特征降维

黑马程序员
www.itheima.com

传智教育旗下
高端IT教育品牌

目录

Contents

学习目标

Learning Objectives

1. 理解特征降维的作用
2. 知道低方差过滤法
3. 知道相关系数法
4. 掌握PCA 进行降维

特征降维

- 为什么要进行特征降维？

  特征对训练模型时非常重要的；用于训练的数据集包含一些不重要的特征，可能导致模型泛化性能不佳

  eg：某些特征的取值较为接近，其包含的信息较少

  eg：希望特征独立存在对预测产生影响，两个特征同增同减非常相关，不会给模型带来更多的信息

- 特征降维目的？

  - 指在某些限定条件下，降低特征个数
  - 特征降维涉及的知识面比较多，当前阶段常用的方法：

    （1）低方差过滤法

    （2）PCA（主成分分析）降维法

    （3）相关系数（皮尔逊相关系数、斯皮尔曼相关系数）

特征降维 - 低方差过滤法

- 低方差过滤法：指的是删除方差低于某些阈值的一些特征
  - 特征方差小：特征值的波动范围小，包含的信息少，模型很难学习到信息
  - 特征方差大：特征值的波动范围大，包含的信息相对丰富，便于模型进行学习

- 低方差过滤API
  - sklearn.feature_selection.VarianceThreshold(threshold = 0.0)

    实例化对象用于删除所有低方差特征
  - variance_obj.fit_transform(X)

    X:numpy array格式的数据[n_samples,n_features]
  - 返回值：训练集差异低于threshold的特征将被删除。

    默认值是保留所有非零方差特征，即删除所有样本中具有相同值的特征

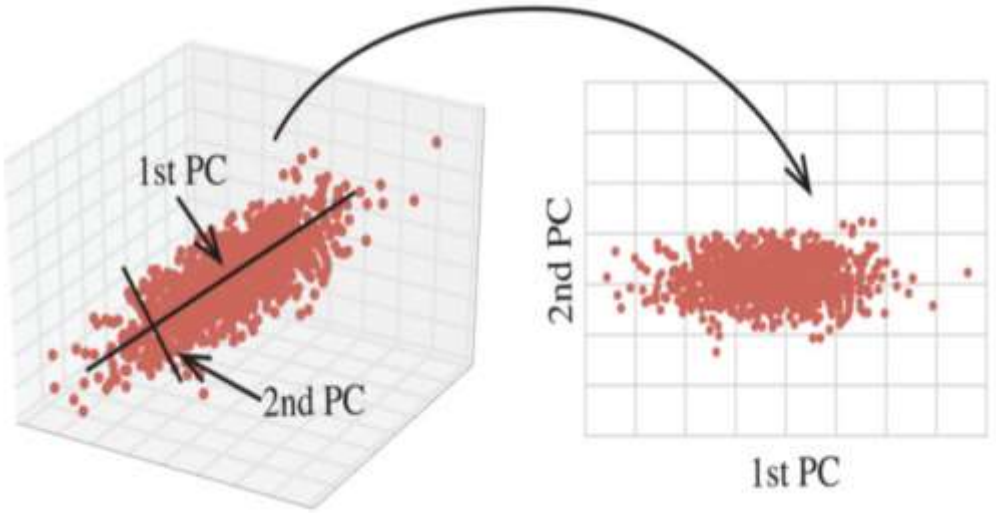
低方差过滤

```
# 1.导入依赖包
from sklearn.feature_selection import VarianceThreshold
import pandas as pd

# 2. 读取数据集
data = pd.read_csv('data/垃圾邮件分类数据.csv')
print(data.shape) # (971, 25734)


# 3. 使用方差过滤法
transformer = VarianceThreshold(threshold=0.1)
data = transformer.fit_transform(data)
print(data.shape) # (971, 1044)
```

主成分分析PCA

- 主成分分析（Principal Component Analysis，PCA）

PCA 通过对数据维数进行压缩，尽可能降低原数据的维数(复杂度)
损失少量信息，在此过程中可能会舍弃原有数据、创造新的变量。



- 主成分分析API
  - sklearn.decomposition.PCA(n_components=None)

    将数据分解为较低维数空间

    n_components: 小数表示保留百分之多少的信息；整数表示减少到多少特征　eg：由20个特征减少到10个
  - mypcaobj.fit_transform(X)
  - 返回值：转换后指定维度的array

主成分分析PCA

```
# 1.导入依赖包
from sklearn.decomposition import PCA
from sklearn.datasets import load_iris

# 2. 加载数据集
x, y = load_iris(return_X_y=True)
print(x[:5])


# 3. PCA,保留指定比例的信息
transformer = PCA(n_components=0.95)
x_pca = transformer.fit_transform(x)
print(x_pca[:5])


# 4. PCA,保留指定数量特征
transformer = PCA(n_components=2)
x_pca = transformer.fit_transform(x)
print(x_pca[:5])
```

特征降维 - 相关系数

- 为什么会使用相关系数?
  - 相关系数：反映特征列之间（变量之间）密切相关程度的统计指标
  - 常见2个相关系数：皮尔逊相关系数、斯皮尔曼相关系数

  - 相关系数的值介于-1与+1之间，即-1 ≤ r ≤ +1。其性质如下：

    当 r > 0 时，表示两变量正相关，r < 0 时，两变量为负相关

    当 |r| = 1 时，表示两变量为完全相关，当r = 0时，表示两变量间无相关关系

    当 0 < |r| < 1时，表示两变量存在一定程度的相关。

    且|r|越接近1，两变量间线性关系越密切；|r|越接近于0，表示两变量的线性相关越弱
  - 一般可按三级划分：

    (1) |r| <0.4为低度相关;

    (2) 0.4≤ |r| <0.7为显著性相关;

    (3) 0.7 ≤ |r| <1为高度线性相关。

皮尔逊相关系数

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}}$$

- 举个例子：已知广告投入x特征与月均销售额y之间的关系，经过皮尔逊相关系数计算，为高度相关

年广告费投入	月均销售额
12.5	21.2
15.3	23.9
23.2	32.9
26.4	34.1
33.5	42.5
34.4	43.2
39.4	49.0
45.2	52.8
55.4	59.4
60.9	63.5

序号	广告投入(万元) x	月均销售额(万元) y	x^2	y^2	xy
1	12.5	21.2	156.25	449.44	265.00
2	15.3	23.9	234.09	571.21	365.67
3	23.2	32.9	538.24	1082.41	763.28
4	26.4	34.1	696.96	1162.81	900.24
5	33.5	42.5	1122.25	1806.25	1423.75
6	34.4	43.2	1183.36	1866.24	1486.08
7	39.4	49.0	1552.36	2401.00	1930.60
8	45.2	52.8	2043.04	2787.84	2386.56
9	55.4	59.4	3069.16	3528.36	3290.76
10	60.9	63.5	3708.81	4032.25	3867.15
合计	346.2	422.5	14304.52	19687.81	16679.09

$$\frac{10 \times 16679.09 - 346.2 \times 422.5}{\sqrt{10 \times 14304.52 - 346.2^2}\sqrt{10 \times 19687.81 - 422.5^2}} = 0.9942$$

斯皮尔曼相关系数

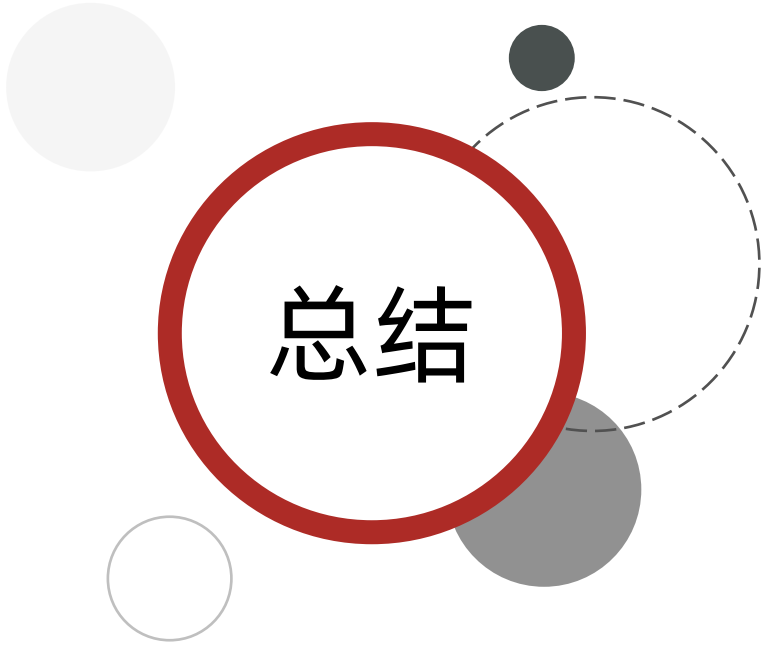
$$RankIC = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

n为等级个数，d为成对变量的等级差数

身高 (X)	等级	睡眠时间 (Y)	等级	d_i	d_i^2
160	3.5	7.6	3	0.5	0.25
168	5	8.0	4	1	1
174	6	8.8	5	1	1
141	1	7.5	2	-1	1
160	3.5	6.9	1	2.5	6,25
159	2	8.9	6	-4	16
176	7	9.0	7	0	0

相关系数

```
# 1.导入依赖包
import pandas as pd
from sklearn.feature_selection import VarianceThreshold
from scipy.stats import pearsonr
from scipy.stats import spearmanr
from sklearn.datasets import load_iris

# 2.读取数据集(鸢尾花数据集)
data = load_iris()
data = pd.DataFrame(data.data, columns=data.feature_names)

# 3. 皮尔逊相关系数
corr = pearsonr(data['sepal length (cm)'], data['sepal width (cm)'])
print(corr, '皮尔逊相关系数:', corr[0], '不相关性概率:', corr[1])
# (-0.11756978413300204, 0.15189826071144918) 皮尔逊相关系数: -0.11756978413300204 不相关性概率:
0.15189826071144918

# 4. 斯皮尔曼相关系数
corr = spearmanr(data['sepal length (cm)'], data['sepal width (cm)'])
print(corr, '斯皮尔曼相关系数:', corr[0], '不相关性概率:', corr[1])
# SpearmanrResult(correlation=-0.166777658283235, pvalue=0.04136799424884587) 斯皮尔曼相关系数: -0.166777658283235 不
相关性概率: 0.04136799424884587
```



1 特征降维

指在某些限定条件下，降低特征个数

2 低方差过滤法

删除方差低于某些阈值的一些特征

3 PCA主成分分析

通过数据压缩实现特征降维，在此过程中去除特征之间的线性相关性

4 相关系数法

皮尔逊相关系数、斯皮尔曼相关系数，通过相关系数法可以实现减少特征的目的

练习

1、下列关于PCA的说法错误的是（单选题）：

A） 它可以通过sklearn.decomposition.PCA来实现降维

B） 它的目的是要找到特征数据中的主要成分，然后删除所有非主要成分数据

C） PCA中的n_components参数可以指定为小数

D） PCA中的n_components参数可以指定为整数

答案：B


黑马程序员
www.itheima.com
传智教育旗下高端IT教育品牌