Spark逻辑回归(LR)模型使用介绍

In [2]:

```
#小案例学习spark LR模型的使用
from pyspark.ml.feature import VectorAssembler
import pandas as pd
sample_dataset = [
    (0, "male", 37, 10, "no", 3, 18, 7, 4),
    (0, "female", 27, 4, "no", 4, 14, 6, 4),
    (0, "female", 32, 15, "yes", 1, 12, 1, 4),
    (0, "male", 57, 15, "yes", 5, 18, 6, 5),
    (0, "male", 22, 0.75, "no", 2, 17, 6, 3),
    (0, "female", 32, 1.5, "no", 2, 17, 5, 5),
    (0, "female", 22, 0.75, "no", 2, 12, 1, 3),
    (0, "male", 57, 15, "yes", 2, 14, 4, 4),
    (0, "female", 32, 15, "yes", 4, 16, 1, 2),
    (0, "male", 22, 1.5, "no", 4, 14, 4, 5),
    (0, "male", 37, 15, "yes", 2, 20, 7, 2),
    (0, "male", 27, 4, "yes", 4, 18, 6, 4),
    (0, "male", 47, 15, "yes", 5, 17, 6, 4),
    (0, "female", 22, 1.5, "no", 2, 17, 5, 4),
    (0, "female", 27, 4, "no", 4, 14, 5, 4),
    (0, "female", 37, 15, "yes", 1, 17, 5, 5),
    (0, "female", 37, 15, "yes", 2, 18, 4, 3),
    (0, "female", 22, 0.75, "no", 3, 16, 5, 4),
    (0, "female", 22, 1.5, "no", 2, 16, 5, 5),
    (0, "female", 27, 10, "yes", 2, 14, 1, 5),
    (1, "female", 32, 15, "yes", 3, 14, 3, 2),
    (1, "female", 27, 7, "yes", 4, 16, 1, 2),
    (1, "male", 42, 15, "yes", 3, 18, 6, 2),
    (1, "female", 42, 15, "yes", 2, 14, 3, 2),
    (1, "male", 27, 7, "yes", 2, 17, 5, 4),
    (1, "male", 32, 10, "yes", 4, 14, 4, 3),
    (1, "male", 47, 15, "yes", 3, 16, 4, 2),
    (0, "male", 37, 4, "yes", 2, 20, 6, 4)
]
columns = ["affairs", "gender", "age", "label", "children", "religiousness", "educati
# pandas构建dataframe，方便
pdf = pd.DataFrame(sample_dataset,columns=columns)
df = spark.createDataFrame(pdf)
# 特征选取：affairs为目标值，其余为特征值
df2 = df.select("affairs","age", "religiousness", "education", "occupation", "rating")
colArray2 = ["age", "religiousness", "education", "occupation", "rating"]
df3 = VectorAssembler().setInputCols(colArray2).setOutputCol('features').transform(df2
print('数据集：')
df3.show(truncate=False)
# 随机切分 训练集和测试集
trainDF,testDF = df3.randomSplit([0.8,0.2])
print('训练集：')
trainDF.show(10,truncate=False)
print('测试集：')
testDF.show(10,truncate=False)
```

数据集：

```
+-------+---+-------------+---------+----------+------+----------------------+
|affairs|age|religiousness|education|occupation|rating|features              |
+-------+---+-------------+---------+----------+------+----------------------+
|0      |37 |3            |18       |7         |4     |[37.0,3.0,18.0,7.0,4.0]|
|0      |27 |4            |14       |6         |4     |[27.0,4.0,14.0,6.0,4.0]|
|0      |32 |1            |12       |1         |4     |[32.0,1.0,12.0,1.0,4.0]|
```

```
|0      |57 |5           |18       |6         |5     |[57.0,5.0,18.0,6.0,5.0]|
|0      |22 |2           |17       |6         |3     |[22.0,2.0,17.0,6.0,3.0]|
|0      |32 |2           |17       |5         |5     |[32.0,2.0,17.0,5.0,5.0]|
|0      |22 |2           |12       |1         |3     |[22.0,2.0,12.0,1.0,3.0]|
|0      |57 |2           |14       |4         |4     |[57.0,2.0,14.0,4.0,4.0]|
|0      |32 |4           |16       |1         |2     |[32.0,4.0,16.0,1.0,2.0]|
|0      |22 |4           |14       |4         |5     |[22.0,4.0,14.0,4.0,5.0]|
|0      |37 |2           |20       |7         |2     |[37.0,2.0,20.0,7.0,2.0]|
|0      |27 |4           |18       |6         |4     |[27.0,4.0,18.0,6.0,4.0]|
|0      |47 |5           |17       |6         |4     |[47.0,5.0,17.0,6.0,4.0]|
|0      |22 |2           |17       |5         |4     |[22.0,2.0,17.0,5.0,4.0]|
|0      |27 |4           |14       |5         |4     |[27.0,4.0,14.0,5.0,4.0]|
|0      |37 |1           |17       |5         |5     |[37.0,1.0,17.0,5.0,5.0]|
|0      |37 |2           |18       |4         |3     |[37.0,2.0,18.0,4.0,3.0]|
|0      |22 |3           |16       |5         |4     |[22.0,3.0,16.0,5.0,4.0]|
|0      |22 |2           |16       |5         |5     |[22.0,2.0,16.0,5.0,5.0]|
|0      |27 |2           |14       |1         |5     |[27.0,2.0,14.0,1.0,5.0]|
+-------+---+------------+---------+----------+------+-----------------------+
only showing top 20 rows
```

训练集：

```
+-------+---+------------+---------+----------+------+-----------------------+
|affairs|age|religiousness|education|occupation|rating|features               |
+-------+---+------------+---------+----------+------+-----------------------+
|0      |32 |1           |12       |1         |4     |[32.0,1.0,12.0,1.0,4.0]|
|0      |37 |3           |18       |7         |4     |[37.0,3.0,18.0,7.0,4.0]|
|0      |57 |5           |18       |6         |5     |[57.0,5.0,18.0,6.0,5.0]|
|0      |32 |2           |17       |5         |5     |[32.0,2.0,17.0,5.0,5.0]|
|0      |57 |2           |14       |4         |4     |[57.0,2.0,14.0,4.0,4.0]|
|0      |22 |4           |14       |4         |5     |[22.0,4.0,14.0,4.0,5.0]|
|0      |37 |2           |20       |7         |2     |[37.0,2.0,20.0,7.0,2.0]|
|0      |22 |2           |17       |5         |4     |[22.0,2.0,17.0,5.0,4.0]|
|0      |27 |4           |14       |5         |4     |[27.0,4.0,14.0,5.0,4.0]|
|0      |37 |1           |17       |5         |5     |[37.0,1.0,17.0,5.0,5.0]|
+-------+---+------------+---------+----------+------+-----------------------+
only showing top 10 rows
```

测试集：

```
+-------+---+------------+---------+----------+------+-----------------------+
|affairs|age|religiousness|education|occupation|rating|features               |
+-------+---+------------+---------+----------+------+-----------------------+
|0      |27 |4           |14       |6         |4     |[27.0,4.0,14.0,6.0,4.0]|
|0      |22 |2           |12       |1         |3     |[22.0,2.0,12.0,1.0,3.0]|
|0      |22 |2           |17       |6         |3     |[22.0,2.0,17.0,6.0,3.0]|
|0      |27 |4           |18       |6         |4     |[27.0,4.0,18.0,6.0,4.0]|
|0      |32 |4           |16       |1         |2     |[32.0,4.0,16.0,1.0,2.0]|
|0      |47 |5           |17       |6         |4     |[47.0,5.0,17.0,6.0,4.0]|
|0      |37 |2           |20       |6         |4     |[37.0,2.0,20.0,6.0,4.0]|
|1      |42 |3           |18       |6         |2     |[42.0,3.0,18.0,6.0,2.0]|
+-------+---+------------+---------+----------+------+-----------------------+
```

In [3]:

```
# 逻辑回归训练模型
from pyspark.ml.classification import LogisticRegression
lr = LogisticRegression()
model = lr.setLabelCol('affairs').setFeaturesCol('features').fit(trainDF)
model.transform(testDF).show(truncate=False)
```

```
+-------+---+-------------+---------+----------+------+----------------------+--------
--------------------------------+----------------------------------------+----------+
|affairs|age|religiousness|education|occupation|rating|features              |rawPredi
ction                           |probability                             |prediction|
+-------+---+-------------+---------+----------+------+----------------------+--------
--------------------------------+----------------------------------------+----------+
|0      |27 |4            |14       |6         |4     |[27.0,4.0,14.0,6.0,4.0]|[1.43701
28406197153,-1.4370128406197153]|[0.8079916464280155,0.19200835357198462] |0.0       |
|0      |22 |2            |12       |1         |3     |[22.0,2.0,12.0,1.0,3.0]|[-2.6357
535531749416,2.6357535531749416]|[0.06687252951201492,0.9331274704879852] |1.0       |
|0      |22 |2            |17       |6         |3     |[22.0,2.0,17.0,6.0,3.0]|[0.65492
44841168171,-0.6549244841168171]|[0.65811932430713,0.34188067569287]      |0.0       |
|0      |27 |4            |18       |6         |4     |[27.0,4.0,18.0,6.0,4.0]|[2.37007
15957681684,-2.3700715957681684]|[0.91451645780968,0.08548354219031987]   |0.0       |
|0      |32 |4            |16       |1         |2     |[32.0,4.0,16.0,1.0,2.0]|[-4.6286
85738947233,4.628685738947233]  |[0.009673104928627526,0.9903268950713724]|1.0       |
|0      |47 |5            |17       |6         |4     |[47.0,5.0,17.0,6.0,4.0]|[3.05035
41549910054,-3.0503541549910054]|[0.954797813935877,0.04520218606412298]  |0.0       |
|0      |37 |2            |20       |6         |4     |[37.0,2.0,20.0,6.0,4.0]|[4.93127
37927811945,-4.9312737927811945]|[0.992834412152514,0.0071655878474860945]|0.0       |
|1      |42 |3            |18       |6         |2     |[42.0,3.0,18.0,6.0,2.0]|[-0.5982
88626751545,0.598288626751545]  |[0.354735326530606,0.6452646734693941]   |1.0       |
+-------+---+-------------+---------+----------+------+----------------------+--------
--------------------------------+----------------------------------------+----------+
```