使用R进行主成分分析

Alex / 2024-03-21 / free_learner@163.com / AlexBrain.cn

本文介绍在R中进行主成分分析（Principal Component Analysis, PCA）的基本方法。

一、主成分分析的基本原理

假设$x = (x_1, x_2, \ldots, x_m)$表示一个随机向量，包含$m$个我们可以观测到的变量（假设称为原始变量），PCA的目标就是找到一个矩阵P，使得$y = Px, Cov(y) = I$，其中Cov表示协方差，所以PCA可以理解为通过对原始变量进行线性组合得到新的m个变量（通常称为主成分分数，PC Score），这m个新变量之间是正交的（相关系数为0），矩阵P中的每一行表示每个原始变量对主成分分数的贡献或者载荷（Loadings）。PCA的另一个特点是，得到的m个新变量对原始变量方差的解释程度是不同的，第一个主成分分数能解释最大的方差，越往后越小。所以在原始变量相关性很高的情况下（意味着可能有很多信息冗余），通过PCA我们一般只需要保留第一个或者前两个主成分分数就可以保留数据中绝大多数变异，从而达到降维的目的。

二、在R中进行主成分分析

这里我们使用R自带的`iris`数据集，包含5个变量，前4个变量分别表示花萼长度、花萼宽度、花瓣长度、花瓣宽度，第5个变量表示花的种类。

```
> data("iris")
> str(iris)
'data.frame':    150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1
1 ...
```

在进行PCA之前，有必要查看一下原始变量之间的相关程度，如果相关程度很低，往往就没有必要进行PCA了。

```
> cor(iris[,-5])
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
```

这里使用`prcomp`函数进行PCA，选项`center=TRUE`和`scale.=TRUE`表示在进行PCA之前，对每个原始变量进行标准化（使得均值为0，方差为1）。返回的结果中，`x`和`rotation`分别表示主成分分数和载荷，`center`和`scale`表示每个原始变量的均值和标准差，`sdev`表示每个主成分分数的标准差。

```
> pca_dat <- prcomp(iris[,-5], center = TRUE, scale. = TRUE)
> names(pca_dat)
[1] "sdev"     "rotation" "center"   "scale"    "x"
```

通过查看载荷，我们才能理解主成分分数和原始变量之间的关系。比如，对于PC1，花萼长度、花瓣长度和花瓣宽度载荷都是正值，而且大小差不多，花萼宽度载荷是负值，那么第一主成分分数越大，就表示花萼长度、花瓣长度和花瓣宽度越大，花萼宽度越小。对于PC2，四个原始变量的载荷都为负值，而且花瓣长度和花瓣宽度的载荷接近于0，那么第二主成分分数越大，就表示花萼长度和花萼宽度越小。

```
> pca_dat$rotation
                    PC1         PC2        PC3        PC4
Sepal.Length  0.5210659 -0.37741762  0.7195664  0.2612863
Sepal.Width  -0.2693474 -0.92329566 -0.2443818 -0.1235096
Petal.Length  0.5804131 -0.02449161 -0.1421264 -0.8014492
Petal.Width   0.5648565 -0.06694199 -0.6342727  0.5235971
```

通过查看每个主成分分数解释的方差比例，我们发现前两个主成分已经可以解释超过95%的方差，那么我们可以只用前两个主成分来表示数据，而不用担心丢失了重要信息。

```
> summary(pca_dat)
Importance of components:
                          PC1    PC2     PC3     PC4
Standard deviation     1.7084 0.9560 0.38309 0.14393
Proportion of Variance 0.7296 0.2285 0.03669 0.00518
Cumulative Proportion  0.7296 0.9581 0.99482 1.00000
```

下面的图表示不同种类的花在前两个主成分分数上的变异，可以看到PC1可以很好的区分`setosa`和`versicolor/virginica`，根据PC1的载荷，我们知道这可能意味着`setosa`的花萼长

度、花瓣长度和花瓣宽度要比`versicolor/virginica`小，而花萼宽度要比`versicolor/virginica`大。