模型-机器学习-降维-主成分分析【hxy】

1. 模型名称

主成分分析（Principal Components Analysis，PCA）

2. 适用范围

将高维且具有很多相关性很高的变量降维成相对低维相互独立的变量

3. 形式

高维且相关性很高的变量

4. 求解方法

4.1 步骤

1. 构造**原始数据矩阵**A

$$A=(a_{ij})_{n\times m}\quad n=number\ of\ samples,\ m=numberof\ indexes$$

2. 进行**标准化处理**

$$\mu_j=\frac{1}{n}\sum_{i=1}^{n}a_{ij}\quad j=1,2,\dots,m$$

$$s_j=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(a_{ij}-\mu_j)^2}\quad j=1,2,\dots,m$$

$$\mu_j\ is\ the\ mean\ value,\ s_j\ is\ the\ sample\ standard\ deviation$$

$$\widetilde{a}_{ij}=\frac{a_{ij}-\mu_j}{s_j}\quad i=1,2,\dots,n,\ j=1,2,\dots,m$$

$$\widetilde{x}_j=\frac{x_j-\mu_j}{s_j}\quad j=1,2,\dots,m$$

$$\widetilde{x}_j\ is\ standardized\ indicator\ variable$$

3. 计算**相关系数矩阵**R

$$r_{ij} = \frac{\sum_{k=1}^{n} \widetilde{a}_{ki} \cdot \widetilde{a}_{kj}}{n-1} \quad i,j = 1,2,\ldots,m$$
$$where\ r_{ii} = 1,\ r_{ij} = r_{ji}$$

4. 计算**特征值**和**特征向量**

$$Eigenvalue \quad \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m \geq 0$$
$$Eigenvector \quad u_1,\ u_2,\ \ldots,\ u_m$$

5. 由特征向量组成m个新的**指标变量**

$$y_1 = u_{11}\widetilde{x}_1 + u_{21}\widetilde{x}_2 + \ldots + u_{m1}\widetilde{x}_m$$
$$y_2 = u_{12}\widetilde{x}_1 + u_{22}\widetilde{x}_2 + \ldots + u_{m2}\widetilde{x}_m$$
$$\ldots$$
$$y_m = u_{1m}\widetilde{x}_1 + u_{2m}\widetilde{x}_2 + \ldots + u_{mm}\widetilde{x}_m$$
$$y_1\ is\ the\ first\ principle\ component,\ etc.$$

通过这个可以看出每个主成分y与原指标x之间的关系，明白每个主成分主要是跟什么有关，例如下图

1	0.34	0.35	0.36	0.36	0.36	0.36	0.22	0.12	0.31	0.24
2	-0.19	0.03	0.02	0.01	-0.05	-0.06	0.58	0.70	-0.19	-0.28
3	-0.16	-0.10	-0.09	-0.11	-0.15	-0.16	-0.03	0.35	0.12	0.86
4	-0.10	-0.22	-0.16	-0.16	-0.04	-0.00	0.08	0.07	0.89	0.24

- 从主成分的系数可以看出，
  - 第一主成分主要反映了前六个指标（学校数、学生数和教师数方面）的信息，
  - 第二主成分主要反映了高校规模和教师中高级职称的比例，
  - 第三主成分主要反映了生均教育经费，
  - 第四主成分主要反映了国家财政预算内普通高教经费占国内生产总值的比重。 分别以四个主成分的贡献率为权重，构建主成分综合评价模型

6. 计算特征值$\lambda_j (j = 1,2,\ldots,m)$**的信息贡献率**$b_j$

$$b_j = \frac{\lambda_j}{\sum_{k=1}^{m} \lambda_k} \quad j = 1,2,\ldots,m$$
$$b_j\ is\ called\ information\ contribution\ rate$$

7. 计算$y_1, y_2, \ldots, y_p$**的累积贡献率**α_p

$$\alpha_p = \frac{\sum_{k=1}^{p} \lambda_k}{\sum_{k=1}^{m} \lambda_k}$$

8. 选择前p个主成分使$\alpha_p \geq 0.85$（接近1）

4.2 实例

年份	投资效果系数（无时滞）	投资效果系数（时滞一年）	全社会固定资产交付使用率	建设项目投产率
1984	0.71	0.49	0.41	0.51
1985	0.4	0.49	0.44	0.57
1986	0.55	0.56	0.48	0.53
1987	0.62	0.93	0.38	0.53
1988	0.45	0.42	0.41	0.54
1989	0.36	0.37	0.46	0.54
1990	0.55	0.68	0.42	0.54
1991	0.62	0.9	0.38	0.56
1992	0.61	0.99	0.33	0.57
1993	0.71	0.93	0.35	0.66
1994	0.59	0.69	0.36	0.57
1995	0.41	0.47	0.4	0.54
1996	0.26	0.29	0.43	0.57
1997	0.14	0.16	0.43	0.55
1998	0.12	0.13	0.45	0.59
1999	0.22	0.25	0.44	0.58
2000	0.71	0.49	0.41	0.51

1. 构造**原始数据矩阵**A

$$A=\begin{pmatrix}0.71 & 0.49 & 0.41 & 0.51 & 0.46\\ 0.4 & 0.49 & 0.44 & 0.57 & 0.5\\ 0.55 & 0.56 & 0.48 & 0.53 & 0.49\\ 0.62 & 0.93 & 0.38 & 0.53 & 0.47\\ 0.45 & 0.42 & 0.41 & 0.54 & 0.47\\ 0.36 & 0.37 & 0.46 & 0.54 & 0.48\\ 0.55 & 0.68 & 0.42 & 0.54 & 0.46\\ 0.62 & 0.9 & 0.38 & 0.56 & 0.46\\ 0.61 & 0.99 & 0.33 & 0.57 & 0.43\\ 0.71 & 0.93 & 0.35 & 0.66 & 0.44\\ 0.59 & 0.69 & 0.36 & 0.57 & 0.48\\ 0.41 & 0.47 & 0.4 & 0.54 & 0.48\\ 0.26 & 0.29 & 0.43 & 0.57 & 0.48\\ 0.14 & 0.16 & 0.43 & 0.55 & 0.47\\ 0.12 & 0.13 & 0.45 & 0.59 & 0.54\\ 0.22 & 0.25 & 0.44 & 0.58 & 0.52\\ 0.71 & 0.49 & 0.41 & 0.51 & 0.46\end{pmatrix}$$

2. 进行**标准化处理**

a) **均值**μ

$$\mu_j=\frac{1}{17}\sum_{i=1}^{17}a_{ij}\quad j=1,2,\dots,5$$

$$\mu=\begin{pmatrix}0.4724 & 0.5435 & 0.4106 & 0.5565 & 0.4759\end{pmatrix}$$

b) **标准差**S

$$s_j=\sqrt{\frac{1}{17-1}\sum_{i=1}^{17}(a_{ij}-\mu_j)^2}\quad j=1,2,\dots,5$$

$$S=\begin{pmatrix}0.1974 & 0.2731 & 0.0404 & 0.0353 & 0.0267\end{pmatrix}$$

c) **标准化指标**$\widetilde{A}$

$$\widetilde{a}_{ij}=\frac{a_{ij}-\mu_j}{s_j}\quad i=1,2,\ldots,n,\ j=1,2,\ldots,m$$

$$\widetilde{A}=\begin{pmatrix}1.2038 & -0.1960 & -0.0146 & -1.3148 & -0.5947\\ -0.3665 & -0.1960 & 0.7283 & 0.3828 & 0.9031\\ 0.3933 & 0.0603 & 1.7188 & -0.7489 & 0.5286\\ 0.7479 & 1.4151 & -0.7574 & -0.7489 & -0.2203\\ -0.1132 & -0.4523 & -0.0146 & -0.4660 & -0.2203\\ -0.5691 & -0.6354 & 1.2235 & -0.4660 & 0.1542\\ 0.3933 & 0.4997 & 0.2331 & -0.4660 & -0.5947\\ 0.7479 & 1.3052 & -0.7574 & 0.0999 & -0.5947\\ 0.6973 & 1.6348 & -1.9955 & 0.3828 & -1.7180\\ 1.2038 & 1.4151 & -1.5003 & 2.9291 & -1.3436\\ 0.5960 & 0.5363 & -1.2527 & 0.3828 & 0.1542\\ -0.3159 & -0.2692 & -0.2622 & -0.4660 & 0.1542\\ -1.0757 & -0.9283 & 0.4807 & 0.3828 & 0.1542\\ -1.6836 & -1.4043 & 0.4807 & -0.1831 & -0.2203\\ -1.7849 & -1.5142 & 0.9759 & 0.9486 & 2.4008\\ -1.2783 & -1.0748 & 0.7283 & 0.6657 & 1.6519\\ 1.2038 & -0.1960 & -0.0146 & -1.3148 & -0.5947\end{pmatrix}$$

3. 计算**相关系数矩阵**R

$$r_{ij}=\frac{\sum_{k=1}^{17}\widetilde{a}_{ki}\cdot\widetilde{a}_{kj}}{17-1}\quad i,j=1,2,\ldots,5$$

$$R=\begin{pmatrix}1.0000 & 0.8097 & -0.5764 & -0.1519 & -0.7141\\ 0.8097 & 1.0000 & -0.7573 & 0.1619 & -0.7005\\ -0.5764 & -0.7573 & 1.0000 & -0.3225 & 0.6862\\ -0.1519 & 0.1619 & -0.3225 & 1.0000 & 0.0499\\ -0.7141 & -0.7005 & 0.6862 & 0.0499 & 1.0000\end{pmatrix}$$

4. 计算**特征值**和**特征向量**

$$Eigenvalue\quad \lambda_1\geq\lambda_2\geq\ldots\geq\lambda_m\geq 0$$

$$Eigenvactor\quad u_1,\ u_2,\ \ldots,\ u_m$$

$$\lambda=(3.1343\quad 1.1683\quad 0.3502\quad 0.2258\quad 0.1213)$$

$$U=\begin{pmatrix}0.4905 & -0.2934 & 0.5109 & 0.1896 & -0.6134\\ 0.5254 & 0.0490 & 0.4337 & -0.1217 & 0.7202\\ -0.4871 & -0.2812 & 0.3714 & 0.6888 & 0.2672\\ 0.0671 & 0.8981 & 0.1477 & 0.3863 & -0.1336\\ -0.4916 & 0.1606 & 0.6255 & -0.5706 & -0.1254\end{pmatrix}$$

5. 由特征向量组成m个新的**指标变量**

$$
\begin{aligned}
y_1 &= 0.4905\widetilde{x}_1 + 0.5254\widetilde{x}_2 - 0.4871\widetilde{x}_3 + 0.0671\widetilde{x}_4 - 0.4916\widetilde{x}_5 \\
y_2 &= -0.2934\widetilde{x}_1 + 0.0490\widetilde{x}_2 - 0.2812\widetilde{x}_3 + 0.8981\widetilde{x}_4 + 0.1606\widetilde{x}_5 \\
y_3 &= 0.5109\widetilde{x}_1 + 0.4337\widetilde{x}_2 + 0.3714\widetilde{x}_3 + 0.1477\widetilde{x}_4 + 0.6255\widetilde{x}_5 \\
y_4 &= 0.1896\widetilde{x}_1 - 0.1217\widetilde{x}_2 + 0.6888\widetilde{x}_3 + 0.3863\widetilde{x}_4 - 0.5706\widetilde{x}_5 \\
y_5 &= -0.6134\widetilde{x}_1 + 0.7202\widetilde{x}_2 + 0.2672\widetilde{x}_3 - 0.1336\widetilde{x}_4 - 0.1254\widetilde{x}_5 \\
&y_1 \ is \ the \ first \ principle \ component, \ etc.
\end{aligned}
$$

6. 计算特征值$\lambda_j(j=1,2,\ldots,m)$的信息贡献率$b_j(\%)$

$$
b_j = \frac{\lambda_j}{\sum_{k=1}^{5}\lambda_k} \quad j = 1,2,\ldots,5
$$

$$
B = (62.6866 \quad 23.3670 \quad 7.0036 \quad 4.5162 \quad 2.4266)
$$

7. 计算$y_1, y_2, \ldots, y_p$的累积贡献率α_p

$$
\alpha_p = \frac{\sum_{k=1}^{p}\lambda_k}{\sum_{k=1}^{m}\lambda_k}
$$

$$
\alpha = (62.6866 \quad 86.0536 \quad 93.0572 \quad 97.5734 \quad 100.0000)
$$

8. 此处选择前3个主成分，累积贡献率达93%，主成分分析效果很好，维数由5维降到3维

$$
\begin{aligned}
y_1 &= 0.4905\widetilde{x}_1 + 0.5254\widetilde{x}_2 - 0.4871\widetilde{x}_3 + 0.0671\widetilde{x}_4 - 0.4916\widetilde{x}_5 \\
y_2 &= -0.2934\widetilde{x}_1 + 0.0490\widetilde{x}_2 - 0.2812\widetilde{x}_3 + 0.8981\widetilde{x}_4 + 0.1606\widetilde{x}_5 \\
y_3 &= 0.5109\widetilde{x}_1 + 0.4337\widetilde{x}_2 + 0.3714\widetilde{x}_3 + 0.1477\widetilde{x}_4 + 0.6255\widetilde{x}_5
\end{aligned}
$$

4.3 代码实现

pca_dimension.m

代码：

```
clc, clear
% Input data with rows of samples and columns of indexes
a = [0.71 0.49  0.41  0.51  0.46
0.40  0.49  0.44  0.57  0.50
0.55  0.56  0.48  0.53  0.49
0.62  0.93  0.38  0.53  0.47
0.45  0.42  0.41  0.54  0.47
0.36  0.37  0.46  0.54  0.48
0.55  0.68  0.42  0.54  0.46
0.62  0.90  0.38  0.56  0.46
0.61  0.99  0.33  0.57  0.43
0.71  0.93  0.35  0.66  0.44
0.59  0.69  0.36  0.57  0.48
0.41  0.47  0.40  0.54  0.48
0.26  0.29  0.43  0.57  0.48
0.14  0.16  0.43  0.55  0.47
0.12  0.13  0.45  0.59  0.54
0.22  0.25  0.44  0.58  0.52
```

```
0.71  0.49  0.41  0.51  0.46];

% Standardize data
standardized_a = zscore(a);
disp('标准化后数据：');
disp(standardized_a);

% Calculate corrcoef matrix
r = corrcoef(standardized_a);
disp('相关系数矩阵：');
disp(r);

% Calculate eigenvalues y, eigenvectors x, contribution p
[x, y, p] = pcacov(r);
% Construct row vector of +1/-1
f = sign(sum(x));
% Modify the sign of eigenvectors x
x = x .* f;
disp('特征值：');
disp(y');
disp('特征向量：');
disp(x');
disp('贡献率(%)：');
disp(p');
% Calculate cummulative contribution p_cum
p_cum = cumsum(p);
disp('累计贡献率(%)：');
disp(p_cum');

% Choose the number of principle components
num = 3;
disp(['PCA选取了前',num2str(num),'个主成分']);
disp(['累计贡献率达',num2str(p_cum(num)),'%']);
disp(' ');
new_p = p((1:num),1);
disp('主成分分析后成分各自贡献率：');
disp(new_p'/100);
new_x = x(:,(1:num));
disp('主成分分析后特征向量：');
disp(new_x');
```

结果：

```
标准化后数据:
    1.2038   -0.1960   -0.0146   -1.3148   -0.5947
   -0.3665   -0.1960    0.7283    0.3828    0.9031
    0.3933    0.0603    1.7188   -0.7489    0.5286
    0.7479    1.4151   -0.7574   -0.7489   -0.2203
   -0.1132   -0.4523   -0.0146   -0.4660   -0.2203
   -0.5691   -0.6354    1.2235   -0.4660    0.1542
    0.3933    0.4997    0.2331   -0.4660   -0.5947
    0.7479    1.3052   -0.7574    0.0999   -0.5947
    0.6973    1.6348   -1.9955    0.3828   -1.7180
    1.2038    1.4151   -1.5003    2.9291   -1.3436
    0.5960    0.5363   -1.2527    0.3828    0.1542
   -0.3159   -0.2692   -0.2622   -0.4660    0.1542
   -1.0757   -0.9283    0.4807    0.3828    0.1542
   -1.6836   -1.4043    0.4807   -0.1831   -0.2203
   -1.7849   -1.5142    0.9759    0.9486    2.4008
   -1.2783   -1.0748    0.7283    0.6657    1.6519
    1.2038   -0.1960   -0.0146   -1.3148   -0.5947

相关系数矩阵:
    1.0000    0.8097   -0.5764   -0.1519   -0.7141
    0.8097    1.0000   -0.7573    0.1619   -0.7005
   -0.5764   -0.7573    1.0000   -0.3225    0.6862
   -0.1519    0.1619   -0.3225    1.0000    0.0499
   -0.7141   -0.7005    0.6862    0.0499    1.0000

特征值:
    3.1343    1.1683    0.3502    0.2258    0.1213

特征向量:
    0.4905    0.5254   -0.4871    0.0671   -0.4916
   -0.2934    0.0490   -0.2812    0.8981    0.1606
    0.5109    0.4337    0.3714    0.1477    0.6255
    0.1896   -0.1217    0.6888    0.3863   -0.5706
   -0.6134    0.7202    0.2672   -0.1336   -0.1254

贡献率(%):
   62.6866   23.3670    7.0036    4.5162    2.4266

累计贡献率(%):
   62.6866   86.0536   93.0572   97.5734  100.0000

PCA选取了前3个主成分
累计贡献率达93.0572%

主成分分析后成分各自贡献率:
    0.6269    0.2337    0.0700

主成分分析后特征向量:
    0.4905    0.5254   -0.4871    0.0671   -0.4916
   -0.2934    0.0490   -0.2812    0.8981    0.1606
    0.5109    0.4337    0.3714    0.1477    0.6255
```

5. 参考资料

1. 数模官网-降维PCA
2. 《数学建模算法与应用》：P427-P430
3. 第三次培训_高晓泓老师PPT：P34-P39
4. Excel相关系数矩阵可视化
5. R相关系数矩阵可视化
6. 主成分分析（PCA）的推导和应用