# 1.什么是特征选择

|  |
| --- |
|  |

# 2.特征选择的方法

|  |
| --- |
|  |

# 3.API模块

|  |
| --- |
|  |

# 4.过滤式特征选择方法

## 1>低方差特征过滤

|  |
| --- |
|  |

### 数值计算

|  |
| --- |
|  |
|  |

### 案例: feature\_selection\_demo.py

|  |
| --- |
| # 低方差特征过滤 from sklearn.feature\_selection import VarianceThreshold import pandas as pd   # 1.加载数据 data = pd.read\_csv("./factor\_returns.csv") # 1.2 把不是特征值的的列过滤掉 # data = data.iloc[:,1:10] data = data.iloc[:,1:-2] # 2.创建转换器对象 # transfer = VarianceThreshold() # 里面有一个此时threshold默认是0 transfer = VarianceThreshold(threshold=15) # 可以根据需要设置阈值,过滤掉没有太多意义的特征 # 3.调用fit\_transform()方法 ret\_data = transfer.fit\_transform(data) print(ret\_data.shape) |

### 结果

|  |
| --- |
|  |

## 2.相关系数

|  |
| --- |
|  |
|  |
|  |

### 扩展:数据是否是有问题?投入12.5万,销售21.2万,投入60.9万,销售才63.5万,赚到的钱都用于做广告?

### 相关系数的特点

|  |
| --- |
|  |

### 相关API模块,居然是scipy的模块?

|  |
| --- |
|  |

### 案例: feature\_selection\_demo\_pearsonr.py

|  |
| --- |
| # 特征过滤:相关系数,居然需要使用scipy.stats模块,不过需要配合VarianceThreshold from sklearn.feature\_selection import VarianceThreshold import pandas as pd from scipy.stats import pearsonr # 相关系数算法模块  # 1.加载数据 data = pd.read\_csv("./factor\_returns.csv") # 1.2 把不是特征值的的列过滤掉 # data = data.iloc[:,1:10] data = data.iloc[:, 2:-2] # 2.创建转换器对象 # transfer = VarianceThreshold() # 里面有一个此时threshold默认是0 transfer = VarianceThreshold(threshold=15) # 可以根据需要设置阈值,过滤掉没有太多意义的特征 # 3.调用fit\_transform()方法 ret\_data = transfer.fit\_transform(data) print(ret\_data.shape) # 4.计算特性的相关性系数 r = pearsonr(data["revenue"], data["total\_expense"]) # 交税额和总费用的相关性 print("相关系数=",r) |

#### 结果

|  |
| --- |
|  |

#### 这两个特征的相关性比较低

#### 注意:老师的代码有一点问题, 000728.XSHE这样子的数据和整数其实无法比较

## 查看特征之间关系图,需要使用Matplotlib库,我们先把它导入

|  |
| --- |
|  |

## 然后我们可以绘制散点图

|  |
| --- |
| # 特征过滤:相关系数,居然需要使用scipy.stats模块,不过需要配合VarianceThreshold from sklearn.feature\_selection import VarianceThreshold import pandas as pd from scipy.stats import pearsonr # 相关系数算法模块 from matplotlib import pyplot as plt  # 1.加载数据 data = pd.read\_csv("./factor\_returns.csv") # 1.2 把不是特征值的的列过滤掉 # data = data.iloc[:,1:10] data = data.iloc[:, 2:-2] # 2.创建转换器对象 # transfer = VarianceThreshold() # 里面有一个此时threshold默认是0 transfer = VarianceThreshold(threshold=15) # 可以根据需要设置阈值,过滤掉没有太多意义的特征 # 3.调用fit\_transform()方法 ret\_data = transfer.fit\_transform(data) print(ret\_data.shape) # 4.计算特性的相关性系数 r = pearsonr(data["revenue"], data["total\_expense"]) # 交税额和总费用的相关性 print("相关系数=",r)  # 绘制散点图 plt.figure(figsize=(20,8),dpi=100) plt.scatter(data["revenue"], data["total\_expense"]) plt.show() |

### 我的图和老师的不太一样.不过,没有关系

## 如果几个特征的相关性很高,我们应该怎么处理?

### 1>选取一个作为代表

### 2>对这几个特征采取加权求和的方法得到一个新特征来代替这几个特征

### 3>主成分分析