

Proportion：方差贡献率

Cumulative：累计方差贡献率

Eigenvectors：特征向量

```
eigenvalue  proportion  cumulative
   1.88237     0.31373     0.31373   0.67 class=crew-0.455sex=female-0.437class=3rd-0.297age=child-0.188class=1st...
   1.3844      0.23073     0.54446   -0.617class=3rd+0.612class=1st+0.304sex=female+0.295class=2nd-0.248age=child...
   1.17705     0.19618     0.74064   -0.811class=2nd+0.509class=1st+0.207class=3rd-0.202age=child+0.016sex=female...
   0.86776     0.14463     0.88526   0.895age=child-0.27class=3rd+0.243class=crew+0.187class=1st-0.177class=2nd...
   0.68842     0.11474     1         0.837sex=female+0.379class=crew-0.34class=1st-0.181class=2nd-0.094age=child...

Eigenvectors
 V1  V2  V3  V4  V5
-0.188   0.6125  0.5085  0.1873 -0.3395 class=1st
-0.1723  0.2953 -0.8107 -0.1774 -0.1806 class=2nd
-0.4371 -0.6167  0.207  -0.2703 -0.0097 class=3rd
 0.6701 -0.0583 -0.0098  0.2433  0.3786 class=crew
-0.2966 -0.2484 -0.2024  0.8947 -0.0937 age=child
-0.455   0.3037  0.0155  0.0246  0.8366 sex=female
```