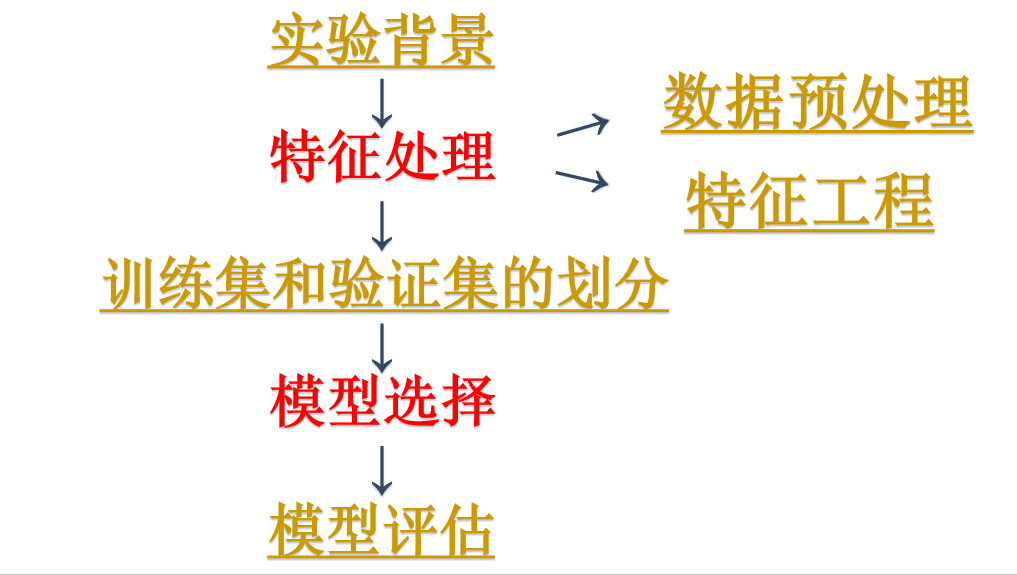
特征工程是通过对原始数据的处理和加工，将原始数据属性通过处理转换为数据特征的过程，属性是数据本身具有的维度，特征是数据中所呈现出来的某一种重要的特性，通常是通过属性的计算，组合或转换得到的。比如主成分分析就是将大量的数据属性转换为少数几个特征的过程。某种程度而言，好的数据以及特征往往是一个性能优秀模型的基础。





import numpy as np

import xlrd #引入xlrd模块

import xlwt #引入xlwd模块

def open\_excel(file):

try:

data = xlrd.open\_workbook(file) #打开excel文件

return data

except Exception as e:

print(str(e))

def split\_age\_range(age):

"""

将特征值年龄进行离散化为8个特征值

:param age: 年龄区间值

:return: 离散化后的特征

"""

if age == 0:

return [1,0,0,0,0,0,0,0,0]

elif age == 1:

return [0,1,0,0,0,0,0,0,0]

elif age == 2:

return [0,0,1,0,0,0,0,0,0]

elif age == 3:

return [0,0,0,1,0,0,0,0,0]

elif age == 4:

return [0,0,0,0,1,0,0,0,0]

elif age == 5:

return [0,0,0,0,0,1,0,0,0]

elif age == 6:

return [0,0,0,0,0,0,1,0,0]

elif age == 7 or age == 8:

return [0,0,0,0,0,0,0,1,0]

# elif age == 8:

# return [0,0,0,0,0,0,0,0,1]

def split\_gender(gender):

"""

将特征值性别进行离散化

:param gender:

:return: 返回离散化的特征

"""

if gender == 0:

return [1,0,0]

elif gender == 1:

return [0,1,0]

elif gender == 2:

return [0,0,1]

def split\_log(Log):

"""

分割数据文件中的Log数据

:param Log: Log数据

:return: 处理后的特征值

"""

items = Log.strip().split('#')

purchase = 0;total = 0

click = 0;add\_to\_card = 0;add\_to\_favourite = 0

for i in range(len(items)):

total += 1

item = items[i].strip().split(':')

if item[4] == '2':

purchase += 1

if item[4] == '1':

add\_to\_card += 1

if item[4] == '3':

add\_to\_favourite += 1

return [float(total),float(round(purchase/total,3)),float(add\_to\_card),float(add\_to\_favourite)]

def loadDataSet(path, training\_sample,colnameindex=0,by\_name=u'Sheet1'):

"""

加载数据

:param path: 数据文件存放路径

:param training\_sample: 数据文件名

:param colnameindex: 文件列名下标

:param by\_name: 表名

:return: 数据集和类别标签

"""

dataMat = [];

labelMat = [] # 定义列表

filename = path + training\_sample

data = open\_excel(filename) #获取文件

table = data.sheet\_by\_name(by\_name) # 获取Sheet1

nrows = table.nrows # 拿到总共行数

colnames = table.row\_values(colnameindex) # 某一行数据 ['user\_id', 'age\_range', 'gender', 'merchant\_id','label']

for rownum in range(1, nrows): # 也就是从Excel第二行开始，第一行表头不算

row = table.row\_values(rownum)

if row[1] == '' or row[2] == '' or row[5] == '':

continue

if row:

app = [] #定义列表

app = split\_age\_range(row[1])+split\_gender(row[2]) + split\_log(row[5]) # 将Log转化为特征值

dataMat.append(app)

labelMat.append(float(row[4])) # 获取类别标签

return dataMat, labelMat

def main():

"""

主函数

:return: null

"""

wb = xlwt.Workbook()

ws = wb.add\_sheet('sheet1',cell\_overwrite\_ok=True)

path = "F:\\AIData\Ch05\\"

training\_sample = 'train\_data.xlsx' # 训练数据文件

trainingSet, trainingLabels = loadDataSet(path, training\_sample) # 取训练数据

# print(len(trainingSet))

num = len(trainingSet)

for i in range(num):

for j in range(16):

ws.write(i,j,trainingSet[i][j])

ws.write(i,j+1,trainingLabels[i])

wb.save('D:\\featuredata.xls')

print("处理完成")

if \_\_name\_\_ == '\_\_main\_\_':

"""

程序入口

"""

main()